

2021

Филипп Картаев

Дружелюбная эконометрика

Оглавление

[Предисловие](#)

[Предисловие для преподавателей. РПД](#)

[Глава 1. Что такое эконометрика и зачем она нужна](#)

[1.1. Чем занимаются эконометристы](#)

[1.2. Корреляция и причинно-следственная связь: некоторые подводные камни](#)

[1.3. Типы данных, используемых в эконометрике](#)

[Глава 2. Парная регрессия](#)

[2.1. Предварительные замечания](#)

[2.2. Метод наименьших квадратов](#)

[2.3. Классическая линейная модель парной регрессии](#)

[2.4. Свойства МНК-оценок](#)

[2.5. Тестирование гипотез и построение доверительных интервалов](#)

[2.6. Прогнозирование](#)

[2.7. Заключение](#)

[Глава 3. Множественная регрессия: основы](#)

[3.1. Почему не стоит ограничиваться парной регрессией](#)

[3.2. Классическая линейная модель множественной регрессии](#)

[3.3. Векторно-матричная форма записи и некоторые доказательства](#)

[3.4. Степень соответствия модели данным](#)

[3.5. Тестирование гипотез и построение доверительных интервалов](#)

[3.6. Тест на линейное ограничение общего вида](#)

[3.7. Обобщающий пример](#)

[Приложение 3.A. таблицы распределения Стьюдента и Фишера](#)

[Глава 4. Множественная регрессия: мультиколлинеарность, фиктивные переменные, нелинейные модели](#)

[4.1. Мультиколлинеарность](#)

[4.2. Фиктивные переменные](#)

[4.3. Нелинейные модели](#)

[4.4. Обобщающий пример](#)

[4.5. Рекомендации по оформлению результатов эконометрических расчетов](#)

[Глава 5. Гетероскедастичность, обобщенная линейная модель](#)

[5.1. Гетероскедастичность: определение и последствия](#)

[5.2. Состоятельные в условиях гетероскедастичности стандартные ошибки](#)

[5.3. Взвешенный метод наименьших квадратов](#)

[5.4. Выявление гетероскедастичности](#)

[5.5. Обобщенная линейная модель и обобщенный МНК](#)

[Глава 6. Модель со стохастическими регрессорами и асимптотический подход в эконометрике](#)

[6.1. Некоторые важные результаты математической статистики](#)

[6.2. Линейная регрессионная модель со стохастическими регрессорами](#)

[6.3. Состоятельность МНК-оценок](#)

[6.4. Асимптотическая нормальность МНК-оценок](#)

[6.5. Тестирование гипотез и построение доверительных интервалов](#)

[Приложение 6.A. Состоятельная в условиях гетероскедастичности стандартная ошибка оценки коэффициента: доказательство состоятельности](#)

[Приложение 6.B. Дельта-метод](#)

[Приложение 6.B. Таблицы стандартного нормального распределения и распределения Хи-квадрат](#)

[Глава 7. Проблемы спецификации уравнения регрессии](#)

[7.1. Эндогенность из-за пропуска существенной переменной](#)

[7.2. Эндогенность из-за выбора неверной функциональной формы связи](#)

[7.3. Эндогенность из-за двусторонней причинно-следственной связи](#)

[7.4. Эндогенность из-за ошибок измерения](#)

[7.5. Другие \(помимо эндогенности\) потенциальные угрозы обоснованности выводов эконометрического исследования](#)

[7.6. Чек-лист эконометриста](#)

Глава 8. Инструментальные переменные

- 8.1. Двухшаговый МНК: парная регрессия
- 8.2. Двухшаговый МНК: множественная регрессия
- 8.3. Векторно-матричная форма записи
- 8.4. Тесты для моделей, оцененных двухшаговым МНК
- 8.5. Где взять подходящие инструменты?

Глава 9. Панельные данные

- 9.1. Модель с фиксированными эффектами
- 9.2. Модель с фиктивными переменными
- 9.3. Внутригрупповое преобразование
- 9.4. Модель в первых разностях
- 9.5. Модель со случайными эффектами
- 9.6. Доступный ОМНК для оценивания модели со случайными эффектами
- 9.7. Спецификационные тесты

Глава 10. Модели бинарного выбора

- 10.1. Линейная модель вероятности
- 10.2. Логит-модель: введение
- 10.3. Логит-модель: оценивание и тестирование гипотез
- 10.4. Пробит-модель

Глава 11. Оценка эффекта воздействия

- 11.1. Оценка эффекта воздействия в идеальном эксперименте
- 11.2. Оценка эффекта воздействия методом разности разностей
- 11.3. Локальный средний эффект воздействия (LATE)
- 11.4. Разрывный регрессионный дизайн

Приложение. Эконометрические кейсы

- Кейс 1. Население и расходы бюджета
- Кейс 2. Динамика стоимости работ

Предисловие

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

В этом учебнике я постарался собрать три ингредиента, нужных для знакомства с эконометрикой:

Во-первых, тут есть основы эконометрической теории. Чтобы с ними разобраться, вам потребуется знание математического анализа, линейной алгебры, теории вероятностей и математической статистики. Впрочем, все самые утомительные доказательства вынесены в отдельные параграфы или приложения. Поэтому если ваша цель состоит в знакомстве с ключевыми идеями, а в технические детали вы погружаться не хотите, то их можно будет пропустить. В этом случае требования к базовой математической подготовке будут гораздо мягче.

Основные идеи изложены не только в виде текста, но и в формате видеолекций. Я советую познакомиться и с видеоматериалами, и с классическим текстовым вариантом.

Во-вторых, в учебнике содержится детальное обсуждение применения эконометрики на практике. При рассказе о разных методах я стараюсь пояснить, для чего каждый из них может быть полезен в ваших собственных изысканиях, и показать, как их можно реализовать в специальных эконометрических программах.

Во второй части книги есть много отсылок к хорошим (широко цитируемым) исследованиям, использующим эмпирические методы. Эти примеры помогают понять, как эконометрика работает в реальной жизни, и позволяют начать применять её сразу после знакомства с данной книгой.

В-третьих, в конце глав предложены задания для самостоятельной работы, а в заключительной части учебника приведены решения **каждого** из них. Формат электронного учебника позволил обеспечить для них автоматическую проверку. Поэтому после прохождения теста вы можете сразу получить обратную связь и узнать, сколько заданий вы решили верно. После этого у вас будет возможность попробовать ещё раз или посмотреть подробные комментарии к решению.

Эти примеры могут быть не только полезным для полноценного освоения эконометрики, но и чертовски интересными. Многие из них опираются на реальные прикладные исследования и требуют использования специальных программ для работы с данными.

Некоторое время я размышлял, с каким из эконометрических пакетов познакомить вас в этой книге. В итоге я не смог выбрать единственную программу, так что в учебнике содержатся видеофрагменты, показывающие, как проводить расчеты сразу в двух пакетах на выбор: **в Gretl** и **в R**. Каждый из них распространяется бесплатно, так что вы легко можете скачать и установить его на свой компьютер. Все необходимые ссылки есть в учебнике.

Для полноценного освоения материала нет нужды работать с обеими программами. Выбирайте ту, которая вам больше нравится (соответствующие видеофрагменты в разных программах дублируют друг друга). Gretl великолепен своей интуитивностью и простотой освоения. Программа R — более мощный и универсальный инструмент, но требует больше усилий и квалификации от пользователя.

В любом случае, обратите внимание, что этот учебник не претендует на статус полноценного пособия по R. В этом нет нужды, так как в сети доступно много хороших альтернативных ресурсов, решающих эту задачу. Здесь мы сконцентрируемся только на функционале, необходимом для применения эконометрических методов.

Благодарности

Я хотел бы выразить признательность моим коллегам, чьи ценные соображения и поддержка сделали возможным появление этой работы: Дарье Елицур, Ольге Ключковой, Евгению Лукашу, Ольге Сучковой, Янине Рожиной, Екатерине Ураковой и многим другим. А также Елизавете Майоровой, взявшей на себя труд неоднократно вычитать черновик книги и сделать дизайн для её электронной версии. Наконец, Владимиру Пшенкину и Валерию Скворцову, реализовавшим эту электронную версию.

Я благодарен всем тем замечательным студентам, которым мне довелось читать вводный курс эконометрики: ваша внимательность помогла мне устранить многие неточности в этом учебнике. Общение с вами служит для меня постоянным источником мотивации и новых идей.

Постоянное обновление

Все ошибки в итоговой версии работы полностью остаются на совести автора. Я буду признателен, если, обнаружив их, вы сообщите мне по электронной почте kartaev@gmail.com. Пожалуйста, дайте мне знать и про содержательные ошибки (например, опечатки в формулах), и про ситуации, когда сайт работает не так, как нужно и у вас на экране что-то отображается не так, как должно.

Учебник находится в режиме постоянного обновления.

Предисловие для преподавателей. РПД

Учебник содержит все темы, необходимые в рамках современного базового курса эконометрики пространственных и панельных данных. В зависимости от начального уровня подготовки студентов и количества часов он может быть полностью изложен в рамках одного семестра, а может стать основой для годичного курса (во втором случае лучше дополнить его введением в методы анализа временных рядов).

Не вполне традиционной для вводных курсов является глава 11, которая посвящена оценке эффектов воздействия. Однако важно познакомить студентов с этой идеологией, так как она стала основой многих современных прикладных работ, посвященных выявлению причинно-следственных связей и оценке последствий политики.

Любой автор вводного учебника по эконометрике вынужден принять несколько важных решений по поводу подхода к изложению материала:

- начать с модели детерминированных регрессоров или стохастических;
- обсудить подробно свойства оценок для конечных выборок или сразу делать акцент на асимптотических свойствах;
- начать со случая гомоскедастичности случайных ошибок или со случая гетероскедастичности?

Большинство вводных учебников (см., например, Магнус, Катышев, Пересецкий, 2007, Dougherty, 2011, Носко, 2011) делают выбор в пользу первого варианта в каждом из трех перечисленных пунктов, то есть начинают с классической линейной модели. Некоторые более поздние работы в каждом из указанных пунктов, напротив, выбирают вторую альтернативу (см., Сток, Ватсон, 2015). Преимущество второго пути заключается в возможности сразу приблизиться к методологии большинства современных прикладных исследований (например, практически все статьи в хороших журналах при работе с пространственными или панельными выборками по умолчанию предполагают, что в данных присутствует гетероскедастичность). Недостаток же этого варианта состоит в том, что он более сложен для студентов, если они впервые в жизни знакомятся с эконометрикой, и если вы хотите излагать материал достаточно строго.

В этом учебнике используется компромиссный путь. Первые главы опираются на предпосылки классической линейной модели с детерминированными регрессорами. Однако довольно быстро мы отказываемся от наименее правдоподобных её предположений (таких как нормальность случайных ошибок, постоянство их дисперсии или неслучайность регрессоров), переходя к более реалистичным моделям. Как показывает мой опыт преподавания вводного курса эконометрики, такая логика, с одной стороны, помогает комфортно освоить ключевые идеи эконометрики студентам с разным уровнем начальной подготовки, а с другой, дает возможность показать, как теория связана с практическими приложениями.

Если вы используете этот учебник в преподавании в вашем вузе — рад предложить [Образец рабочей программы дисциплины «Эконометрика»](#), напишите мне, поделюсь доступом к дополнительным методическим материалам (примерам задач, контрольным работам), которые помогут в организации учебного курса. Если формальные правила вашего университета не позволяют ссылаться на интерактивный учебник, можно указать в списке литературы его «традиционную» бумажную версию (которая тоже доступна по запросу):

Картаев Ф. С. Введение в эконометрику. — Экономический факультет МГУ имени М. В. Ломоносова Москва, 2019. — 472 с.

Глава 1. Что такое эконометрика и зачем она нужна

В первом параграфе главы вы можете найти описание того, что такое эконометрика и какие задачи она позволяет решать. Второй параграф призван показать некоторые типичные трудности, с которыми сталкиваются эконометристы, чтобы продемонстрировать, что в этой науке всё не так уж и просто, и мотивировать вас приложить усилия для изучения продвинутых методов. Наконец, в третьем параграфе речь идёт о том, без чего прикладная эконометрика невозможна — о данных.

[1.1. Чем занимаются эконометристы](#)

[1.2. Корреляция и причинно-следственная связь: некоторые подводные камни](#)

[1.3. Типы данных, используемых в эконометрике](#)

[Задания для самостоятельного решения](#)

1.1. Чем занимаются эконометристы

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

В литературе можно встретить много определений эконометрической науки. Например, такое:

Эконометрика— это наука, изучающая количественные и качественные экономические взаимосвязи с помощью математических и статистических методов и моделей.

Однако лучше всего для понимания того, что представляет собой наука эконометрика, выяснить, какие задачи можно решать с ее помощью:

1. описательный (дескриптивный) анализ,
2. выявление причинно-следственных связей между переменными,
3. прогнозирование.

Поясним суть каждого из этих пунктов и приведем примеры.

Описательный (дескриптивный) анализ

Речь идет о количественных оценках зависимостей между переменными *без выявления направления причинно-следственных связей*.

Об этом вы уже много знаете из курса математической статистики. Например, скорее всего, вы умеете вычислять коэффициент корреляции, сравнивать средние значения в выборках и тестировать гипотезы о соотношении средних друг с другом. В рамках данного учебника мы обсудим и более продвинутые техники.

Рассмотрим такой пример. Некоторым исследователям интересно, есть ли корреляция между здоровым образом жизни (скажем, количеством времени, которое индивид в течение месяца посвящает тренировкам) и заработной платой.

Более тонкий вопрос: сохранится ли эта корреляция, если учесть влияние прочих важных факторов, которые могут быть связаны и со склонностью к занятиям физкультурой, и с доходами (например, возраст, здоровье, страна проживания)? Для ответа потребуется нечто большее, чем простой парный коэффициент корреляции, так как он не позволяет учесть влияние прочих факторов. Например, тут может пригодиться множественная регрессия (см. главу 3). Тем не менее, какой бы инструмент вы ни использовали (коэффициент корреляции, диаграмму рассеяния, регрессию и т. д.), пока вы не зададитесь вопросом о том, где причина, а где следствие¹, подобный анализ остается дескриптивным.

Выявление причинно-следственных связей между переменными

В отличие от предыдущего случая здесь речь идет не просто о наличии корреляции, а о попытке ответа на вопрос: является ли изменение переменной X причиной изменения переменной Y ? Идея о том, что корреляция и причинно-следственная связь — это совсем не одно и то же — одна из ключевых идей эконометрики, и в рамках нашего курса мы будем возвращаться к ней снова и снова, сопровождая рассказ примерами (в том числе уже в следующем параграфе).

Вот несколько вопросов о причинно-следственных связях, на которые умеют отвечать эконометристы:

1. *Что произойдет с уровнем преступности, если принять закон, разрешающий гражданам носить личное огнестрельное оружие?* Это нетривиальный вопрос, дискуссия по поводу которого ведется и среди политиков, и среди экспертов, и среди простых обывателей. Сторонники закона утверждают, что его введение позволит снизить преступность, так как граждане получают возможность защититься от злоумышленников. Их оппоненты возражают, что в результате введения такого закона преступность, наоборот, вырастет из-за избыточного количества огнестрельного оружия на руках у населения и его спонтанного использования. Оказывается, что при наличии достаточного массива данных ответ на этот вопрос вполне может быть получен с помощью подходящих эконометрических методов. Мы обратимся к этому примеру в главе, посвященной панельным данным.

2. *Увеличится ли уровень инфляции в результате ускорения темпов роста денежной массы? Если да, то на сколько процентных пунктов? Произойдет ли это сразу или через некоторое время? Через какое?* Макроэкономическая теория подсказывает нам, что когда центральный банк наращивает количество денег в экономике, общий уровень цен должен расти. Проверка этой гипотезы — достаточно трудная задача, потому что на уровень инфляции в стране оказывают влияние не только решения центрального банка, но и многие другие факторы: скажем, изменения тарифов на услуги естественных монополий (на перевозки грузов по железной дороге или на жилищно-коммунальное хозяйство). Тем не менее, в рамках эконометрики временных рядов возможно не только выявить эффект воздействия изменений денежной массы на инфляцию, но и понять его распределение во времени. Например, выяснить, как сильно изменится общий уровень цен через три месяца, если увеличить предложение денег сегодня.

Отдельно отметим, что ответ на этот вопрос интересен не только государству, но и частному сектору. Например, коммерческим банкам, которым для назначения оптимальных процентных ставок нужно понимать, как будущая инфляция среагирует на сегодняшние действия центрального банка.

3. *Влияет ли образование индивида на уровень его доходов?* Сложность ответа тут в том, что обычный подсчет средних уровней дохода более и менее образованных людей вряд ли даст корректный результат. Образование обычно коррелировано с ненаблюдаемыми характеристиками (например, с уровнем таланта, интеллекта и мотивации), которые тоже влияют на заработную плату индивида. Например, более талантливым людям легче поступить в университет. Поэтому они чаще это делают.

Таким образом, может оказаться, что более образованные люди получают более высокую зарплату *не потому, что они более образованные, а потому, что они более талантливые*.

В результате сравнение средних приведет к преувеличенной оценке эффекта от образования. Такую ситуацию называют смещением из-за самоотбора (selection bias).

Поскольку важные факторы оказываются ненаблюдаемыми (получить надёжный измеритель уровня таланта очень трудно), чтобы отделить эффект их влияния от эффекта самого образования, приходится использовать довольно тонкие методы, которые мы обсудим в главе, посвященной инструментальным переменным.

Прогнозирование

Речь идет о прогнозировании/предсказании значений тех или иных переменных, и примеры тут привести проще всего.

- Сколько будет стоить однокомнатная квартира с заданными характеристиками на вторичном рынке недвижимости Москвы через полгода?
- С какой вероятностью в следующем году в России начнется экономический спад?
- Если выдать кредит этому клиенту с известными характеристиками, вернёт ли он его в будущем или нет?

Все эти вопросы в том или ином смысле являются вопросами о прогнозировании, и нет сомнений в практической пользе от умения отвечать на них.

Лирическое отступление о моих личных наблюдениях по поводу востребованности эконометрики

Вокруг экономического факультета МГУ, где я со своими коллегами преподаю эконометрику, сформировалось значительное сообщество выпускников, для которых применение эконометрических методов стало основой профессии, источником вдохновения (да и денег тоже).

Они используют эконометрику и другие продвинутые методы анализа данных в совершенно разных областях:

1. В корпоративном секторе. Например, в Яндексе, ВТБ Капитале, в ведущих российских операторах мобильной связи и многих других компаниях, принимающих решения на основе работы с данными.
2. В Центральном банке Российской Федерации, Министерстве экономического развития и других подобных государственных структурах².
3. Наконец, в науке — в ведущих российских университетах и исследовательских центрах, а также в лучших университетах по всему миру, например, в Принстоне, Чикаго и Мадриде.

Истории всех этих людей заставляют меня верить в то, что инвестиции времени и сил в изучение эконометрики определенно окупаются, а знание этой науки является важным преимуществом на соответствующем рынке труда.

Можно ли быть успешным экономистом, совсем не зная эконометрики? Конечно, можно. Однако с эконометрикой эта задача становится гораздо более реалистичной.

1. Мы сознательно выбрали пример, в котором причинно-следственная связь может быть направлена в обе стороны: с одной стороны, люди в хорошей спортивной форме могут иметь более высокую производительность на работе, что способствует увеличению зарплаты; с другой стороны, более состоятельные люди имеют больше возможностей для занятий спортом, так как им легче позволить себе необходимую экипировку или абонемент в спортзал. ↵
2. Мой дипломник рассказывал, что директор одного из департаментов минэкономразвития на собеседовании неожиданно стал задавать ему вопросы по поводу свойств эконометрических оценок. Меня, впрочем, это не удивило, так как я знал, что этот директор не только чиновник, но и автор хороших академических статей, использующих эконометрические методы. ↵

1.2. Корреляция и причинно-следственная связь: некоторые подводные камни

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

В этом параграфе мы обсудим несколько важных эконометрических идей. Обойдемся без продвинутых методов (которые нам только предстоит изучить в будущем). Вместо этого ограничимся пока анализом простых таблиц и графиков.

Пример 1.1. Подготовительные курсы

В таблице 1.1 приведены статистические данные, которые характеризуют платные курсы по подготовке к поступлению в магистратуру экономического факультета одного из ведущих российских университетов.

Интуитивно кажется, что ходить на курсы полезно. Однако разглядывание таблицы 1.1 подталкивает нас к противоположному выводу. Курсы выглядят не просто бесполезными, а даже вредными: их посетители пишут экзамен хуже тех, кто готовился к экзамену сам. В таблице приведена информация о 150 абитуриентах, 50 из которых ходили на эти подготовительные курсы, а остальные 100 не ходили, готовясь к поступлению как-то иначе (скорее всего, самостоятельно). Кроме того, можно видеть, что представители первой группы получили в среднем 43 балла за вступительный экзамен (экзамен оценивался по стобалльной шкале). Представители второй группы, в свою очередь, в среднем набрали на этом экзамене 48,5 баллов.

Таблица 1.1. Результаты вступительного экзамена в магистратуру для 150 абитуриентов

	Ходили на курсы	Не ходили на курсы
Средний балл за экзамен	43 балла (50 человек)	48,5 баллов (100 человек)

Примечание: экзамен оценивался по 100-балльной шкале.

Иными словами, между посещением курсов и результатом экзамена наблюдается отрицательная корреляция. Означает ли это, что посещение курсов является *причиной* более скверных результатов экзамена? На самом деле вовсе не обязательно. И если мы немного поразмыслим, откуда возникла такая корреляция, то найдем альтернативное объяснение.

Одна из причин может состоять в том, что на курсы обычно ходят менее подготовленные и, следовательно, менее уверенные в своих силах абитуриенты.

Можем ли мы как-то учесть данный фактор? Один из способов сделать это — разделить всех абитуриентов на две группы: выпускники бакалавриата данного факультета данного университета и выпускники других вузов. Естественно предположить, что выпускники этого университета лучше готовы к экзамену: во-первых, университет, данные по которому мы анализируем, является одним из ведущих учебных заведений; во-вторых, выпускники, закончившие его бакалавриат, лучше представляют требования экзаменаторов, и поэтому лучше готовы к поступлению в его магистратуру.

Подобное разделение учтено в таблице 1.2. Здесь данные про тех же абитуриентов, что и в таблице 1.1, представлены несколько иначе. Из новой таблицы видно, что для каждой из групп *в отдельности* подготовительные курсы полезны. Действительно, бакалавры этого факультета, посетившие курсы, получают в среднем на 5 баллов больше не посетивших (55 баллов вместо 50). А для выпускников других вузов посещение курсов соответствует увеличению результата экзамена аж на 20 баллов (40 баллов вместо 20).

Таблица 1.2. Результаты вступительного экзамена в магистратуру для 150 абитуриентов (с учетом дополнительного фактора)

	Ходили на курсы	Не ходили на курсы
Выпускники бакалавриата данного университета	55 баллов (10 человек)	50 баллов (95 человек)
Выпускники других вузов	40 баллов (40 человек)	20 баллов (5 человек)
Средний балл за экзамен	43 балла (50 человек)	48,5 баллов (100 человек)

Примечание: экзамен оценивался по 100-балльной шкале.

Ещё раз подчеркнем, что это те же самые 150 абитуриентов, что и в таблице 1.1. Если посчитать средние взвешенные по каждому из столбцов, то мы получим числа из этой таблицы (см. нижнюю строчку таблицы 1.2).

Откуда же возникла видимость негативного эффекта от посещения курсов в таблице 1.1? Дело в том, что, как мы и предположили, абитуриенты из «сильной» группы гораздо реже ходят на курсы, чем абитуриенты из «слабой» группы: среди этих групп курсы посетили 10 и 40 человек, соответственно. Поэтому и общий средний балл всех посетителей курсов оказался не слишком высок, несмотря на то, что для каждой отдельной группы абитуриентов курсы были полезны.

Из этой простой истории можно извлечь два важных вывода:

Вывод №1: если вы будете игнорировать существенные переменные, вы получите смещенные результаты.¹

Такой эффект также называют смещением из-за пропуска существенных переменных (omitted-variable bias). Он подробно анализируется в главе 3.

Вывод №2: корреляция — это не то же самое, что причинно-следственная связь. В первой таблице между посещением курсов и результатами отрицательная корреляция, однако на самом деле это ничего не говорит о качестве курсов. Посчитать корреляцию,

как правило, легко. Были бы данные. Выявить причинно-следственную связь — сложно. Пожалуй, это самая сложная (но и самая интересная!) задача в современной эконометрике.

Пример 1.2. Источник роста

Какие факторы определяют рост валового внутреннего продукта (ВВП)? Это важный вопрос, так как ВВП, несмотря на некоторые недостатки, является одним из ключевых показателей состояния экономики страны. Предположим, мы задались этим вопросом применительно к России и, собрав данные, построили график, характеризующий зависимость ВВП от некоторой переменной (см. рис. 1.1).

Каждая точка на рисунке соответствует данным за определенный год. Прямая линия представляет тенденцию, отражающую эту взаимосвязь. На рис. 1.1 также есть уравнение этой прямой и коэффициент R^2 . Подробнее о том, как определять это уравнение и вычислять этот коэффициент, мы обсудим в главе 2. Пока отметим, что значение R^2 , близкое к единице (как это наблюдается в нашем случае) говорит о хорошем соответствии модели данным. Это соответствие видно невооруженным взглядом: на рисунке большему значению фактора N соответствует большее значение ВВП, и зависимость очень близка к линейной. Коэффициент корреляции между двумя рассматриваемыми переменными также близок к единице и составляет примерно 0,96.

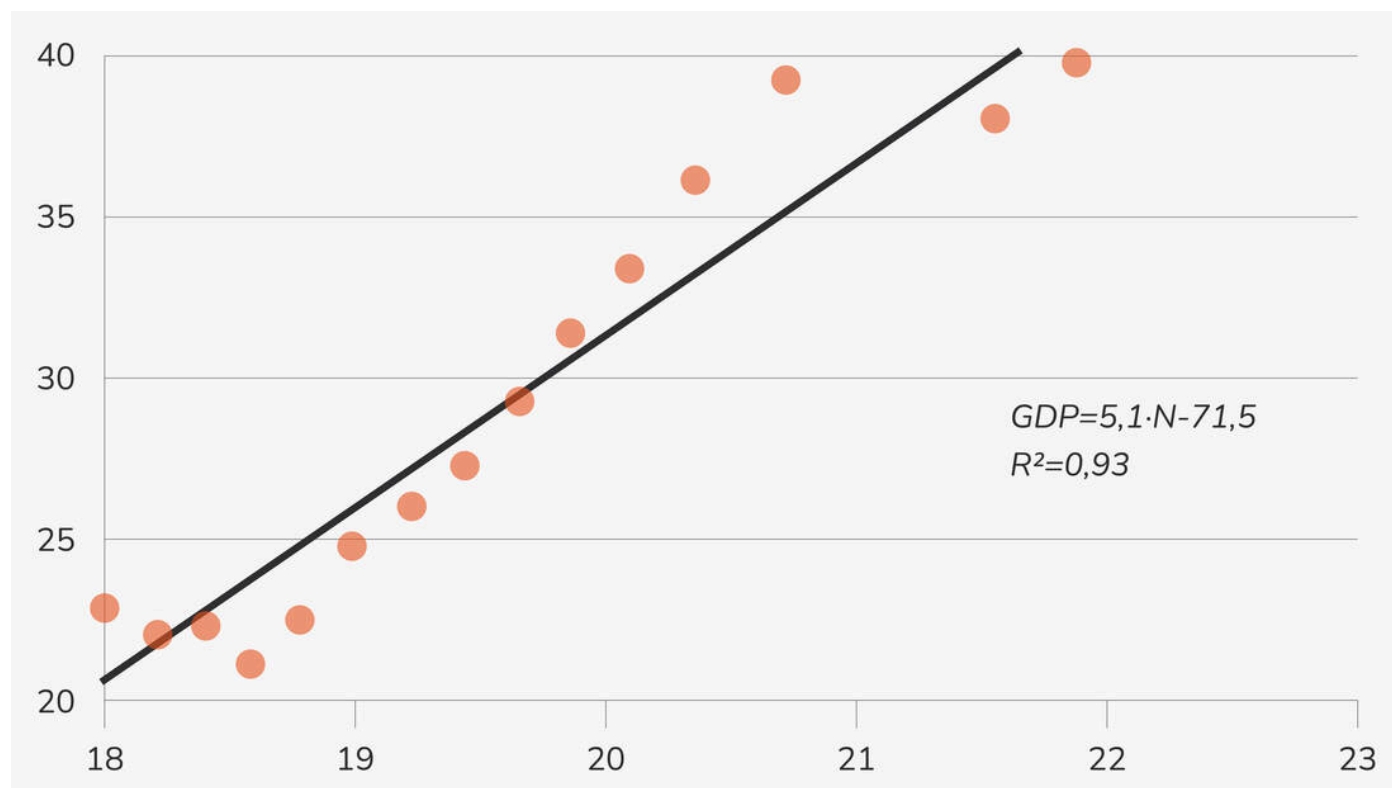


Рисунок 1.1. Зависимость реального ВВП России (вертикальная ось) от некоторого фактора N (горизонтальная ось)

Примечания: Реальный ВВП в ценах 2008 года измерен в трлн руб. Использованы данные за 1995–2012 гг.

После анализа рисунка 1.1 есть соблазн сказать, что мы нашли тот самый важный фактор, который определяет динамику ВВП России. Достаточно увеличить его, чтобы и совокупный выпуск конечных товаров и услуг на территории страны также стал больше. Иногда, читая лекцию, я показываю этот график на экране и спрашиваю слушателей: что это за такой важный для российской экономики фактор N? Самый популярный ответ — цены на нефть — невероятно далек от истины.

В действительности переменная N — это... численность населения Австралии. Здравый смысл подсказывает, что такая переменная вряд ли критична для российского ВВП, а значит, вывод о наличии тесной причинно-следственной связи на основе рисунка 1.1 является ошибочным. В данном случае источником ошибки является одна из типичных ловушек, с которыми сталкиваются начинающие эконометристы при работе с временными рядами — так называемая ложная регрессия (spurious regression).

Чтобы понять источник проблемы, обратимся к рисункам 1.2а и 1.2б, где изображены графики ВВП России и численности населения Австралии по отдельности. Легко видеть, что каждая из этих переменных характеризуется возрастающим трендом. Поэтому, когда мы считаем корреляцию между указанными переменными, технически она оказывается чрезвычайно высокой. Однако в действительности возрастающие тенденции этих переменных определяются совершенно разными факторами и механизмами. А потому существенной причинно-следственной связи между ВВП России и населением Австралии, конечно, нет.

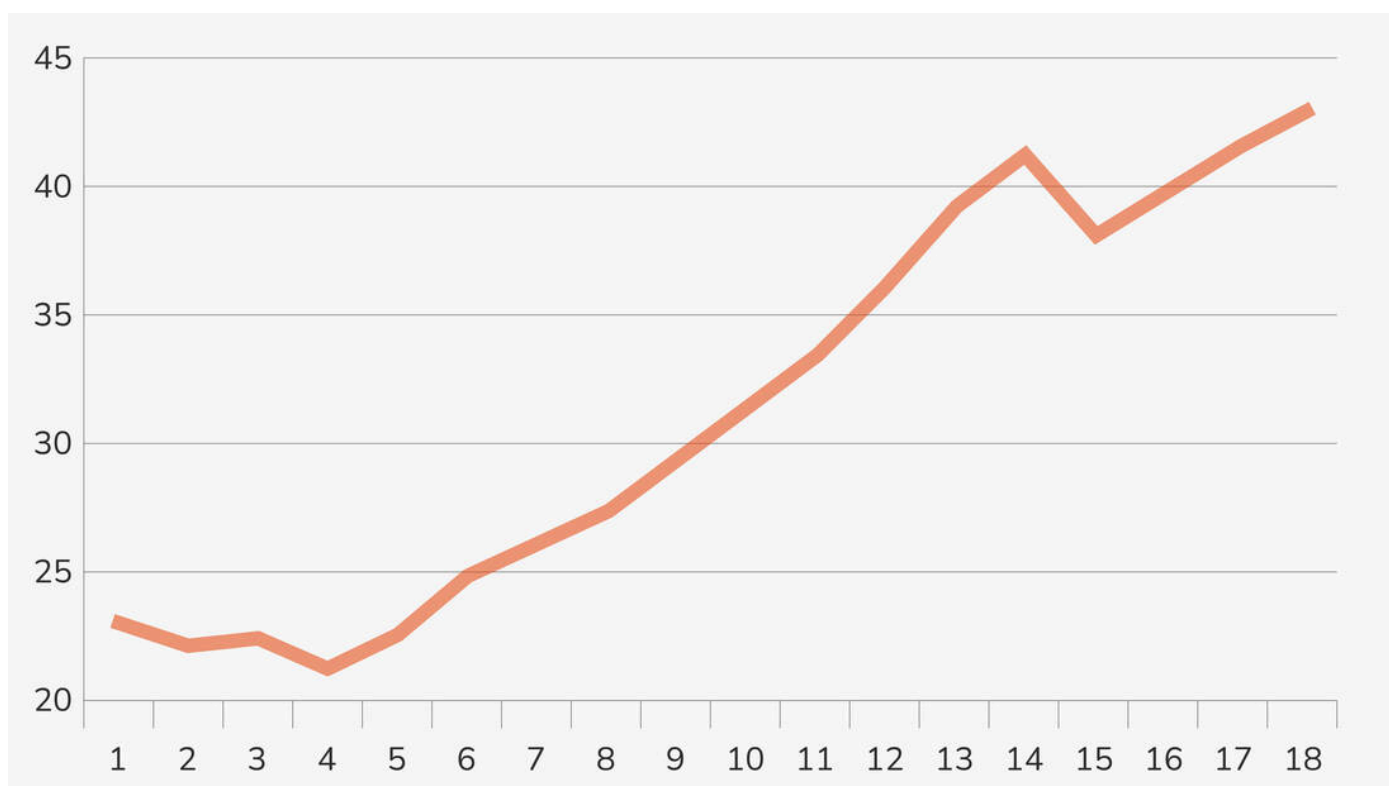


Рисунок 1.2а. Динамика реального ВВП России в 1995–2012 гг.

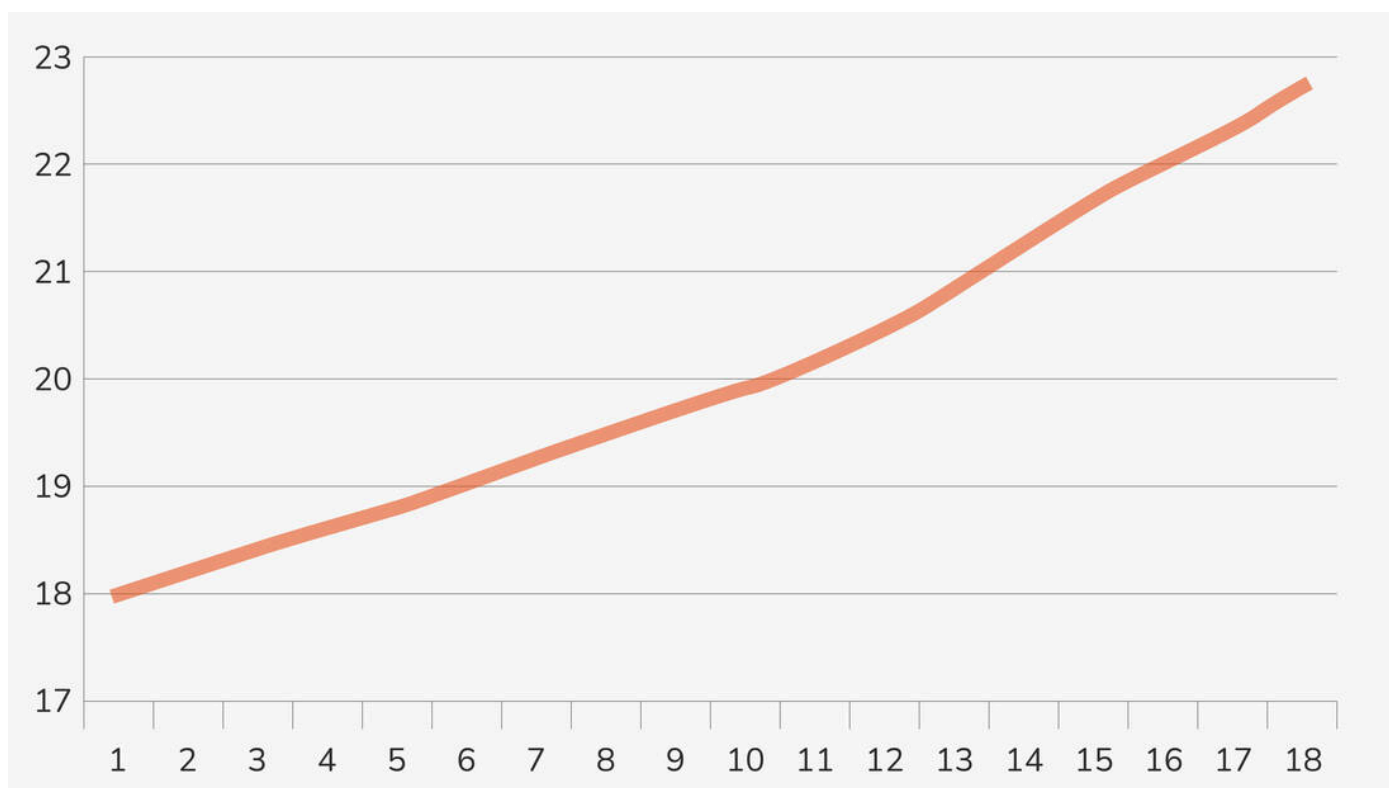


Рисунок 1.2б. Динамика численности населения Австралии в 1995–2012 гг.

Ложная регрессия — ситуация, когда между объясняющей и зависимой переменной в действительности нет причинно-следственной связи, однако коэффициент корреляции между ними по модулю близок к единице, а уравнение, описывающее их взаимосвязь, с высокой точностью соответствует данным. Эта ситуация обычно возникает в случае работы с временными рядами, которые характеризуются наличием тренда, детерминированного или случайного. Эконометристы называют такие временные ряды нестационарными.

Избавиться от возникновения ложной зависимости можно, устранив из данных указанные тренды. Для этого, например, вместо самих переменных можно анализировать их изменения. В нашем случае — изменение ВВП в году t по сравнению с годом $t - 1$ и изменение численности населения Австралии. Такой приём в эконометрике называют переходом к первым разностям

переменных. Результат этого перехода представлен на рис. 1.3. Легко видеть, что в этом случае «злые чары» ложной регрессии рассеиваются, и кажущаяся связь между в действительности не связанными переменными пропадает.

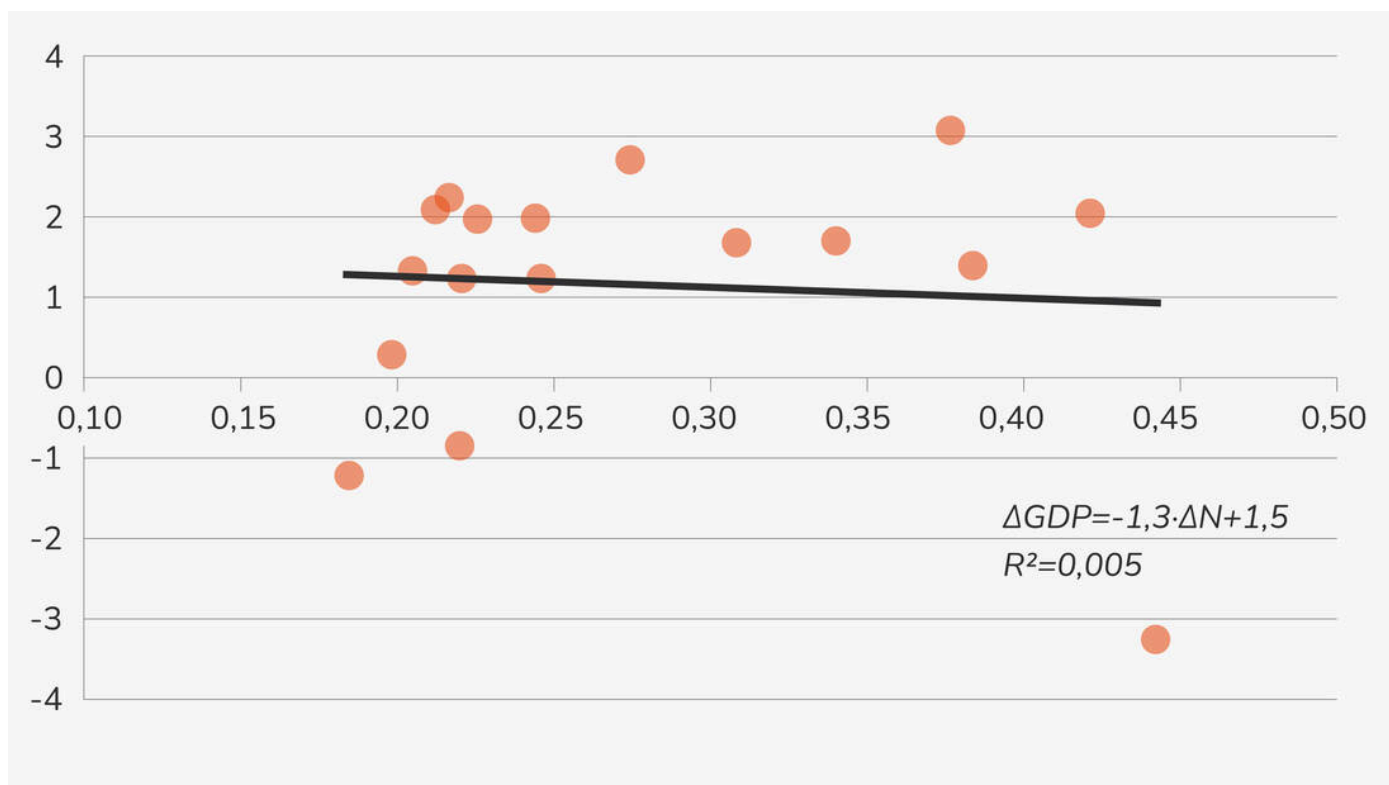


Рисунок 1.3. Изменение реального ВВП России (вертикальная ось) и изменение некоторого фактора N (горизонтальная ось)

Примечание: Можно заметить, что после перехода к изменениям переменных между ними больше не наблюдается явной взаимосвязи (в отличие от ситуации на рис. 1.1.)

Из этой истории также можно извлечь важную мораль.

Вывод №3: если вы будете игнорировать свойства временных рядов, с которыми работаете (наличие трендов и нестационарность), вы получите искаженные результаты.

Много других примеров ложных зависимостей есть тут: <https://tylervigen.com/spurious-correlations>.

Приведенные кейсы охватывают лишь малую долю подводных камней, которые могут возникать в процессе попыток выявить причинно-следственные связи. Например, пока за кадром остался вопрос об определении того, какая из переменных является причиной, а какая — следствием (скажем, влияет ли религиозность нации на её благосостояние, или, наоборот, рост благосостояния является первопричиной изменения популярности религии в обществе?). Эта проблема рассматривается в главе 7.

Но даже те ситуации, которые мы уже обсудили, позволяют проиллюстрировать важность корректного применения методов работы с данными. Неправильное их использование опасно тем, что вместо истинных ответов на исследовательские вопросы вы получите полную чепуху. Именно неверное применение статистических методов, приводящее к сомнительным результатам, сделало популярной присказку про ложь, наглую ложь и статистику². Главы этого учебника призваны помочь вам избежать этой опасности и научиться, используя эконометрику, узнавать о мире нечто ценное.

1. Слово «смещенный» тут можно понимать в математико-статистическом смысле. Напомним, что оценка называется смещенной, если её математическое ожидание не совпадает с истинным значением оцениваемого параметра (в данном случае оцениваемым параметром является изменение балла за экзамен в результате посещения курсов). Более детально формальный смысл термина «смещенная оценка» мы обсудим в главах 2 и 3. [↩](#)
2. «Существуют три вида лжи: ложь, наглая ложь и статистика». Выражение известно благодаря Марку Твену, который приписывал его премьер-министру Великобритании Бенджамину Дизраэли. [↩](#)

1.3. Типы данных, используемых в эконометрике

Один из важных способов классификации данных в эконометрике — это классификация с точки зрения структуры данных. В ней выделяют следующие типы:

- *Пространственные данные (cross section data)*. Пространственными называются данные, собранные о множестве объектов за один момент времени. Например, данные о ценах однокомнатных квартир в Москве в мае 2020 года. Или данные о росте и весе тысячи индивидов по состоянию на 1 сентября 2020 года.
- *Временные ряды (time series)*. Под временным рядом понимаются данные об одном объекте, собранные в течение нескольких последовательных тактов времени. Например, ежедневные данные о курсе доллара, собранные за год. Или данные о росте и весе Ивана Петровича Сидорова, которые собирались 1 числа каждого месяца на протяжении пяти лет.
- *Панельные данные (panel data)*. Панельными называются данные о нескольких объектах, измеренные в течение нескольких тактов времени. Например, ежегодные данные об уровне инфляции в 50 развивающихся странах, собранные за 10 лет. Или данные о росте и весе тысячи индивидов, по каждому из которых доступна информация за 12 месяцев.

В ближайших главах мы сконцентрируемся в основном на пространственных данных, так как они лучше всего подходят для первого знакомства с эконометрикой. В главе 9 мы обсудим преимущества, дополнительные трудности и специальные методы, которые возникают при работе с панельными данными. Подробный анализ временных рядов выходит за рамки первого издания этого учебника, и с ними лучше знакомиться в рамках отдельного продвинутого курса, так как работа с данными подобной структуры имеет много специфических особенностей. Тем не менее, мы будем обращаться к временным рядам с целью иллюстрации некоторых важных идей (как мы уже сделали в примере 1.2).

Альтернативная классификация данных определяется источником их возникновения. По этому критерию выделяют экспериментальные данные и наблюдаемые статистические данные (их ещё называют историческими).

Экспериментальные данные получаются из контролируемых случайных экспериментов (randomized controlled experiments). Наблюдаемые данные, в свою очередь, возникают исторически в течение развития неконтролируемых экспериментатором процессов.

Удобно пояснить различие между двумя этими типами данных, вспомнив пример 1.1 про подготовительные курсы. В нём каждый абитуриент сам решал, посещать ему подготовительные курсы или нет. В итоге на это решение воздействовали разные характеристики абитуриента (скажем, уровень его подготовки), которые также влияли и на результат вступительного экзамена. Это порождало трудности с выявлением истинной пользы от посещения подготовительных курсов в связи с проблемой самоотбора и проблемой пропуска существенных факторов. Возникновение таких трудностей весьма характерно для неэкспериментальных данных.

Представим теперь на минуту, что абитуриенты лишились возможности самостоятельно выбирать, посещать ли им курсы. Теперь мы решаем за каждого абитуриента, будет ли он ходить на курсы. Мы делаем это при помощи специальной лотереи, победители которой в принудительном порядке отправляются на занятия, а всем остальным запрещается это делать. Ясно, что в этом случае индивидуальные характеристики каждого из абитуриентов перестанут влиять на то, попал он на курсы или нет. Следовательно, полезный эффект от посещения курсов будет гораздо проще измерить, так как проблемы, описанные в примере 1.1, автоматически пропадут. Такую лотерею, если она реализована аккуратно, можно считать контролируемым экспериментом, а данные, полученные на основе такого исследования, — экспериментальными.

Второй (более традиционный) пример экспериментальных данных — это медицинские данные, собираемые в ходе тестирования эффективности новых лекарств. Процедура такова: все испытуемые случайным образом разбиваются на две группы, и одной группе выдается новое лекарство, а другой — плацебо.

Из этих примеров следуют два вывода:

Во-первых, ясно, почему эконометристы очень любят работать именно с экспериментальными данными. В этом случае пропадает ряд типичных подводных камней, которые мешают получить корректные результаты (проблема самоотбора, проблема смещения из-за пропуска существенных факторов и другие проблемы, которые обсуждаются в последующих главах учебника). Это позволяет надёжно получать новые знания о мире, используя элементарные методы и модели.

Во-вторых, становится понятным, почему на практике эконометристам гораздо чаще приходится работать с неэкспериментальными историческими данными. Действительно, во многих ситуациях проведение экспериментов либо аморально, либо очень дорого, либо просто невозможно. Скажем, в примере 1.1 история с отбором участников подготовительных курсов по лотерее вряд ли бы вызвала энтузиазм у абитуриентов. А представьте, что вас интересует менее безобидный вопрос: например, влияет ли введение смертной казни на уровень преступности? Маловероятно, что общество благосклонно отнеслось бы к избирательному случайному применению смертной казни в рамках эксперимента.

Всё это приводит к тому, что в большинстве случаев экспериментальные данные остаются для эконометристов недостижимым идеалом. А все продвинутые эконометрические методы направлены на то, чтобы «заставить» исторические данные давать ответы на вопросы так же, как если бы они являлись экспериментальными.

Тем не менее, в некоторых случаях экспериментальные данные эконометристам все-таки доступны. А иногда обстоятельства складываются столь удачно, что и без вмешательства исследователя ситуация оказывается очень похожей на контролируемый эксперимент. Подобное счастливое стечение обстоятельств называется квазиэкспериментом или естественным экспериментом (natural experiment). Про эксперименты и квазиэксперименты мы подробно поговорим в главе 11.

Глава 2. Парная регрессия

В этой главе мы обсудим парную регрессию. Это простой и удобный инструмент, который обычно является отправной точкой в знакомстве с эконометрической наукой. Конечно, кроме преимуществ, у модели парной регрессии есть и существенные ограничения, которые вынудят нас в последующих главах изучить более совершенные методы и модели.

[2.1. Предварительные замечания](#)

[2.2. Метод наименьших квадратов](#)

[2.3. Классическая линейная модель парной регрессии](#)

[2.4. Свойства МНК-оценок](#)

[2.5. Тестирование гипотез и построение доверительных интервалов](#)

[2.6. Прогнозирование](#)

[2.7. Заключение](#)

[Задания для самостоятельного решения](#)

2.1. Предварительные замечания

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

Нам будет удобно использовать следующее определение выборочной ковариации:

$$\widehat{cov}(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}), \text{ где } \bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} \text{ и, аналогично, } \bar{y} = \frac{y_1 + y_2 + \dots + y_n}{n}$$

Здесь и далее значок «крышки» над некоторой величиной будет означать, что эта величина посчитана по выборке. В данном случае, например, $\widehat{cov}(x, y)$ — ковариация между двумя переменными, посчитанная по выборке из n наблюдений. Её не следует путать с теоретической ковариацией между двумя случайными величинами x и y , которую мы будем обозначать $cov(x, y)$ и которая, напомним, определяется так:

$$cov(x, y) = E((x - Ex)(y - Ey)).$$

Принципиальное отличие между теоретической и выборочной ковариацией состоит, в частности, в том, что первая на практике почти никогда не известна и не может быть точно вычислена, в то время как вторая может быть посчитана для каждой конкретной выборки.

Выборочная ковариация обладает рядом удобных свойств, каждое из которых может быть доказано путем непосредственных вычислений:

1. $\widehat{cov}(x, b) = 0$,
2. $\widehat{cov}(x, by) = b * \widehat{cov}(x, y)$,
3. $\widehat{cov}(x, y + b) = \widehat{cov}(x, y)$,
4. $\widehat{cov}(x, y + z) = \widehat{cov}(x, y) + \widehat{cov}(x, z)$.

Здесь b — некоторая константа.

Кроме того, нам будет полезен альтернативный способ вычисления выборочной ковариации: $\widehat{cov}(x, y) = \overline{xy} - \bar{x} * \bar{y}$.

Также нам пригодится выборочная дисперсия переменной, которую мы будем обозначать так:

$$\widehat{var}(x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \overline{x^2} - (\bar{x})^2$$

Свойства выборочной дисперсии, которые нам понадобятся:

1. $\widehat{var}(b) = 0$
2. $\widehat{var}(bx) = b^2 * \widehat{var}(x)$
3. $\widehat{var}(x + b) = \widehat{var}(x)$
4. $\widehat{var}(x + y) = \widehat{var}(x) + \widehat{var}(y) + 2\widehat{cov}(x, y)$

Здесь b — снова некоторая константа.

Наконец, выборочный коэффициент корреляции договоримся обозначать следующим образом:

$$\widehat{corr}(x, y) = \frac{\widehat{Cov}(x, y)}{\sqrt{\widehat{Var}(x) * \widehat{Var}(y)}}$$

2.2. Метод наименьших квадратов

Рассмотрим следующую задачу: у нас есть данные про две переменные, и требуется численно оценить, как одна переменная зависит от другой.

Для удобства введем следующие обозначения: x — объясняющая переменная (её также называют независимой переменной или регрессором), y — объясняемая переменная (её также называют зависимой), n — количество наблюдений, которое имеется в нашем распоряжении (размер выборки). Например, x — площадь однокомнатной квартиры (в квадратных метрах), y — цена квартиры (в миллионах рублей). Понятно, что большие квартиры в среднем стоят дороже, чем маленькие, однако было бы ценно получить конкретное уравнение, которое описывает эту зависимость. Другой пример: x — образование индивида (число лет обучения), y — его ежегодный доход (в рублях). Есть гипотеза, что более образованные индивиды в среднем зарабатывают больше, чем менее образованные, и интересно было бы выяснить, действительно ли это так, и если да, то какую прибавку к доходу дает один дополнительный год обучения.

Тогда формально нашу задачу можно записать так: имеется выборка из n пар наблюдений $(x_i, y_i), i = 1, \dots, n$, требуется подобрать функцию, которая наилучшим образом описывает зависимость переменной y от переменной x .

Парный регрессионный анализ как раз и состоит в решении такой задачи.

Для начала предположим, что искомая функция является линейной. Конечно, это сильное упрощение, поэтому в будущем мы от него откажемся. Мир устроен сложно, и далеко не все зависимости являются линейными.

Про эту задачу удобно думать, глядя на картинку. Если мы каждой паре наблюдений (x_i, y_i) поставим в соответствие точку на плоскости, то у нас получится что-то похожее на рисунок 2.1а. В терминах примера про квартиры можно считать, что каждая точка на этом рисунке соответствует определенной квартире (по оси абсцисс отложена площадь квартиры, а по оси ординат — её цена). Тогда нашу задачу поиска линейной функции, описывающей влияние переменной x на переменную y , можно воспринимать как задачу подбора прямой линии на графике, которая наилучшим образом описывает изображенное на нём облако точек. Пример такой прямой изображен на рисунке 2.1б. Мы будем называть её линией регрессии.

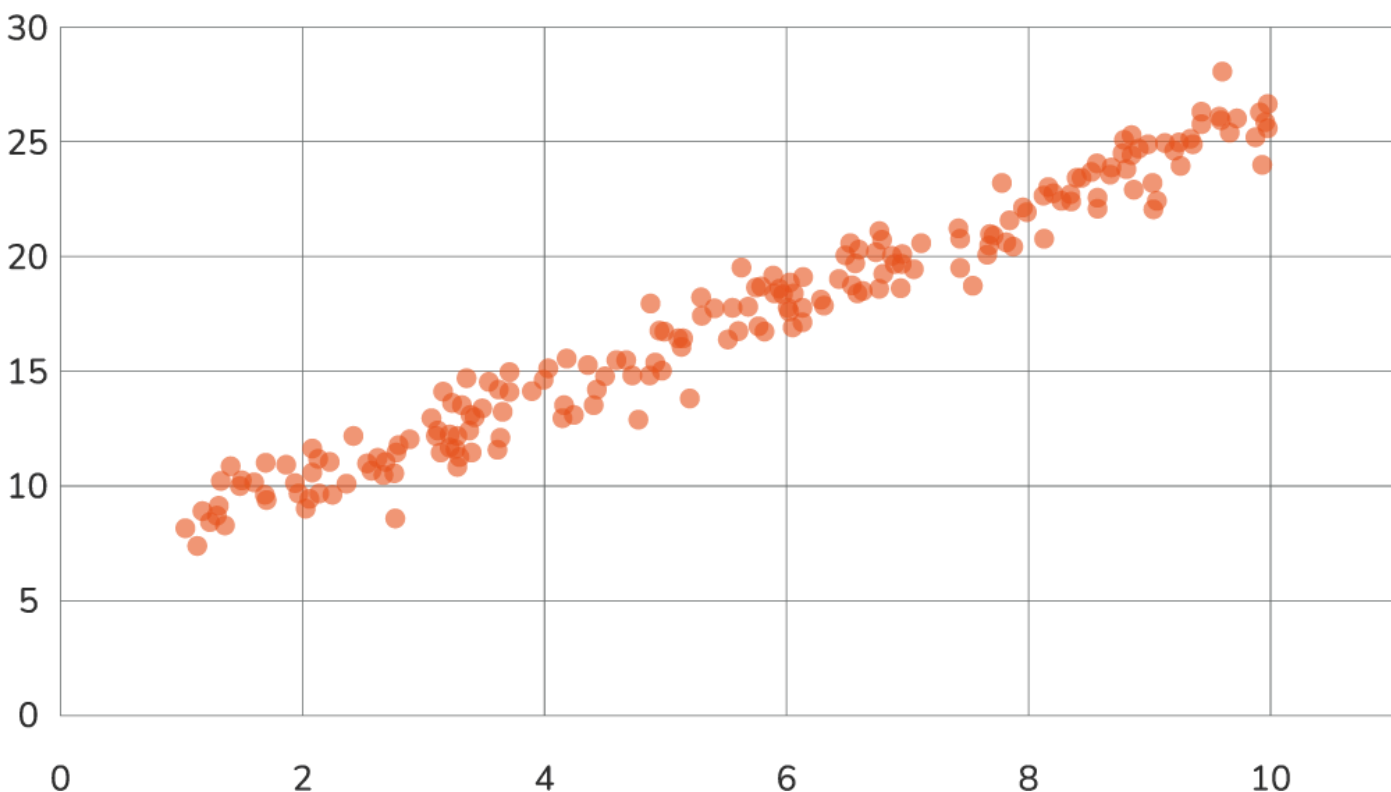


Рисунок 2.1а. Диаграмма рассеяния

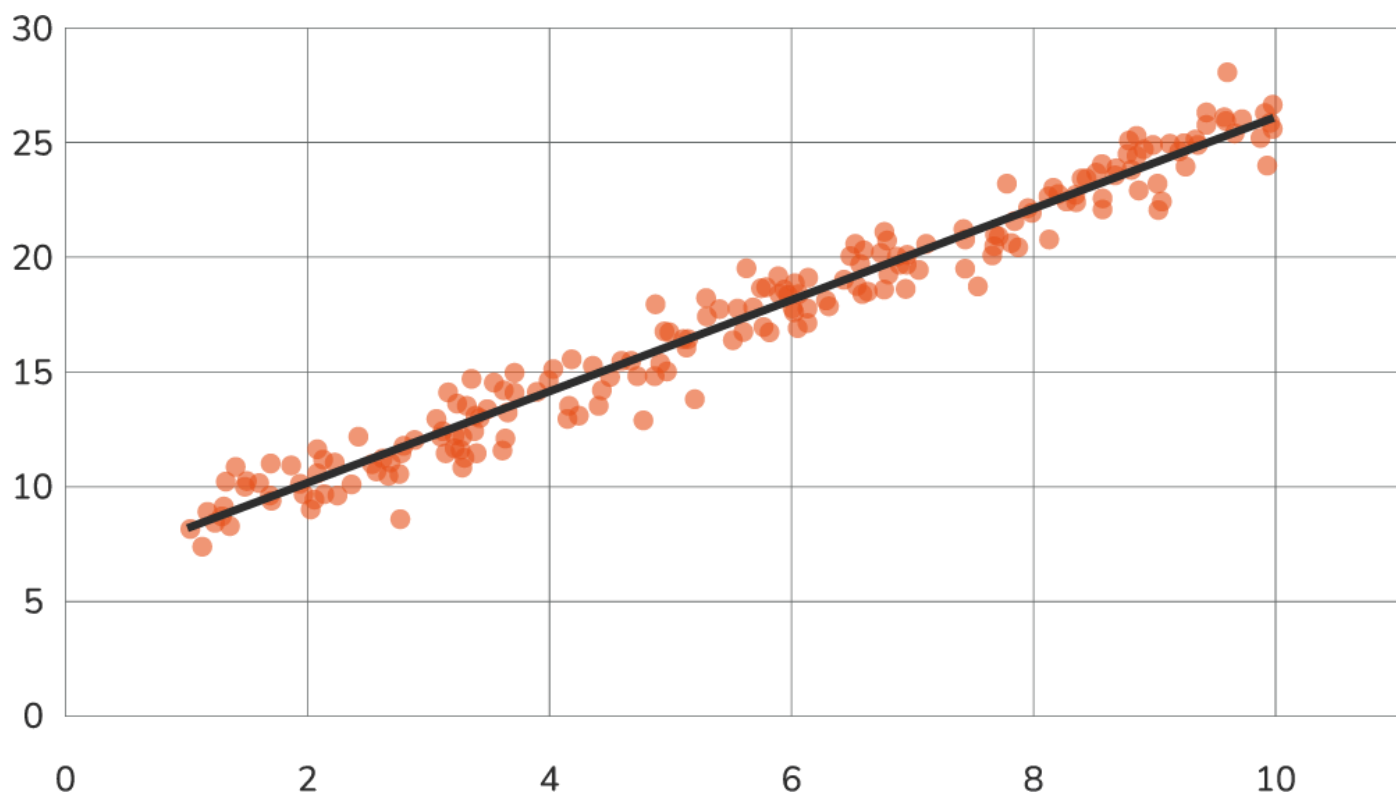


Рисунок 2.16. Диаграмма рассеяния с линией регрессии

Если вы нарисовали картинку, как в нашем примере, то линию регрессии можно провести и «на глазок». Это будет быстро и не вызовет трудностей. Однако в большинстве случаев удобно делать это более технологично: во-первых, чтобы за вас это мог сделать компьютер, во-вторых, чтобы вы могли быть уверены в том, что, если другой исследователь воспользуется теми же самыми данными, то у него получится в точности та же самая линия (с рисованием «на глазок» это требование, очевидно, не выполняется). Давайте разберемся, как этого можно достичь.

Чтобы получить уравнение прямой, нам нужно подобрать численные значения двух коэффициентов: свободного члена, который мы обозначим $\hat{\beta}_1$, и коэффициента наклона $\hat{\beta}_2$. Тогда нашу линию регрессии можно записать так:

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i.$$

Здесь $\hat{y}_i$ — предсказываемое нашим уравнением значение переменной y . Поскольку наша линия лишь приближенно описывает облако точек, предсказанное значение $\hat{y}_i$ может не совпадать с фактическим значением y_i (то есть с тем значением, которое действительно наблюдается в данных). Отклонения предсказанных значений от фактических будем называть остатками регрессии и обозначать $e_i = y_i - \hat{y}_i$. Геометрически остатки регрессии характеризуют отклонение соответствующих наблюдений от линии регрессии по вертикали. Положительным остаткам соответствует отклонение вверх, а отрицательным — вниз.

Естественно подбирать уравнение таким образом, чтобы отклонения предсказаний от фактических значений были не слишком большими, то есть чтобы остатки e_i были маленькими. На первый взгляд кажется, что хорошей идеей в этом случае будет идея подбирать $\hat{\beta}_1$ и $\hat{\beta}_2$ таким образом, чтобы сумма остатков $e_1 + e_2 + \dots + e_n$ была как можно ближе к нулю. Но ограничиваться только этим подходом не стоит. Проблема в том, что в этом случае возможна ситуация, когда есть очень большие положительные отклонения и очень большие по абсолютному значению отрицательные отклонения (то есть наше уравнение сильно «ошибается» в обе стороны), однако эти положительные и отрицательные отклонения взаимно компенсируют друг друга так, что сумма остатков равна нулю.

Поэтому обычно вместо простой суммы остатков минимизируют сумму квадратов остатков, то есть решают такую задачу:

$$e_1^2 + e_2^2 + \dots + e_n^2 \rightarrow \min_{\hat{\beta}_1, \hat{\beta}_2}$$

Такой метод поиска интересующих нас величин $\hat{\beta}_1$ и $\hat{\beta}_2$ называется методом наименьших квадратов (МНК, ordinary least squares, OLS). Внимательный читатель спросит: почему нужно минимизировать именно сумму квадратов, а не, например, сумму модулей остатков $|e_1| + |e_2| + \dots + |e_n|$? Оказывается, что применение МНК приводит к получению результатов, которые обладают очень хорошими свойствами, часть из которых мы обсудим уже в этой главе (а часть — несколько позже). Поэтому МНК и стал одной из главных «рабочих лошадок» эконометристов.

Решим задачу минимизации суммы квадратов остатков.

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2 \rightarrow \min_{\hat{\beta}_1, \hat{\beta}_2}$$

Возьмем производные по $\hat{\beta}_1$ и $\hat{\beta}_2$ и получим необходимые условия экстремума функции:

$$\begin{cases} -2 \sum_{i=1}^n (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0 \\ -2 \sum_{i=1}^n x_i (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0 \end{cases} \quad (2.1)$$

Поскольку анализируемая нами функция представляет собой положительно определенную квадратичную форму (сумма квадратов уж точно не может быть отрицательной), решение этой системы будет именно точкой минимума.

Раскроем скобки:

$$\begin{cases} \sum_{i=1}^n y_i - n\hat{\beta}_1 - \hat{\beta}_2 \sum_{i=1}^n x_i = 0 \\ \sum_{i=1}^n x_i y_i - \hat{\beta}_1 \sum_{i=1}^n x_i - \hat{\beta}_2 \sum_{i=1}^n x_i^2 = 0 \end{cases}$$

Поделим каждое уравнение системы на n :

$$\begin{cases} \bar{y} - \hat{\beta}_1 - \hat{\beta}_2 \bar{x} = 0 \\ \overline{xy} - \hat{\beta}_1 \bar{x} - \hat{\beta}_2 \overline{x^2} = 0 \end{cases}$$

Наконец, выразим из этой системы $\hat{\beta}_1$ и $\hat{\beta}_2$:

$$\begin{aligned} \hat{\beta}_2 &= \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\widehat{Cov}(x, y)}{\widehat{Var}(x)} \\ \hat{\beta}_1 &= \bar{y} - \hat{\beta}_2 \bar{x} = \bar{y} - \frac{\widehat{Cov}(x, y)}{\widehat{Var}(x)} \bar{x} \end{aligned}$$

Пример 2.1. Имеются следующие данные о переменных x и y .

y	1	3	1	3	7
x	2	2	6	4	6

Найдите МНК-оценки параметров в линейной регрессии y на x .

Решение:

Конечно, на практике МНК-оценки всегда вычисляются при помощи компьютера, но для того чтобы разобраться, как это работает, полезно разок-другой проделать это вручную. Для этого удобно организовать промежуточные вычисления в виде такой таблицы:

	y	x	x^2	$x*y$
	1	2	4	2
	3	2	4	6
	1	6	36	6
	3	4	16	12
	7	6	36	42
Сумма	15	20	96	68
Среднее	3	4	19,2	13,6

Теперь можно посчитать МНК-оценки:

$$\begin{aligned} \hat{\beta}_2 &= \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{13,6 - 4*3}{19,2 - 4^2} = 0,5 \\ \hat{\beta}_1 &= \bar{y} - \hat{\beta}_2 \bar{x} = 3 - 4*0,5 = 1 \end{aligned}$$

Ответ: $\hat{y}_i = 1 + 0,5x_i$.

Остатки регрессии, получаемые в результате применения МНК, обладают рядом свойств, которые пригодятся нам в будущем:

Свойство №1. $\sum_{i=1}^n e_i = 0$,

Свойство №2. $\sum_{i=1}^n x_i e_i = 0$,

Свойство №3. $\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{y}_i$,

Свойство №4. $\sum_{i=1}^n (\hat{y}_i - \bar{y}) e_i = 0$, или, что то же самое, $\widehat{cov}(\hat{y}, e) = 0$

Для доказательства первых двух свойств достаточно взглянуть на необходимое условие экстремума в нашей задаче (см. систему (2.1)). С учетом того, что $e_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i$, из первого уравнения этой системы следует первое свойство, а из второго — второе.

Третье свойство следует из первого:

$$\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{y}_i$$

$$\sum_{i=1}^n (y_i - \hat{y}_i) = 0$$

$$\sum_{i=1}^n e_i = 0$$

Наконец, четвертое свойство следует из первых двух:

$$\begin{aligned} \widehat{cov}(\hat{y}, e) &= \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - \bar{y}) (e_i - \bar{e}) = \\ &= \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - \bar{y}) e_i = \frac{1}{n} \sum_{i=1}^n \hat{y}_i e_i - \frac{1}{n} \bar{y} \sum_{i=1}^n e_i = \\ \frac{1}{n} \sum_{i=1}^n \hat{y}_i e_i - 0 &= \frac{1}{n} \sum_{i=1}^n (\hat{\beta}_1 + \hat{\beta}_2 x_i) e_i = \frac{1}{n} \hat{\beta}_1 \sum_{i=1}^n e_i + \frac{1}{n} \hat{\beta}_2 \sum_{i=1}^n x_i e_i = \\ &= \frac{1}{n} \hat{\beta}_1 0 + \frac{1}{n} \hat{\beta}_2 0 = 0 \end{aligned}$$

Мы научились при помощи МНК находить функцию, описывающую влияние переменной x на переменную y . Следующий важный шаг — научиться измерять то, насколько полученное нами уравнение хорошо соответствует данным. Например, глядя на рисунки 2.2, легко видеть, что в первом случае (рисунок 2.2а) наше уравнение очень хорошо соответствует данным, так как все фактические значения зависимой переменной оказываются очень близки к предсказанным. Иными словами, все точки лежат близко к нашей линии регрессии. Однако на рисунке 2.2б точно такая же линия регрессии соответствует данным довольно скверно, так как мы видим, что многие точки лежат от неё очень далеко. Зная только значения $\hat{\beta}_1$ и $\hat{\beta}_2$, оценить степень этого (не)соответствия не получится, поэтому нам понадобится специальный измеритель.

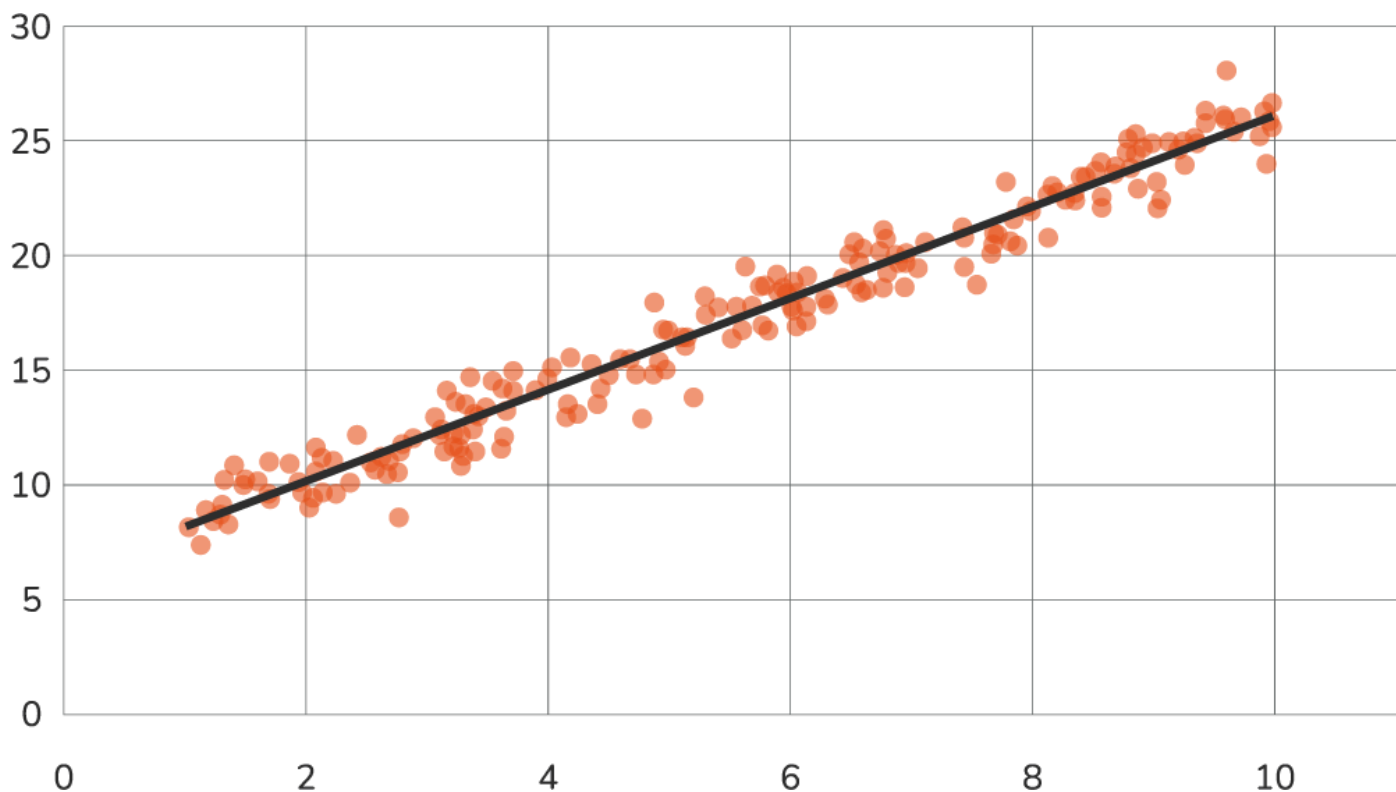


Рисунок 2.2а. Хорошее соответствие линии регрессии данным

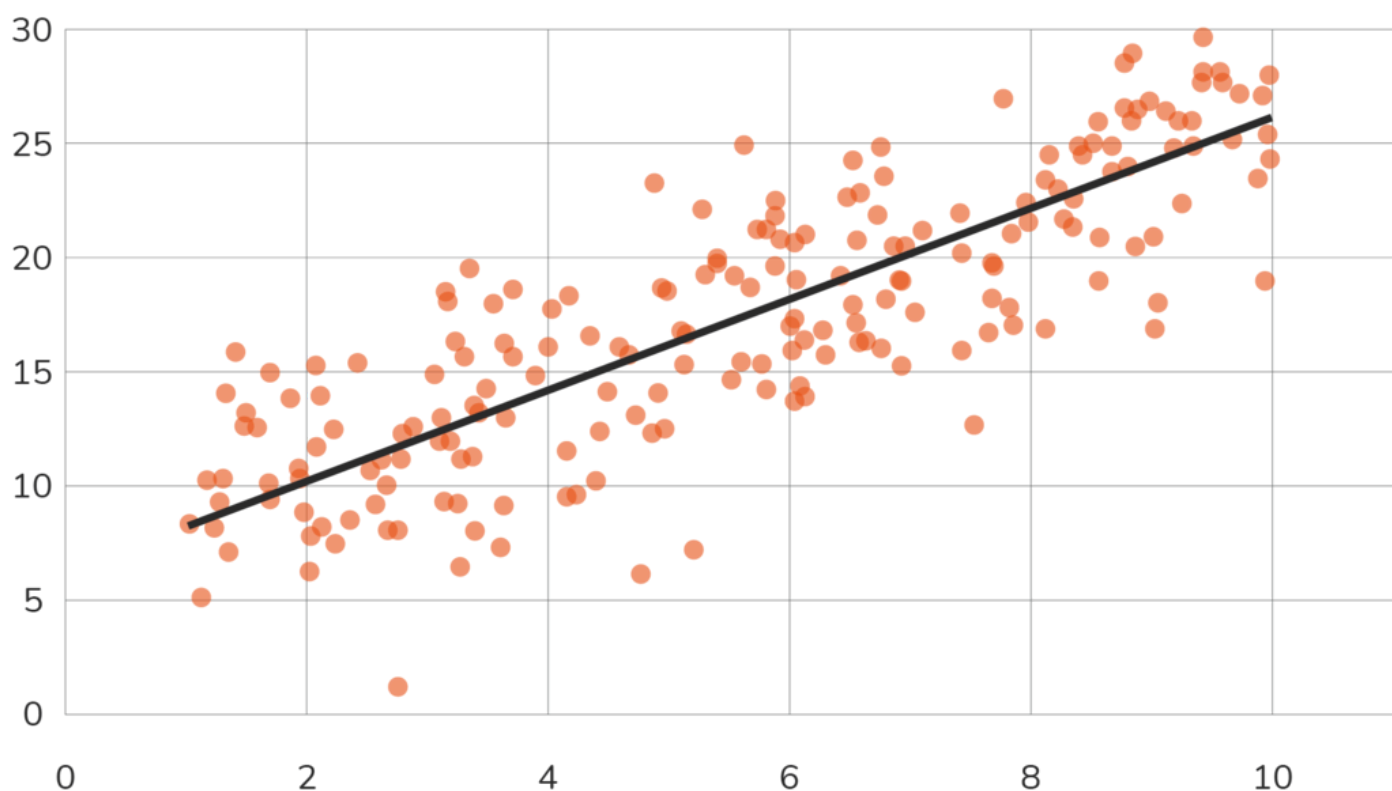


Рисунок 2.26. Плохое соответствие линии регрессии данным

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

Чтобы его получить, выразим переменную y через остатки и $\hat{y}$:

$$e_i = y_i - \hat{y}_i \Rightarrow y_i = e_i + \hat{y}_i$$

Теперь посчитаем выборочную дисперсию этой переменной, используя стандартные свойства выборочной дисперсии:

$$\begin{aligned} \widehat{\text{var}}(y) &= \widehat{\text{var}}(e + \hat{y}) = \widehat{\text{var}}(e) + \widehat{\text{var}}(\hat{y}) + 2 \cdot \widehat{\text{cov}}(e, \hat{y}) = \\ &= \widehat{\text{var}}(e) + \widehat{\text{var}}(\hat{y}) + 2 \cdot 0 = \widehat{\text{var}}(e) + \widehat{\text{var}}(\hat{y}) \end{aligned}$$

В последней строчке мы воспользовались свойством остатков №4 ($\text{Cov}(\hat{y}, e) = 0$). Таким образом, мы доказали, что:

$$\widehat{\text{var}}(y) = \widehat{\text{var}}(e) + \widehat{\text{var}}(\hat{y})$$

Или, что то же самое:

$$\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n} \sum_{i=1}^n (e_i - \bar{e})^2 + \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

Если домножить правую и левую часть данного равенства на n , а также вспомнить, что среднее значение остатков равно нулю (так как по свойству №1 сумма остатков равна нулю), то мы получим следующее тождество:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n e_i^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

Иными словами, верно такое выражение: общая сумма квадратов = сумма квадратов остатков + объясненная сумма квадратов.

Отметим, что если наше уравнение очень хорошо соответствует данным (то есть если наблюдается картинка, похожая на рисунок 2.2а), то сумма квадратов остатков в модели должна быть очень маленькой (так как каждый из остатков регрессии близок к нулю), и, следовательно, величина $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ должна быть очень близка к $\sum_{i=1}^n (y_i - \bar{y})^2$. Иными словами, если наше уравнение хорошо соответствует данным, то дробь $\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$ должна быть близка к единице. В крайнем случае, когда уравнение идеально соответствует данным, и сумма квадратов остатков равна нулю, эта дробь будет в точности равна единице. Напротив, если линия регрессии плохо описывает фактические данные (то есть если наблюдается картинка, похожая на рисунок 2.2б), то сумма квадратов остатков $\sum_{i=1}^n e_i^2$ будет большой. Тогда выражение $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$, наоборот, будет сравнительно

маленьким, и, следовательно, дробь $\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$ будет близка к нулю.

На этой идее основывается использование коэффициента R-квадрат (его ещё иногда называют коэффициентом детерминации).

$$R^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n e_i^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2} =$$

$$= \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\widehat{\text{var}}(\hat{y})}{\widehat{\text{var}}(y)}$$

В силу соображений, которые мы обсудили выше, $R^2 \in [0, 1]$, причем чем лучше наша линия регрессии соответствует данным, тем ближе этот коэффициент к единице. И наоборот, чем хуже наше уравнение согласуется с фактическими наблюдениями, тем ближе он к нулю.

Лирическое отступление о разнообразии обозначений

Для общей суммы квадратов в литературе обычно используется обозначение TSS (total sum of squares). Сумма квадратов остатков в некоторых источниках (Dougherty, 2011) обозначается RSS (residual sum of squares), в некоторых (Сток, Уотсон, 2015) она называется SSR (sum of squared residuals), и, наконец, в (Магнус, Катышев, Пересецкий, 2004) используется обозначение ESS (error sum of squares).

Объясненная сумма квадратов в некоторых учебниках (Dougherty, 2011) тоже обозначается ESS (estimated sum of squares), а в некоторых (Магнус, Катышев, Пересецкий, 2004) — наоборот, RSS (regression sum of squares).

Поэтому мы советуем читателю быть внимательным при работе с разными учебниками. Мы же, чтобы избежать путаницы, как правило, вместо аббревиатур будем использовать соответствующие формулы в явном виде или их расшифровки словами.

Пример 2.2 (продолжение примера 2.1) Для полученного в примере 2.1 уравнения вычислите сумму квадратов остатков и R^2 .

Решение:

Для вычисления суммы квадратов остатков удобно дополнить нашу таблицу еще несколькими столбцами. В этой таблице $\hat{y}_i$ вычисляется по формуле, которую мы нашли в примере 2.1, а остатки регрессии вычисляются по определению: $e_i = y_i - \hat{y}_i$.

	y	x	x^2	$x \cdot y$	$\hat{y}$	e	e^2
	1	2	4	2	2	-1	1
	3	2	4	6	2	1	1
	1	6	36	6	4	-3	9
	3	4	16	12	3	0	0
	7	6	36	42	4	3	9
Сумма	15	20	96	68	15	0	20
Среднее	3	4	19,2	13,6			

Таким образом, сумма квадратов остатков равна 20. Обратите внимание, что сумму предсказанных значений зависимой переменной $\hat{y}_i$ и сумму остатков (без квадратов) вычислять для решения задания было не обязательно. Мы сделали это для проверки. Если все вычисления верны, то в нашей регрессии сумма $\sum \hat{y}_i$ должна совпадать с суммой $\sum y_i$, а сумма остатков всегда должна быть равна нулю. В нашем случае так и получается.

Коэффициент детерминации R^2 можно вычислять разными способами. Поскольку у нас уже посчитана сумма квадратов остатков, удобно воспользоваться той формулой для R^2 , где эта сумма используется:

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2} =$$

$$1 - \frac{20}{(1-3)^2 + (3-3)^2 + (1-3)^2 + (3-3)^2 + (7-3)^2} = \frac{1}{6}$$

Ответ: $\sum e_i^2 = 20$, $R^2 = \frac{1}{6}$.

2.3. Классическая линейная модель парной регрессии

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

В предыдущем параграфе мы научились описывать облако точек некоторой прямой, не делая никаких предположений по поводу природы анализируемых данных. Иными словами, мы не предполагали никакой конкретной модели, описывающей процесс порождения наших данных. Для дальнейшего продвижения, однако, это будет нам необходимо.

Вернемся к нашему примеру с квартирами. Естественно ожидать, что на цену квартиры (y) влияет её площадь (x), а также прочие факторы, например, удаленность квартиры от метро, этаж, наличие балкона и так далее. Обозначим эти прочие факторы переменной ε . С учетом этих соображений естественно предположить следующую модель цены квартиры:

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i, i = 1, 2, \dots, n$$

Здесь x_i — площадь i -ой квартиры в квадратных метрах, y_i — цена i -ой квартиры в миллионах рублей, ε_i — прочие факторы, которые оказывают влияние на цену квартиры y_i . Переменную ε принято называть случайной ошибкой модели. Буквой n будем обозначать число наблюдений в доступной нам выборке.

В целом такая модель выглядит достаточно разумно. Если бы мы знали точные значения коэффициентов β_1 и β_2 , мы могли бы использовать её в практических целях. Например, зная, что $\beta_2 = 0,3$, строительная компания могла бы учитывать при планировании продаж, что один дополнительный квадратный метр площади квартиры оценивается рынком в 0,3 млн рублей. К сожалению, на практике значения параметров β_1 и β_2 нам не известны, зато мы можем собрать статистические данные и получить их (приблизительные) оценки.

Здесь уместно подчеркнуть важное различие между

- параметрами β_k (без «крышек») в выражении $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$,
- и их оценками $\hat{\beta}_k$ («крышками») в выражении $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$.

Это различие состоит в том, что β_1 и β_2 — это некоторые истинные значения параметров модели, которые на практике никогда не известны исследователю. Все, что исследователь в силах сделать — собрать данные и эти значения оценить приближенно. $\hat{\beta}_1$ и $\hat{\beta}_2$ — это оценки истинных значений, которые мы получаем, используя наши выборочные данные. Так как $\hat{\beta}_1$ и $\hat{\beta}_2$ рассчитываются на основе случайной выборки, то они являются случайными величинами.

Естественно, мы хотим, чтобы оценки $\hat{\beta}_1$ и $\hat{\beta}_2$ были близки к истинным значениям оцениваемых параметров. Поэтому нам важно знать: при каких условиях мы можем доверять этим оценкам, то есть рассчитывать на то, что результат использования МНК будет близок к истине? Эти условия называют предпосылками классической линейной модели парной регрессии.

Предпосылки классической линейной модели парной регрессии (КЛМНР):

1. Модель линейна по параметрам и корректно специфицирована
 $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i, i = 1, 2, \dots, n$.
2. $x_1, x_2, \dots, x_n$ — детерминированные (неслучайные) величины, не все одинаковые.
3. Математическое ожидание случайных ошибок равно нулю
 $E\varepsilon_i = 0$.
4. Дисперсия случайной ошибки одинакова для всех наблюдений $var(\varepsilon_i) = \sigma^2$.
5. Случайные ошибки, относящиеся к разным наблюдениям, взаимно независимы.
6. Случайные ошибки имеют нормальное распределение $\varepsilon_i \sim N(0, \sigma^2)$.

Мы обсудили выше соображения, исходя из которых может быть сформулирована предпосылка №1. Как мы увидим в дальнейшем, правильная спецификация подразумевает в первую очередь отсутствие среди прочих факторов других переменных, которые одновременно влияют на y и коррелируют с x . Нарушение этого требования приводит к серьезным проблемам, которые мы осветим в конце данной главы.

Предпосылка №2 касается двух важных аспектов. Во-первых, мы предполагаем, что регрессоры x_i являются неслучайными величинами. Это техническое предположение, которое упростит некоторые выкладки в этом разделе. Обратите внимание, что ε_i в отличие от регрессоров являются случайными величинами, а следовательно, и y_i тоже случайны, так как представляют собой сумму неслучайной компоненты $\beta_1 + \beta_2 x_i$ и случайной величины ε_i . В терминах нашего примера с квартирами про эту предпосылку можно думать так: представим, что вы собрали случайную выборку из 100 квартир площадью 30 м², 100 квартир площадью 35 м² и 100 квартир площадью 40 м². Если вы соберете другую выборку из трехсот квартир с такими же площадями, то значения регрессоров x_i останутся теми же самыми, а вот значения объясняемой переменной y_i поменяются, поэтому в данном примере разумно думать про регрессоры как про неслучайные величины, а про величины y_i — как про случайные.

Во-вторых, в рамках предпосылки №2 мы предполагаем, что не все значения регрессоров одинаковы. Нетрудно понять, зачем нужно это предположение, если взглянуть на формулу оценки коэффициента

$\hat{\beta}_2 = \frac{Cov(x, y)}{Var(x)}$. Обратите внимание, что в знаменателе этой формулы стоит выборочная дисперсия переменной x , но если все значения этой переменной в выборке будут одинаковы, то эта дисперсия окажется равной нулю, и из-за этого мы не сможем рассчитать МНК-оценку $\hat{\beta}_2$.

Предпосылка №3 говорит о том, что прочие факторы могут приводить к отклонению y_i от величины $\beta_1 + \beta_2 x_i$ как вверх, так и вниз, но в среднем эти отклонения компенсируют друг друга.

Предпосылка №4 требует, чтобы разброс случайных ошибок в среднем был постоянен для всех наблюдений. Её смысл удобно пояснить, используя картинку. Посмотрите на рисунки 2.3а и 2.3б. В первом случае предпосылка о постоянстве дисперсии случайной ошибки выполнена, а во втором — нет, так как разброс точек вокруг линии регрессии растет по мере увеличения

объясняющей переменной, следовательно, мы можем заключить, что дисперсия случайной ошибки не является одинаковой для всех наблюдений. Ситуация, когда предпосылка №4 выполнена (то есть ситуация, соответствующая рисунку 2.3а) называется гомоскедастичностью случайных ошибок. Альтернативная ситуация называется гетероскедастичностью случайных ошибок.

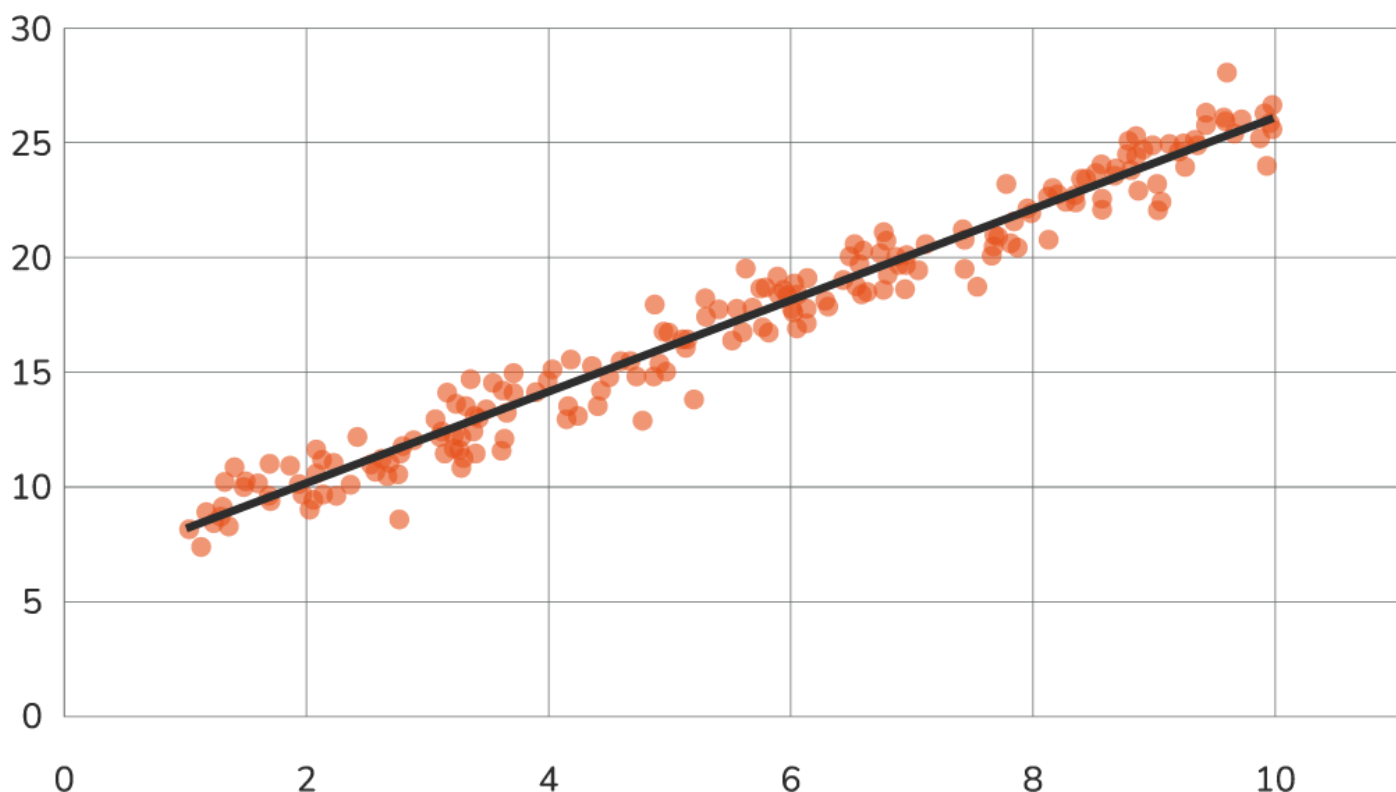


Рисунок 2.3а. Гомоскедастичность случайных ошибок

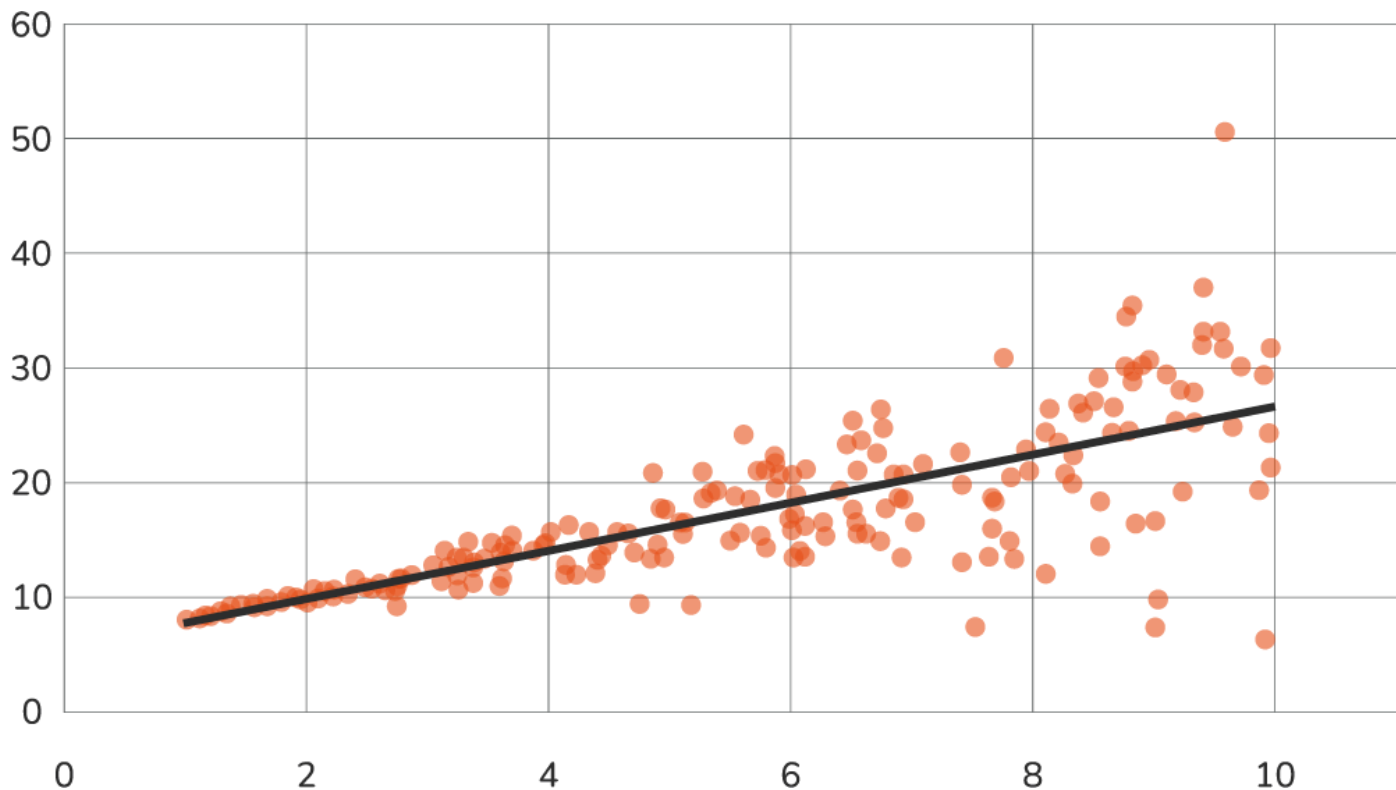


Рисунок 2.3б. Гетероскедастичность случайных ошибок

Из предпосылки №5 следует, что случайные ошибки, относящиеся к разным наблюдениям, не коррелированы друг с другом: $cov(\epsilon_i, \epsilon_j) = 0$ при $i \neq j$.

Предпосылка №6 не требуется для обеспечения хороших свойств оценок коэффициентов (обратите внимание, что ниже, в формулировке теоремы Гаусса — Маркова, она не фигурирует), однако будет полезна для тестирования гипотез и построения доверительных интервалов.

Теорема Гаусса — Маркова. Если выполнены предпосылки 1-5 классической линейной модели парной регрессии, то МНК-оценки коэффициентов $\widehat{\beta}_1$ и $\widehat{\beta}_2$ будут:

- (а) несмещенными,
- (б) эффективными в классе всех несмещенных и линейных по y оценок¹.

Напомним, что оценка называется несмещенной, если её математическое ожидание совпадает с истинным значением оцениваемого параметра: $E\widehat{\beta}_2 = \beta_2$. Свойство эффективности означает, что оценка характеризуется минимальной дисперсией среди всех альтернативных оценок в данном классе, то есть является «наиболее точной» оценкой интересующего нас параметра. Линейность по y означает, что мы рассматриваем все оценки, которые могут быть представлены в виде линейной комбинации значений объясняемой переменной, то есть записаны в виде $\sum_{i=1}^n c_i * y_i$.

Если переформулировать свойства несмещенности и эффективности нестрого, то можно сказать, что при выполнении предпосылок 1-5 МНК-оценки параметров окажутся хорошими: они будут «в среднем правильными» и наиболее точными. Теорема Гаусса — Маркова дает нам важную мотивацию для того, чтобы оценивать параметры нашей модели именно методом наименьших квадратов, а не каким-то альтернативным способом.

Лирическое отступление о предпосылках

Каждый раз, когда я рассказываю студентам об этой теореме, в моей голове разыгрывается примерно такой диалог между двумя эконометристами (назовем их Филипп и Дима).

Дима: Реалистичны ли предпосылки КЛМНР?

Филипп: Не очень. Например, в реальных исследованиях на пространственных данных ты почти всегда будешь сталкиваться с нарушением требования постоянства дисперсии случайной ошибки (нарушением предпосылки №4). Во многих прикладных исследованиях также окажется более целесообразным думать про регрессоры как про случайные, а не детерминированные случайные величины (это отклонение от предпосылки №2). На нормальность случайных ошибок (предпосылка №6) я бы тоже не рассчитывал...

Дима: Зачем же тогда мы её изучаем? Давайте сразу перейдем к более реалистичной модели.

Филипп: Мы начинаем с КЛМНР, так как это самая простая модель, на примере которой мы можем обсудить ряд важных эконометрических идей и при этом не погрязнуть в технических трудностях. В последующих главах мы будем постепенно отказываться от предпосылок КЛМНР и в результате получим набор моделей и методов, которые хорошо подходят для реальных исследований на живых данных. Кроме того, мы научимся проверять выполнение тех или иных предположений КЛМНР, чтобы понять, когда стоит их использовать, а когда — нет.

В частности, последствия нарушения предпосылки №3 читателю предлагается проанализировать уже в этом параграфе, в одном из заданий для самостоятельного решения.

1. Несмещенность и эффективность — это свойства оценок при фиксированном объеме выборки (при фиксированном n). Во многих случаях удобно также использовать асимптотические свойства оценок, то есть свойства, которые имеют место при $n \rightarrow \infty$ (например, состоятельность). Об асимптотических свойствах МНК-оценок мы подробно поговорим в одной из последующих глав. [↩](#)

2.4. Свойства МНК-оценок

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

Этот параграф содержит ряд формальных доказательств. Читатель, не заинтересованный в технических деталях, может сразу перейти к следующему параграфу, где полученные здесь результаты используются для тестирования гипотез и построения доверительных интервалов.

Математическое ожидание и дисперсия МНК-оценок. Прежде чем непосредственно доказать сформулированную теорему, сконцентрируемся на анализе ряда важных свойств МНК-оценок. Для этого нам пригодится следующий факт: $\Sigma(\mathbf{x}_i - \bar{\mathbf{x}}) = \mathbf{0}$. Действительно:

$$\Sigma(\mathbf{x}_i - \bar{\mathbf{x}}) = \Sigma \mathbf{x}_i - \Sigma \bar{\mathbf{x}} = \Sigma \mathbf{x}_i - n\bar{\mathbf{x}} = \Sigma \mathbf{x}_i - n \frac{\Sigma \mathbf{x}_i}{n} = \mathbf{0}$$

Перепишем формулу для оценки $\widehat{\beta}_2$, используя это соображение (мы применяем его в самом последнем переходе):

$$\begin{aligned} \widehat{\beta}_2 &= \frac{\widehat{cov}(\mathbf{x}, \mathbf{y})}{\widehat{var}(\mathbf{x})} = \frac{\frac{1}{n} \Sigma(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{y}_i - \bar{\mathbf{y}})}{\frac{1}{n} \Sigma(\mathbf{x}_i - \bar{\mathbf{x}})^2} = \\ &= \frac{\Sigma(\mathbf{x}_i - \bar{\mathbf{x}})(\beta_1 + \beta_2 \mathbf{x}_i + \varepsilon_i - \beta_1 - \beta_2 \bar{\mathbf{x}} - \bar{\varepsilon})}{\Sigma(\mathbf{x}_i - \bar{\mathbf{x}})^2} = \\ &= \frac{\Sigma(\mathbf{x}_i - \bar{\mathbf{x}})(\beta_2(\mathbf{x}_i - \bar{\mathbf{x}}) + \varepsilon_i - \bar{\varepsilon})}{\Sigma(\mathbf{x}_i - \bar{\mathbf{x}})^2} = \\ &= \frac{\beta_2 \Sigma(\mathbf{x}_i - \bar{\mathbf{x}})^2 + \Sigma(\mathbf{x}_i - \bar{\mathbf{x}})(\varepsilon_i - \bar{\varepsilon})}{\Sigma(\mathbf{x}_i - \bar{\mathbf{x}})^2} = \\ &= \beta_2 + \frac{\Sigma(\mathbf{x}_i - \bar{\mathbf{x}})\varepsilon_i - \bar{\varepsilon} \Sigma(\mathbf{x}_i - \bar{\mathbf{x}})}{\Sigma(\mathbf{x}_i - \bar{\mathbf{x}})^2} = \beta_2 + \frac{\Sigma(\mathbf{x}_i - \bar{\mathbf{x}})\varepsilon_i - \bar{\varepsilon} \cdot \mathbf{0}}{\Sigma(\mathbf{x}_i - \bar{\mathbf{x}})^2} \end{aligned}$$

Таким образом, мы получили следующее представление для МНК-оценки:

$$\widehat{\beta}_2 = \beta_2 + \frac{\Sigma(\mathbf{x}_i - \bar{\mathbf{x}})\varepsilon_i}{\Sigma(\mathbf{x}_i - \bar{\mathbf{x}})^2} \quad (2.2)$$

Еще раз подчеркнем, что МНК-оценка является случайной величиной (хотя для каждой конкретной реализации данных это будет какое-то конкретное число). Полученное представление дает нам возможность удобно исследовать свойства этой случайной величины, в частности, вычислить её математическое ожидание и дисперсию.

Начнем с математического ожидания:

$$E(\widehat{\beta}_2) = E\left(\beta_2 + \frac{\Sigma(\mathbf{x}_i - \bar{\mathbf{x}})\varepsilon_i}{\Sigma(\mathbf{x}_i - \bar{\mathbf{x}})^2}\right)$$

β_2 — это число (неслучайная величина), которое по свойству математического ожидания можно вынести из-под знака математического ожидания. $\mathbf{x}_i - \bar{\mathbf{x}}$ — тоже неслучайные величины (предпосылка №2), с которыми можно сделать то же самое. В результате под знаком математического ожидания осталось только ε_i . Но в силу предпосылки №3 $E\varepsilon_i = 0$, следовательно:

$$E(\widehat{\beta}_2) = \beta_2 + \Sigma(\mathbf{x}_i - \bar{\mathbf{x}}) \frac{E(\varepsilon_i)}{\Sigma(\mathbf{x}_i - \bar{\mathbf{x}})^2} = \beta_2 + \frac{\Sigma(\mathbf{x}_i - \bar{\mathbf{x}})0}{\Sigma(\mathbf{x}_i - \bar{\mathbf{x}})^2} = \beta_2$$

Мы выяснили, что $E(\widehat{\beta}_2) = \beta_2$, то есть доказали несмещенность оценки $\widehat{\beta}_2$. По аналогии можно доказать то, что $\widehat{\beta}_1$ также является несмещенной оценкой. Стоит отметить, что в этих рассуждениях нам потребовались не все предпосылки теоремы Гаусса-Маркова. Были использованы первая предпосылка о спецификации модели, вторая предпосылка о том, что $\mathbf{x}_i$ — детерминированные (неслучайные) величины, и третья предпосылка о математическом ожидании случайной ошибки.

Для вычисления дисперсии $\widehat{\beta}_2$ нам потребуются дополнительно предпосылки №4 и №5. Также придется вспомнить некоторые свойства дисперсии: добавление константы к случайной величине на дисперсию не влияет, значит, справедливо следующее:

$$var(\widehat{\beta}_2) = var\left(\beta_2 + \frac{\Sigma(\mathbf{x}_i - \bar{\mathbf{x}})\varepsilon_i}{\Sigma(\mathbf{x}_i - \bar{\mathbf{x}})^2}\right) = var\left(\frac{\Sigma(\mathbf{x}_i - \bar{\mathbf{x}})\varepsilon_i}{\Sigma(\mathbf{x}_i - \bar{\mathbf{x}})^2}\right)$$

Дисперсия константы, умноженной на случайную величину, равна квадрату константы, умноженному на дисперсию этой случайной величины. Используем это свойство и вынесем знаменатель из-под знака дисперсии как константу:

$$\text{var}(\widehat{\beta}_2) = \frac{1}{\left(\sum (x_i - \bar{x})\right)^2} * \text{var}(\sum (x_i - \bar{x})\varepsilon_i)$$

Остается дисперсия суммы $\sum (x_i - \bar{x})\varepsilon_i$. В случае независимости слагаемых дисперсия суммы равняется сумме дисперсий. В силу пятой предпосылки о независимости случайных ошибок, соответствующих разным наблюдениям, можно утверждать, что слагаемые действительно независимы и поэтому:

$$\text{var}(\widehat{\beta}_2) = \frac{1}{\left(\sum (x_i - \bar{x})\right)^2} * \sum \text{var}((x_i - \bar{x})\varepsilon_i)$$

Следующий шаг — вынести $(x_i - \bar{x})$ из-под знака дисперсии, что законно, так как в силу второй предпосылки x_i — это неслучайные величины:

$$\text{var}(\widehat{\beta}_2) = \frac{1}{\left(\sum (x_i - \bar{x})\right)^2} * \left(\sum (x_i - \bar{x})^2 * \text{var}(\varepsilon_i)\right)$$

Воспользуемся четвертой предпосылкой о том, что для всех наблюдений дисперсия случайной ошибки равна константе σ^2 :

$$\text{var}(\widehat{\beta}_2) = \frac{1}{\left(\sum (x_i - \bar{x})\right)^2} * \left(\sum (x_i - \bar{x})^2 * \sigma^2\right)$$

Величину σ^2 можно вынести за скобки:

$$\text{var}(\widehat{\beta}_2) = \frac{\sum (x_i - \bar{x})^2}{\left(\sum (x_i - \bar{x})\right)^2} * \sigma^2 = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

Таким образом, мы нашли дисперсию МНК-оценки $\widehat{\beta}_2$: $\text{var}(\widehat{\beta}_2) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$ (2.3)

Используя аналогичные рассуждения, можно вычислить математическое ожидание и дисперсию оценки $\widehat{\beta}_1$, а также ковариацию между оценками $\widehat{\beta}_1$ и $\widehat{\beta}_2$:

$$E(\widehat{\beta}_1) = \beta_1$$

$$\text{var}(\widehat{\beta}_1) = \frac{\frac{\sigma^2}{n} * \sum x_i^2}{\sum (x_i - \bar{x})^2} \quad (2.4)$$

$$\text{cov}(\widehat{\beta}_1, \widehat{\beta}_2) = -\bar{x} * \text{var}(\widehat{\beta}_2) = \frac{-\bar{x} * \sigma^2}{\sum (x_i - \bar{x})^2} \quad (2.5)$$

Читателю предлагается сделать это самостоятельно.

Доказательство теоремы Гаусса — Маркова

Докажем теорему для МНК-оценки параметра β_2 . Рассмотрим произвольную линейную по y оценку. Обозначим её $\widetilde{\beta}_2$. В силу линейности:

$$\widetilde{\beta}_2 = \sum_{i=1}^n c_i * y_i$$

Так как эта оценка должна быть несмещенной, то для нее при любых данных должно выполняться следующее условие:

$$E\widetilde{\beta}_2 = \beta_2$$

$$E\left(\sum_{i=1}^n c_i * y_i\right) = \beta_2$$

$$E\left(\sum_{i=1}^n c_i * (\beta_1 + \beta_2 x_i + \varepsilon_i)\right) = \beta_2$$

$$\beta_1 \left(\sum_{i=1}^n c_i\right) + \beta_2 \left(\sum_{i=1}^n c_i * x_i\right) = \beta_2$$

Это равенство верно при любом наборе $x_1, x_2, \dots, x_n$ тогда и только тогда, когда $\sum_{i=1}^n c_i = 0$ и $\sum_{i=1}^n c_i * x_i = 1$.

Теперь сформулируем задачу минимизации дисперсии несмещенной оценки параметра:

$$\text{var}(\widetilde{\beta}_2) \rightarrow \min_{c_1, \dots, c_n}$$

при условиях

$$\sum_{i=1}^n c_i = 0$$

и

$$\sum_{i=1}^n c_i x_i = 1$$

Вспользуемся тем, что

$$\text{var}(\widetilde{\beta}_2) = \text{var}\left(\sum_{i=1}^n c_i (\beta_1 + \beta_2 x_i + \varepsilon_i)\right) = \sigma^2 \sum_{i=1}^n c_i^2$$

Последнее равенство корректно в силу предпосылок №4 и №5 КЛМНР. Так как мы имеем дело с задачей на условный экстремум, то мы можем выписать соответствующую функцию Лагранжа:

$$L = \sum_{i=1}^n c_i^2 + \lambda \left(1 - \sum_{i=1}^n c_i x_i\right) + \mu \left(0 - \sum_{i=1}^n c_i\right)$$

Возьмем частные производные по λ, μ и $c_i, i = 1, 2, \dots, n$. И приравняем их к нулю. Получим необходимые условия экстремума (так как мы минимизируем положительно определенную квадратичную форму при линейных ограничениях, то это условие будет и достаточным условием минимума):

$$\begin{cases} \sum_{i=1}^n c_i x_i = 1, \\ \sum_{i=1}^n c_i = 0, \\ 2c_i - \lambda x_i - \mu = 0, i = 1, 2, \dots, n. \end{cases}$$

Решая эту систему, находим: $c_i = \frac{x_i - \bar{x}}{\sum_{j=1}^n (x_j - \bar{x})^2}$.

Следовательно:

$$\widetilde{\beta}_2 = \sum_{i=1}^n c_i y_i = \sum_{i=1}^n \frac{x_i - \bar{x}}{\sum_{j=1}^n (x_j - \bar{x})^2} y_i$$

Осталось заметить, что полученная нами оценка — это и есть МНК-оценка, записанная немного другим способом. Действительно:

$$\begin{aligned} \widetilde{\beta}_2 &= \sum_{i=1}^n \frac{x_i - \bar{x}}{\sum_{j=1}^n (x_j - \bar{x})^2} y_i = \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{j=1}^n (x_j - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i - 0}{\sum_{j=1}^n (x_j - \bar{x})^2} = \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i - \bar{y} \sum_{i=1}^n (x_i - \bar{x})}{\sum_{j=1}^n (x_j - \bar{x})^2} = \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x}) (y_i - \bar{y})}{\sum_{j=1}^n (x_j - \bar{x})^2} = \\ &= \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) (y_i - \bar{y})}{\frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})^2} = \frac{\widehat{cov}(x, y)}{\widehat{var}(x)} \end{aligned}$$

Таким образом, мы доказали, что решение задачи минимизации дисперсии несмещенной оценки параметра приводит нас в точности к МНК-оценке. Следовательно, МНК-оценка является эффективной и несмещенной, что и требовалось доказать.

Для МНК-оценки параметра β_1 доказательство может быть осуществлено аналогичным образом.

2.5. Тестирование гипотез и построение доверительных интервалов

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

В предыдущем параграфе мы выяснили, что дисперсия оценки $\widehat{\beta}_2$ равна:

$$\text{var}(\widehat{\beta}_2) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}.$$

Это полезная информация, так как дисперсия $\widehat{\beta}_2$ характеризует точность результатов оценивания соответствующего параметра (чем меньше дисперсия, тем точнее наша оценка). Проблема в том, что непосредственно величину $\text{var}(\widehat{\beta}_2)$ мы вычислить не можем: хотя мы наблюдаем значения $x_i, i = 1, 2, \dots, n$, но мы не наблюдаем величину σ^2 . Этот параметр является неизвестным параметром классической линейной модели подобно величинам β_1 и β_2 . Впрочем, как и в случае с β_1 и β_2 , мы можем получить оценку неизвестного параметра σ^2 . Несмещенная оценка дисперсии случайной ошибки σ^2 имеет вид:

$$S^2 = \frac{1}{n-2} * \sum_{i=1}^n e_i^2$$

Чтобы доказать её несмещенность, достаточно осуществить выкладки, аналогичные преобразованиям из предыдущего параграфа, и убедиться, что $E(S^2) = \sigma^2$.

Если в формуле для $\text{var}(\widehat{\beta}_2)$ вместо дисперсии случайной ошибки σ^2 подставить её оценку S^2 , мы получим несмещенную оценку дисперсии МНК-оценки $\widehat{\beta}_2$, которая будет иметь вид:

$$\widehat{\text{var}}(\widehat{\beta}_2) = \frac{S^2}{\sum (x_i - \bar{x})^2}$$

Корень из этой величины называется стандартной ошибкой оценки коэффициента $\widehat{\beta}_2$:

$$se(\widehat{\beta}_2) = \sqrt{\widehat{\text{var}}(\widehat{\beta}_2)} = \sqrt{\frac{S^2}{\sum (x_i - \bar{x})^2}}$$

Аналогичным образом вычисляется стандартная ошибка оценки коэффициента $\widehat{\beta}_1$ (здесь мы опираемся на равенство 2.4, заменяя в нем дисперсию случайной ошибки её оценкой).

$$se(\widehat{\beta}_1) = \sqrt{\widehat{\text{var}}(\widehat{\beta}_1)} = \sqrt{\frac{\frac{S^2}{n} * \sum x_i^2}{\sum (x_i - \bar{x})^2}}$$

Стандартные ошибки оценок коэффициентов пригодятся нам для тестирования гипотез.

Представим, что мы хотим выяснить, влияет ли уровень образования (переменная x) на заработную плату работника в некоторой отрасли (переменная y)? Ответы на такого сорта вопросы, как мы обсудили в первой главе, и есть одна из главных задач эконометрики.

Представим также, что все предпосылки классической линейной модели парной регрессии выполнены. Тогда в терминах нашей модели вопрос «Верно ли, что образование **не влияет** на заработную плату?» эквивалентен вопросу «Верно ли, что в регрессии $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ коэффициент β_2 равен нулю?».

Как мы могли бы ответить на этот вопрос?

Естественная идея состоит в том, чтобы посмотреть оценки коэффициентов $\widehat{\beta}_1$ и $\widehat{\beta}_2$ и увидеть, равен ли коэффициент $\widehat{\beta}_2$ нулю.

Однако при этом возникает следующая проблема: $\widehat{\beta}_1$ и $\widehat{\beta}_2$ — оценки, полученные при помощи МНК на основе случайной выборки. Следовательно, они сами являются случайными величинами, которые могут принимать значения лишь «приблизительно» равные истинным. Поэтому, даже если истинное значение коэффициента β_2 равно нулю, его оценка $\widehat{\beta}_2$, скорее всего, будет отклоняться от нуля.

Следовательно, нужно уметь определять, достаточно ли сильно $\widehat{\beta}_2$ отличается от нуля для того, чтобы можно было с уверенностью утверждать, что и истинное значение коэффициента β_2 также не равно нулю. Опишем процедуру, которая позволяет это сделать.

Процедура тестирования незначимости коэффициента:

Формулируем тестируемую гипотезу $H_0: \beta_2 = 0$ («переменная x не влияет на переменную y ») и альтернативную гипотезу $H_1: \beta_2 \neq 0$ («переменная x влияет на переменную y »)

Находим расчетное значение тестовой статистики по формуле $\frac{\widehat{\beta}_2}{se(\widehat{\beta}_2)}$.

Выбираем уровень значимости α . Уровнем значимости в математической статистике называется вероятность ошибки первого рода, то есть вероятность отклонить тестируемую гипотезу при условии, что в действительности эта гипотеза верна. Разумеется, нам хотелось бы ошибаться не слишком часто, поэтому данную вероятность обычно выбирают маленькой. Чаще всего в эконометрике используются уровни значимости 1% и 5%.

Из таблиц распределения Стюдента находим критическое значение тестовой статистики t_{n-2}^α для выбранного уровня значимости и так называемого числа степеней свободы, которое в нашем случае равно $(n - 2)$.

Если $\left| \frac{\hat{\beta}_2}{se(\hat{\beta}_2)} \right| > t_{n-2}^\alpha$, то есть $\hat{\beta}_2$ достаточно велик по абсолютной величине, следует отвергнуть гипотезу $H_0 : \beta_2 = 0$ и сделать вывод в пользу альтернативной гипотезы, то есть заключить, что переменная x влияет на переменную y . В этом случае переменную x называют статистически значимой при уровне значимости α . В противном случае, соответственно, гипотеза H_0 не может быть отвергнута, и переменную x называют статистически незначимой при уровне значимости α .

Здесь и далее во всех тестах, если явно не указано иное, мы подразумеваем альтернативную гипотезу " β_2 не равно c ", а не $\beta_2 < c$ или $\beta_2 > c$. Поэтому под критическими значениями из таблиц распределения Стюдента по умолчанию подразумеваются критические значения для двусторонних (а не односторонних) тестов. Все стандартные эконометрические пакеты используют такой же подход.

Замечание 1. В этой процедуре мы опираемся на тот факт, что тестовая статистика имеет t -распределение Стюдента. Чтобы это было верно, как раз и нужна предпосылка №6 КЛМНР, которую мы до этого никак не использовали.

В соответствии с этой предпосылкой случайные ошибки имеют нормальное распределение. Мы показали (см. равенство (2.2)), что $\hat{\beta}_2$ — это линейная комбинация случайных ошибок, то есть независимых, одинаково и нормально распределенных случайных величин.

Из математической статистики известно, что отсюда следуют два утверждения:

Во-первых, $\hat{\beta}_2$ имеет нормальное распределение (так как линейная комбинация нормальных случайных величин является нормальной случайной величиной), дисперсию и математическое ожидание которого мы вычислили в предыдущем параграфе. Иными словами $\hat{\beta}_2$ имеет вот такое распределение:

$$N\left(\beta_2, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right)$$

Во-вторых, случайная величина $\frac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)}$ имеет t -распределение Стюдента. В нашем случае это будет распределение с $n - 2$ степенями свободы: $\frac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)} \sim t_{n-2}$

В частности, если верна сформулированная нами гипотеза $\beta_2 = 0$, то распределение Стюдента имеет дробь $\frac{\hat{\beta}_2}{se(\hat{\beta}_2)}$, которую мы используем в нашей процедуре. В этом случае критическое значение определяется из вот такого условия (его геометрическая интерпретация представлена в примере 2.3):

$$P\left(\left| \frac{\hat{\beta}_2}{se(\hat{\beta}_2)} \right| < t_{n-2}^\alpha\right) = 1 - \alpha.$$

Замечание 2. Аналогичным образом можно тестировать гипотезу $H_0 : \beta_2 = c$ (против альтернативной гипотезы $H_0 : \beta_2 \neq c$), где c — это некоторая константа. В этом случае процедура тестирования остается такой же с одним исключением: расчетное значение тестовой статистики будет иметь вид $\frac{\hat{\beta}_2 - c}{se(\hat{\beta}_2)}$.

Замечание 3. Раньше для определения величины критического значения t_{n-2}^α было необходимо использовать таблицы распределения Стюдента. Сейчас этот способ тоже доступен (например, соответствующая таблица представлена в Приложении 3.A в конце третьей главы), однако теперь это значение можно рассчитать непосредственно в эконометрическом пакете или, например, в MS Excel (см. пример ниже).

Альтернативным способом является использование для тестирования гипотезы так называемого p -значения. P -значением называют такой уровень значимости, при котором тестируемая гипотеза находится на грани между отвержением и принятием.

Поэтому использовать p -значение при принятии решения очень просто: если оно меньше заранее выбранного уровня значимости α , то тестируемая гипотеза отвергается при уровне значимости α . Например, если при тестировании незначимости коэффициента вы используете пятипроцентный уровень значимости ($\alpha = 0,05$), а p -значение оказалось равно 0,0002, следует заключить, что соответствующий коэффициент является значимым. Удобство использования p -значения состоит в том, что эта величина автоматически рассчитывается всеми стандартными эконометрическими пакетами, поэтому для принятия решения о значимости или незначимости того или иного коэффициента (а также для проведения любых других тестов, которые мы обсудим далее) вам не требуется никаких таблиц распределения и никаких дополнительных расчетов.

Рассмотрим для большей наглядности еще один пример.

Пример 2.3. Тестирование незначимости коэффициента и графическая иллюстрация

Представим, что у нас 10 наблюдений ($n = 10$), оценка коэффициента оказалась равна $\hat{\beta}_2 = 8,0$, а ее стандартная ошибка $se(\hat{\beta}_2) = 4,0$. Если использовать подход, связанный с критическими значениями, нужно открыть таблицу распределения Стьюдента (см. Приложение 3.А), и найти критическое значение для пятипроцентного уровня значимости и $(n - 2) = 8$ степеней свободы¹. Это критическое значение $t_{кр} = t_8^{0,05} \approx 2,3$. Расчетное значение t-статистики здесь тоже посчитать несложно

$$t_{расч} = \frac{\hat{\beta}_2}{se(\hat{\beta}_2)} = \frac{8}{4} = 2$$

Если мы нанесем все указанные значения на картинку, у нас получится рисунок 2.4а. Критическое значение отсекает по 2,5% слева и справа (всего 5%). Следовательно, вероятность попасть между $-t_{кр}$ и $t_{кр}$ будет 95%. Нанесем также $-t_{расч}$ и $t_{расч}$. Эти значения отсекают по 3% справа и слева, как это показано на рисунке 2.4б.

Обозначим ξ — случайную величину, имеющую распределение стьюдента с $(n - 2) = 8$ степенями свободы. Тогда формально Р-значение в нашем случае — это вот такая вероятность:

р-значение $P(|\xi| > 2)$.

То есть в нашем примере р-значение — это вероятность такого события, что случайная величина, имеющая t-распределение Стьюдента с 8 степенями свободы, по модулю превысит $t_{расч} = 2$. Как видно из рисунка, в нашем случае эта вероятность равна $0,03 + 0,03 = 0,06$.

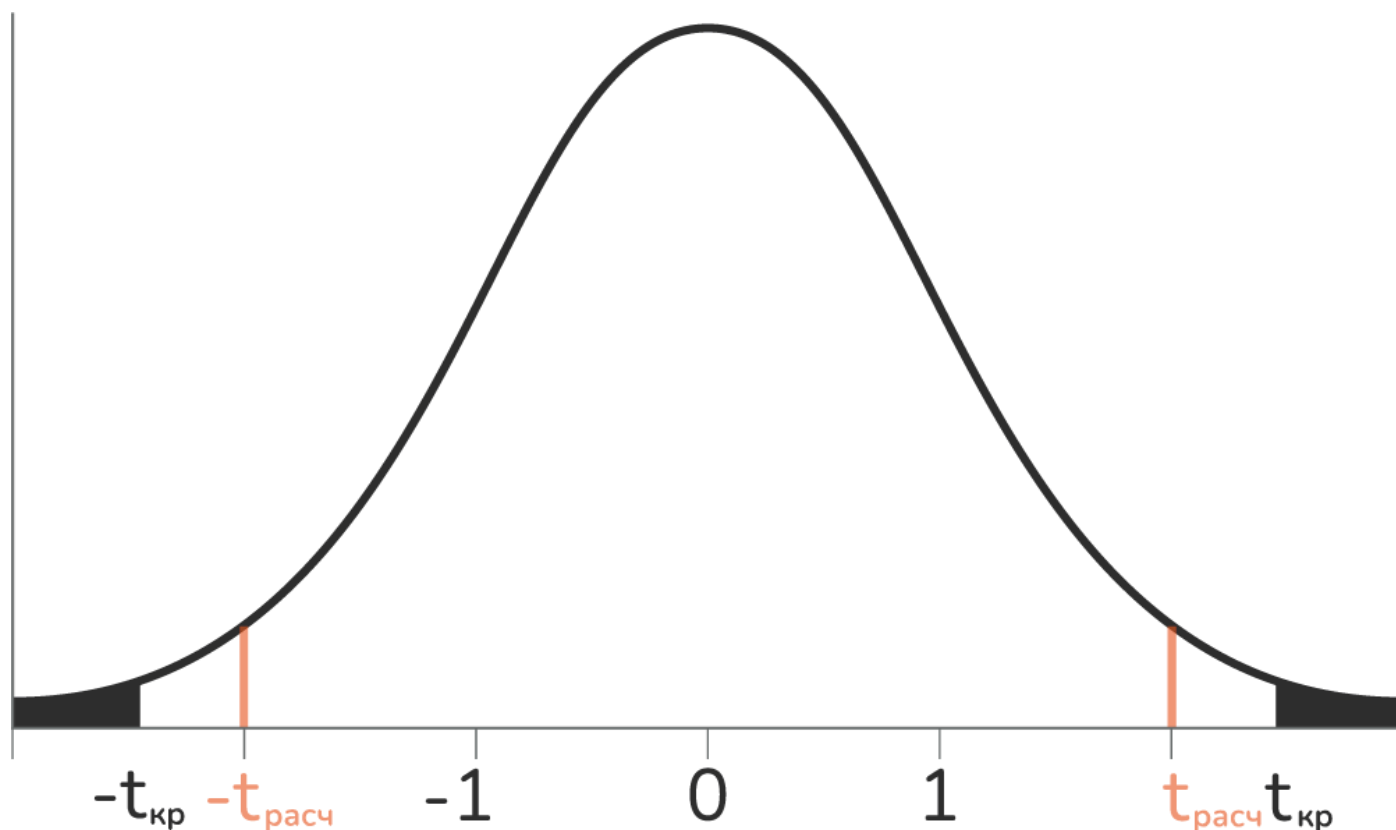


Рисунок 2.4а. Расчетное и критическое значения тестовой статистики для примера 2.3.

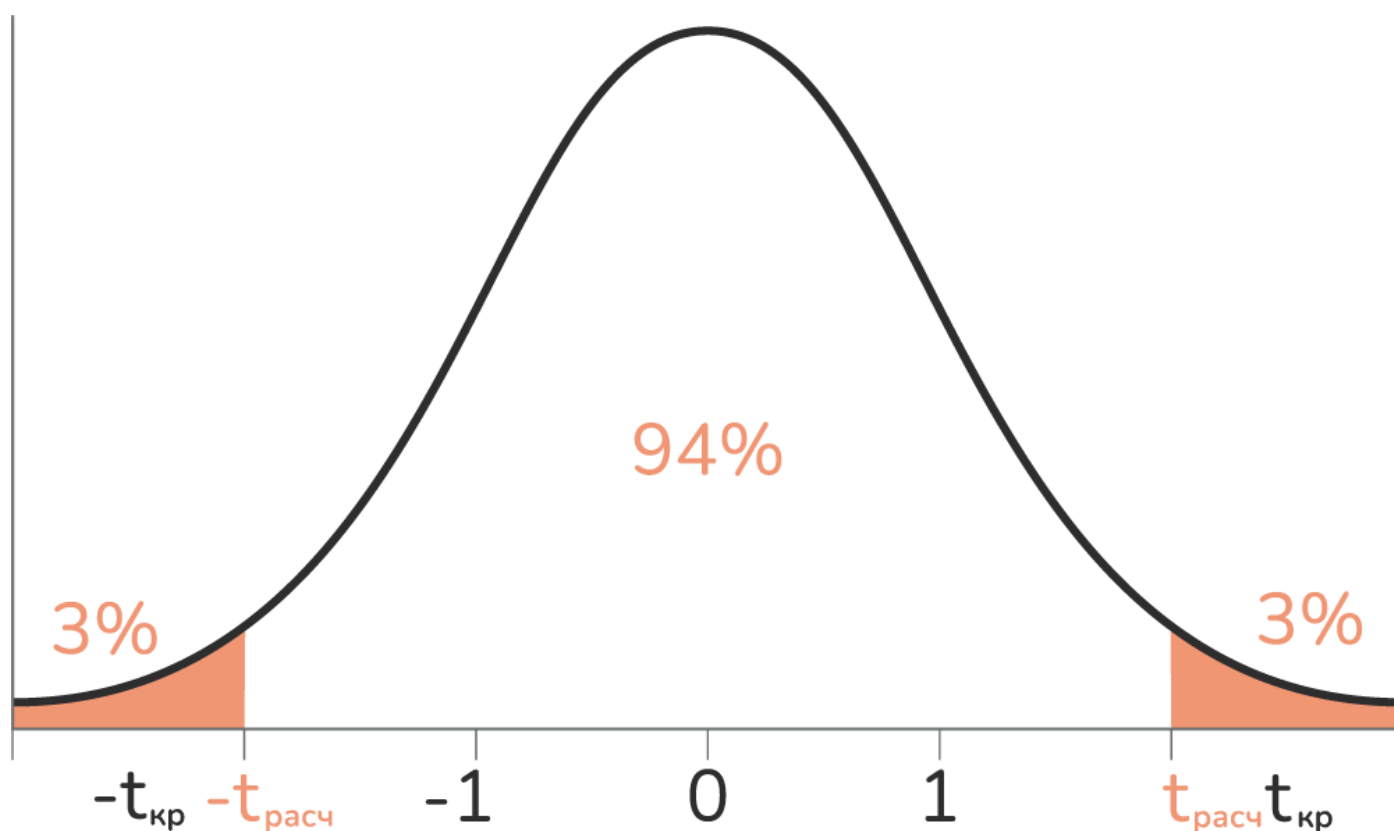


Рисунок 2.46. Р-значение для примера 2.3.

Как видно из нашего примера, Р-значение больше заранее выбранного уровня значимости только тогда, когда $|t_{расч}| < t_{кр}$, что подтверждает сформулированное нами правило принятия решения при помощи Р-значения: если Р-значение больше уровня значимости, то нулевая гипотеза не отвергается. Если Р-значение меньше уровня значимости, то нулевая гипотеза отвергается.

Решив неравенство $\left| \frac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)} \right| < t_{n-2}^\alpha$ относительно β_2 , получим:

$$\hat{\beta}_2 - se(\hat{\beta}_2) * t_{n-2}^\alpha < \beta_2 < \hat{\beta}_2 + se(\hat{\beta}_2) * t_{n-2}^\alpha$$

Иными словами, с вероятностью $1 - \alpha$ интервал $(\hat{\beta}_2 - se(\hat{\beta}_2) * t_{n-2}^\alpha; \hat{\beta}_2 + se(\hat{\beta}_2) * t_{n-2}^\alpha)$ содержит истинное значение оцениваемого параметра. Например, если $\alpha = 0,05$ и, следовательно, $1 - \alpha = 0,95$, этот интервал и называют 95-процентным доверительным интервалом для параметра β_2 .

Возможность построения доверительных интервалов важна с практической точки зрения. Дело в том, что, так как $\hat{\beta}_2$ является лишь приблизительной оценкой параметра β_2 , эта точечная оценка сама по себе несет гораздо меньше информации, чем интервал. Ведь без доверительного интервала невозможно понять, насколько эта оценка на самом деле (не)точная. Например, утверждение « $\hat{\beta}_2$ равно 23,4» куда менее информативно, чем утверждение «истинное значение оцениваемого параметра с вероятностью 95 процентов содержится в пределах от 23,1 до 23,7».

Завершим раздел еще двумя примерами. В первом из них все расчеты проделаны вручную, чтобы, проследив их, можно было еще раз разобраться во взаимосвязях между введенными нами понятиями. Во втором примере используется эконометрический пакет, что позволяет продемонстрировать, как подобные вычисления осуществляются в реальных прикладных исследованиях.

Пример 2.4. Доходы индивидов и потребление риса

Исследователь анализирует зависимость потребления риса от уровня дохода (кривую Энгеля) для однородной группы из 20 потребителей. Все потребители из этой группы сталкиваются с одинаковыми ценами на рис и другие товары, и только уровни дохода у них различны, поэтому исследователь использует модель парной регрессии.

Обозначим:

x_i — ежемесячный располагаемый доход i -го потребителя (в тысячах денежных единиц),

y_i — ежемесячное потребление риса i -м потребителем (в килограммах).

Имеются следующие данные о переменных x и y :

$$\sum_{i=1}^{20} x_i = 20, \sum_{i=1}^{20} x_i^2 = 40, \sum_{i=1}^{20} y_i = 42, \sum_{i=1}^{20} y_i^2 = 108,$$

$$\sum_{i=1}^{20} x_i * y_i = 60$$

(а) Вычислите МНК-оценки коэффициентов в регрессии

$$y_i = \beta_1 + \beta_2 * x_i + \varepsilon_i.$$

Выпишите полученное уравнение регрессии и коэффициент R^2 .

(б) При уровне значимости 5% проверьте значимость переменной x .

(в) Дайте содержательную интерпретацию коэффициента при переменной x .

(г) Вспомнив соответствующие определения из курса микроэкономики и вычислив необходимую эластичность, определите: является ли рис для этой группы потребителей низкокачественным товаром, товаром первой необходимости или предметом роскоши?

(д) При уровне значимости 5% проверьте гипотезу о том, что коэффициент β_2 равен единице.

(е) Постройте 95-процентный доверительный интервал для коэффициента β_2 .

Решение:

(а) Вычислим средние значения:

$$\bar{x} = 1, \bar{x^2} = 2, \bar{y} = 2, \bar{y^2} = 5, \bar{xy} = 3$$

Найдем оценки коэффициентов:

$$\hat{\beta}_2 = \frac{\bar{xy} - \bar{x} * \bar{y}}{\bar{x^2} - \bar{x}^2} = \frac{3 - 1 * 2, 1}{2 - 1} = 0, 9$$

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 * \bar{x} = 2, 1 - 1 * 0, 9 = 1, 2$$

Таким образом, $\hat{y}_i = 1, 2 + 0, 9 * x_i$.

Теперь вычислим R^2 . Для этого воспользуемся тем, что по определению он равен отношению объясненной суммы квадратов к общей сумме квадратов:

$$R^2 = \frac{\sum_{i=1}^{20} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{20} (y_i - \bar{y})^2}.$$

Вычислим каждую из этих сумм по отдельности. Сначала найдем общую сумму квадратов:

$$\begin{aligned} TSS &= \sum_{i=1}^{20} (y_i - \bar{y})^2 = \sum_{i=1}^{20} y_i^2 - 2 * \sum_{i=1}^{20} y_i * \bar{y} + \sum_{i=1}^{20} \bar{y}^2 = \\ &= \sum_{i=1}^{20} y_i^2 - 2 * \bar{y} * \sum_{i=1}^{20} y_i + 20 * \bar{y}^2 = 108 - 2 * 2, 1 * 42 + 20 * 2, 1^2 = 19, 8 \end{aligned}$$

Теперь найдем объясненную сумму квадратов:

$$\begin{aligned} \sum_{i=1}^{20} (\hat{y}_i - \bar{y})^2 &= \sum_{i=1}^{20} (1, 2 + 0, 9 * x_i - 2, 1)^2 = \sum_{i=1}^{20} (0, 9 * x_i - 0, 9)^2 = \\ &= 0, 9^2 \sum_{i=1}^{20} (x_i - 1)^2 = 0, 81 * \left(\sum_{i=1}^{20} x_i^2 - 2 * \sum_{i=1}^{20} x_i + 20 \right) = 16, 2 \end{aligned}$$

Теперь можно вычислить коэффициент детерминации:

$$R^2 = \frac{\sum_{i=1}^{20} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{20} (y_i - \bar{y})^2} = \frac{16, 2}{19, 8} = 0, 82$$

Ответ на пункт (а): $\hat{y}_i = 1, 2 + 0, 9 * x_i, R^2 = 0, 82$

(б) Тестируемая гипотеза $H_0 : \beta_2 = 0$. Альтернативная гипотеза $H_1 : \beta_2 \neq 0$.

Чтобы проверить значимость, нам понадобится стандартная ошибка оценки коэффициента. Для этого нам придется оценить сумму квадратов остатков. Воспользуемся тем фактом, что для регрессии с константой верно равенство:

$$\sum_{i=1}^{20} (y_i - \bar{y})^2 = \sum_{i=1}^{20} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{20} e_i^2$$

В этой формуле мы вычислили все элементы, кроме суммы квадратов остатков:

$$19,8 = 16,2 + \sum_{i=1}^{20} e_i^2$$

Следовательно, $\sum_{i=1}^{20} e_i^2 = 19,8 - 16,2 = 3,6$.

Вычислим оценку дисперсии случайной ошибки:

$$S^2 = \frac{\sum_{i=1}^{20} e_i^2}{n-2} = \frac{3,6}{20-2} = 0,2$$

Теперь вычислим стандартную ошибку оценки коэффициента:

$$\begin{aligned} se(\hat{\beta}_2) &= \sqrt{\frac{S^2}{\sum_{i=1}^{20} (x_i - \bar{x})^2}} = \\ &= \sqrt{\frac{0,2}{\sum_{i=1}^{20} x_i^2 - n \cdot (\bar{x})^2}} = \sqrt{\frac{0,2}{40 - 20}} = 0,1 \end{aligned}$$

Расчетное значение t-статистики равно $\frac{\hat{\beta}_2}{se(\hat{\beta}_2)} = \frac{0,9}{0,1} = 9$.

Критическое значение t-статистики из таблицы распределения Стьюдента при уровне значимости 5% и $(20-2)=18$ степенях свободы составляет 2,101. Расчетное значение больше критического, следовательно, мы отклоняем нулевую гипотезу и делаем вывод о том, что уровень дохода индивида значимо влияет на его спрос на рис.

Ответ на пункт (б): Переменная значима.

Ответ на пункт (в): При увеличении располагаемого дохода потребителя на одну тысячу денежных единиц его спрос на рис увеличивается в среднем на 0,9 кг.

(г) Вычислим эластичность спроса на рис по доходу. По определению эластичность равна:

$$\frac{\frac{dy}{dx} \cdot x}{\hat{y}} = \frac{0,9 \cdot x}{1,2 + 0,9 \cdot x}$$

Легко видеть, что при любых положительных значениях x эластичность спроса по доходу лежит между нулем и единицей, следовательно, для рассматриваемой группы потребителей рис является товаром первой необходимости. Что, в общем-то, неудивительно.

Ответ на пункт (г): Товар первой необходимости.

(д) Тестируемая гипотеза $H_0 : \beta_2 = 1$. Альтернативная гипотеза $H_1 : \beta_2 \neq 1$.

Для проверки значимости нам понадобится стандартная ошибка оценки коэффициента.

Расчетное значение t-статистики равно $\frac{\hat{\beta}_2 - 1}{se(\hat{\beta}_2)} = \frac{0,9 - 1}{0,1} = -1$.

Критическое значение t-статистики из таблицы распределения Стьюдента при уровне значимости 5% и $(20-2)=18$ степенях свободы составляет 2,101. Расчетное значение по модулю меньше критического, следовательно, мы принимаем (не отклоняем) нулевую гипотезу.

Ответ на пункт (д): Гипотеза не отклоняется.

(е) В рамках предпосылок классической линейной модели парной регрессии доверительный интервал может быть посчитан следующим образом:

$$\begin{aligned} &(\hat{\beta}_2 - se(\hat{\beta}_2) \cdot t_{n-2}; \hat{\beta}_2 + se(\hat{\beta}_2) \cdot t_{n-2}) \\ &(0,9 - 0,1 \cdot 2,101; 0,9 + 0,1 \cdot 2,101) \end{aligned}$$

Таким образом, с вероятностью 95% интервал $(0,69; 1,11)$ содержит истинное значение коэффициента β_2 .

Ответ на пункт (е): $(0,69; 1,11)$.

Пример 2.5. Площадь однокомнатной квартиры и её цена

В этом задании вам предлагается проанализировать взаимосвязь между площадью квартиры и ее ценой. Вам доступны следующие данные о московском рынке недвижимости в 2012 году (файл Price2012):

Price — рыночная цена однокомнатной квартиры в Москве (в тысячах руб.), выкуп которой был осуществлен с 10.01.2012 по 28.09.2012

TotalArea — общая площадь квартиры (кв. м)

(а) Оцените регрессию переменной *Price* на переменную *TotalArea*. Запишите оцененное уравнение регрессии, указав коэффициент детерминации и (в скобках под соответствующими коэффициентами) стандартные ошибки. Постройте диаграмму рассеяния с линией регрессии.

(б) Является ли коэффициент при переменной *TotalArea* статистически значимым при уровне значимости 1%? Дайте содержательную интерпретацию для этого коэффициента.

Решение:

(а) Ниже представлена распечатка результатов оценивания уравнения в эконометрическом пакете Gretl². (Любой стандартный эконометрический пакет, например R, Stata или Econometric Views, выдаст аналогичную табличку. Пользуйтесь тем из них, который вам больше нравится. Ну или тем, который есть под рукой.)

Модель 1: МНК, использованы наблюдения 1-121

Зависимая переменная: Price

	Коэффициент	Ст. ошибка	t-статистика	P-значение	
const	786,456	583,051	1,349	0,1799	
TotalArea	135,317	16,5144	8,194	0,0001	***

Среднее зав. перемен	5540,335		Ст. откл. зав. перемен	792,7127
Сумма кв. остатков	48208118		Ст. ошибка модели	636,4827
R-квадрат	0,360696		Испр. R-квадрат	0,355324
F(1, 119)	67,14000		P-значение (F)	3,30e-13
Лог. правдоподобие	−951,8540		Крит. Акаике	1907,708
Крит. Шварца	1913,300		Крит. Хеннана-Куинна	1909,979

В столбце «Коэффициент» указаны оценки коэффициентов, а в столбце «Ст. ошибка» — их стандартные ошибки. В нижней части таблицы среди прочих показателей можно найти и коэффициент R-квадрат.

Общепринятый формат записи полученных результатов имеет следующий вид (в скобках под оценками коэффициентов указаны соответствующие стандартные ошибки):

$$\widehat{Price}_i = 786,456 + 135,317 TotalArea_i, R^2 = 0,36$$

(583,051) (16,514)

Обратите внимание, что в скобках под оценками коэффициентов мы указали их стандартные ошибки. Такой формат является хорошим тоном при записи результатов эконометрического моделирования, так как позволяет читателю оценить точность ваших результатов и прикинуть доверительные интервалы для коэффициентов.

(б) В столбце «P-значение» указано, что P-значение для оценки коэффициента при переменной *TotalArea* меньше, чем 0,0001 (и тем более меньше, чем 0,01). Следовательно, этот коэффициент является статистически значимым при уровне значимости 1%.

Содержательная интерпретация: при увеличении общей площади квартиры на один квадратный метр ее цена в среднем при прочих равных условиях увеличивается на 135 тысяч рублей.

Отметим, что свободное слагаемое в данном случае отличается от нуля статистически незначимо, так как соответствующее P-значение равно 0,18, что больше любого разумного уровня значимости. Да и если бы даже эта константа была значима, все равно отдельно интерпретировать её смысла не было бы, ведь константа показывает значение зависимой переменной при условии, что регрессор *TotalArea* равен нулю (то есть при условии, что анализируемая квартира имеет нулевую площадь). Вряд ли кто-то всерьез интересуется ценой квартиры площадью 0 квадратных метров.

1. Вместо использования готовых таблиц распределения можно, например, ввести в MS Excel формулу =СТЫЮДЕНТ.ОБР(1 - 0,05 / 2; 10 - 2) [↗](#)

2. Для получения этого результата достаточно запустить Gretl; используя пункт меню «Импорт», импортировать данные из файла MS Excel (или просто мышкой «перетащить» нужный файл в рабочую область эконометрического пакета); выбрать в меню «Модель» пункт «Метод наименьших квадратов» и указать в качестве зависимой переменной переменную Price, а в качестве объясняющей — переменную TotalArea. [↗](#)

2.6. Прогнозирование

В зависимости от контекста термин «прогнозирование» в эконометрике может трактоваться по-разному. Применительно к данным временных рядов речь обычно идет о прогнозировании будущего значения зависимой переменной, например, курса рубля или ВВП. Когда же речь идет о пространственных выборках, под прогнозированием понимают предсказание значения зависимой переменной для заданных значений объясняющих переменных. Например, предсказание цены квартиры с заданной жилой площадью.

Формально задачу построения прогноза можно представить следующим образом. Имеется модель, для которой выполнены все предпосылки КЛМГР:

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

Представим, что мы уже воспользовались МНК и получили оцененную на основе n наблюдений линию регрессии:

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$$

Теперь пусть у нас есть известное $(n+1)$ -ое наблюдение регрессора x_{n+1} , но неизвестно соответствующее значение зависимой переменной y_{n+1} и нужно построить его прогноз. Естественной идеей будет подставить известное значение в оцененную регрессию:

$$\hat{y}_{n+1} = \hat{\beta}_1 + \hat{\beta}_2 x_{n+1}$$

Оказывается, что это хорошая мысль: такой прогноз будет несмещенным и эффективным (то есть будет характеризоваться минимальной ожидаемой квадратичной ошибкой прогноза).

Докажем несмещенность этого прогноза.

Вычислим математическое ожидание фактического значения y_{n+1} и нашего прогноза $\hat{y}_{n+1}$. Если прогноз несмещенный, то эти математические ожидания будут совпадать.

Воспользуемся тем, что, как мы доказали выше, $\hat{\beta}_1$ и $\hat{\beta}_2$ — несмещенные оценки коэффициентов β_1 и β_2 :

$$E(\hat{y}_{n+1}) = E(\hat{\beta}_1 + \hat{\beta}_2 x_{n+1}) = E(\hat{\beta}_1) + E(\hat{\beta}_2) x_{n+1} = \beta_1 + \beta_2 x_{n+1}$$

Кроме того:

$$\begin{aligned} E(y_{n+1}) &= E(\beta_1 + \beta_2 x_{n+1} + \varepsilon_{n+1}) = \\ &= \beta_1 + \beta_2 x_{n+1} + E(\varepsilon_{n+1}) = \beta_1 + \beta_2 x_{n+1} \end{aligned}$$

Следовательно, $E(y_{n+1}) = E(\hat{y}_{n+1})$.

Кроме самого прогноза нас интересует его точность. Чтобы её оценить, целесообразно вычислить математические ожидания квадрата ошибки прогноза:

$$\begin{aligned} E(\hat{y}_{n+1} - y_{n+1})^2 &= E(\hat{\beta}_1 + \hat{\beta}_2 x_{n+1} - \beta_1 - \beta_2 x_{n+1} - \varepsilon_{n+1})^2 = \\ &= E((\hat{\beta}_1 - \beta_1) + (\hat{\beta}_2 - \beta_2) x_{n+1} - \varepsilon_{n+1})^2 = \\ &= E(\hat{\beta}_1 - \beta_1)^2 + x_{n+1}^2 E(\hat{\beta}_2 - \beta_2)^2 + E(\varepsilon_{n+1})^2 + \\ &\quad + 2x_{n+1} E((\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2)) - 2E((\hat{\beta}_1 - \beta_1)\varepsilon_{n+1}) - \\ &\quad - 2x_{n+1} E((\hat{\beta}_2 - \beta_2)\varepsilon_{n+1}) = \\ &= \text{var}(\hat{\beta}_1) + x_{n+1}^2 \text{var}(\hat{\beta}_2) + \sigma^2 + 2x_{n+1} \text{cov}(\hat{\beta}_1, \hat{\beta}_2) - 0 - 0 = \\ &= \frac{\frac{\sigma^2}{n} * \sum x_i^2}{\sum (x_i - \bar{x})^2} + x_{n+1}^2 \frac{\sigma^2}{\sum (x_i - \bar{x})^2} + \sigma^2 - 2x_{n+1} \frac{\bar{x} * \sigma^2}{\sum (x_i - \bar{x})^2} = \\ &= \sigma^2 * \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right) \end{aligned}$$

Здесь в предпоследнем равенстве мы воспользовались формулами для $\text{var}(\hat{\beta}_1)$, $\text{var}(\hat{\beta}_2)$ и $\text{cov}(\hat{\beta}_1, \hat{\beta}_2)$, представленными выше.

Дисперсия ошибки прогноза σ^2 , неизвестная нам в реальности, может быть заменена несмещенной оценкой S^2 . Если проделать эту замену, а затем извлечь из полученного результата корень, то получим стандартную ошибку прогноза:

$$\delta = \sqrt{s^2 * \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right)}$$

Эту стандартную ошибку прогноза можно использовать для построения доверительного интервала прогноза.

95-процентный доверительный интервал для прогноза — это такой интервал, который накрывает истинное прогнозное значение зависимой переменной с вероятностью 95%. Он имеет вид:

$$(\hat{y}_{n+1} - \delta * t_{n-2}^\alpha, \hat{y}_{n+1} + \delta * t_{n-2}^\alpha)$$

Обратите внимание, что величина стандартной ошибки прогноза зависит от соотношения x_{n+1} и $\bar{x}$. Если $x_{n+1} = \bar{x}$, то последняя дробь в этой большой формуле окажется равной нулю, и стандартная ошибка прогноза будет минимальной. Чем сильнее x_{n+1} отличается от $\bar{x}$, тем больше будет эта дробь. Таким образом, чем меньше наблюдение, для которого вы строите прогноз, похоже на вашу исходную выборку, тем менее точным этот прогноз окажется.

Пример 2.6. Построение прогноза

Рассматривается классическая линейная модель парной регрессии $y_i = \beta_1 + \beta_2 * x_i + \varepsilon_i$. Имеется следующая информация о 10 наблюдениях анализируемых переменных:

$$\sum_{i=1}^{10} x_i = 20, \sum_{i=1}^{10} x_i^2 = 50, \sum_{i=1}^{10} y_i = 8, \sum_{i=1}^{10} y_i^2 = 26,$$

$$\sum_{i=1}^{10} x_i * y_i = 10$$

Для одиннадцатого наблюдения дано $x_{11} = 5$. Предполагая, что это наблюдение удовлетворяет исходной модели, вычислите наилучший линейный несмещенный прогноз y_{11} и оцените его точность, построив для него 95-процентный доверительный интервал.

Решение:

$$\hat{\beta}_2 = \frac{\overline{xy} - \bar{x} * \bar{y}}{\overline{x^2} - \bar{x}^2} = -0,6$$

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 * \bar{x} = 2$$

Прогноз $\hat{y}_{11} = \hat{\beta}_1 + \hat{\beta}_2 * x_{11} = 2 - 0,6 * 5 = -1$.

Сумма квадратов остатков равна:

$$\sum_{i=1}^{10} e_i^2 = \sum_{i=1}^{10} e_i * (y_i - \hat{\beta}_1 - \hat{\beta}_2 * x_i) =$$

$$\sum_{i=1}^{10} e_i y_i - \hat{\beta}_1 \sum_{i=1}^{10} e_i - \hat{\beta}_2 \sum_{i=1}^{10} e_i x_i = \sum_{i=1}^{10} e_i y_i - \hat{\beta}_1 * 0 - \hat{\beta}_2 * 0$$

Последнее равенство верно в силу свойств остатков регрессии. Таким образом:

$$\sum_{i=1}^{10} e_i^2 = \sum_{i=1}^{10} e_i y_i = \sum_{i=1}^{10} (y_i - \hat{\beta}_1 - \hat{\beta}_2 * x_i) y_i =$$

$$\sum_{i=1}^{10} y_i^2 - \hat{\beta}_1 \sum_{i=1}^{10} y_i - \hat{\beta}_2 * \sum_{i=1}^{10} x_i y_i = 26 - 2 * 8 + 0,6 * 10 = 16$$

$$\delta = \sqrt{s^2 * \left(1 + \frac{1}{n} + \frac{(x_{11} - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right)} =$$

$$= \sqrt{\frac{\sum e_i^2}{n-2} * \left(1 + \frac{1}{n} + \frac{(x_{11} - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right)} =$$

$$= \sqrt{\frac{16}{10-2} * \left(1 + \frac{1}{10} + \frac{(5-2)^2}{10}\right)} = 2$$

Теперь можно посчитать доверительный интервал прогноза:

$$\begin{aligned} &(\hat{y}_{11} - \delta * t_8, \hat{y}_{11} + \delta * t_8) \\ &(-1 - 2 * 2, 306, -1 + 2 * 2, 306) \\ &(-5, 612, 3, 612) \end{aligned}$$

Заметим, что в этом примере точность прогноза не слишком высока, что объясняется маленьким количеством наблюдений и тем, что x_{11} довольно далек от среднего по выборке значения переменной x .

Для получения более точного прогноза лучше, конечно, использовать больше данных.

Ответ: $\hat{y}_{11} = -1$, доверительный интервал: $(-5, 612, 3, 612)$

2.7. Заключение

В этой главе мы выяснили, как, используя статистические данные, анализировать влияние одной переменной на другую в рамках классической линейной модели парной регрессии:

- как оценивать параметры уравнения, описывающего эту связь;
- каким образом тестировать гипотезы по поводу этих параметров и строить доверительные интервалы;
- наконец, мы выяснили, как на основе полученного уравнения можно строить прогнозы и оценивать точность этих прогнозов.

Еще раз подчеркнем, что в этой главе мы предполагали выполнение довольно жесткого набора предпосылок. В прикладных исследованиях часть из них, как правило, нарушается. Например, во многих ситуациях естественно анализировать влияние на зависимую переменную не одного, а нескольких факторов. Это приводит нас к необходимости перейти к следующей главе, где мы обсудим модель множественной регрессии.

В этом видео представлен пример эконометрических расчетов в программе Gretl. Ссылка для скачивания программы:

<https://gretl.sourceforge.net/ru.html>

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

[Файл с данными для этого примера.](#)

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

Глава 3. Множественная регрессия: основы

В первом параграфе этой главы мы обсудим мотивацию для изучения множественной регрессии. Во втором — сформулируем предпосылки классической линейной модели множественной регрессии (КЛММР).

Третий и шестой параграфы являются техническими: они содержат векторно-матричную форму записи для КЛММР и ряд доказательств. Если вы хотите ограничиться прикладными аспектами множественной регрессии, а все соответствующие вычисления готовы доверить эконометрическому пакету, можете пропустить эти параграфы.

В четвертом и пятом параграфах рассматриваются способы измерения качества соответствия модели данным, тестирование гипотез и построение доверительных интервалов.

В последнем, седьмом, параграфе мы кратко обобщим основные идеи главы.

[3.1. Почему не стоит ограничиваться парной регрессией](#)

[3.2. Классическая линейная модель множественной регрессии](#)

[3.3. Векторно-матричная форма записи и некоторые доказательства](#)

[3.4. Степень соответствия модели данным](#)

[3.5. Тестирование гипотез и построение доверительных интервалов](#)

[3.6. Тест на линейное ограничение общего вида](#)

[3.7. Обобщающий пример](#)

[Задания для самостоятельного решения](#)

[Приложение 3.A. таблицы распределения Стьюдента и Фишера](#)

3.1. Почему не стоит ограничиваться парной регрессией

[Внешняя ссылка]: [RuTube](#), [YouTube](#)

Предположим, что выполнены все предпосылки классической линейной модели парной регрессии за одним исключением — на зависимую переменную влияют не один, а два регрессора:

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 w_i + \varepsilon_i$$

Например, мы заинтересованы в оценке влияния уровня образования индивида (переменная x) на его уровень дохода (переменная y). Иными словами, мы заинтересованы в получении корректной оценки коэффициента β_2 . Однако вполне естественно ожидать, что на уровень дохода работника влияет еще и стаж его работы (переменная w).

Представим, что мы игнорируем второй фактор и оцениваем парную регрессию переменной y по переменной x .

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i.$$

Будет ли в этом случае оценка коэффициента $\hat{\beta}_2$ несмещенной? Для ответа на этот вопрос преобразуем её следующим образом:

$$\begin{aligned}\hat{\beta}_2 &= \frac{\widehat{cov}(x, y)}{\widehat{var}(x)} = \frac{\widehat{cov}(x, \beta_1 + \beta_2 x + \beta_3 w + \varepsilon)}{\widehat{var}(x)} = \\ &= \frac{\beta_2 \widehat{cov}(x, x) + \beta_3 \widehat{cov}(x, w) + \widehat{cov}(x, \varepsilon)}{\widehat{var}(x)} = \\ &= \beta_2 + \beta_3 \frac{\widehat{cov}(x, w)}{\widehat{var}(x)} + \frac{\widehat{cov}(x, \varepsilon)}{\widehat{var}(x)}\end{aligned}$$

Теперь для проверки несмещенности следует вычислить её математическое ожидание:

$$\begin{aligned}E\hat{\beta}_2 &= E\left(\beta_2 + \beta_3 \frac{\widehat{cov}(x, w)}{\widehat{var}(x)} + \frac{\widehat{cov}(x, \varepsilon)}{\widehat{var}(x)}\right) = \\ &= \beta_2 + \beta_3 \frac{\widehat{cov}(x, w)}{\widehat{var}(x)} + \frac{E(\widehat{cov}(x, \varepsilon))}{\widehat{var}(x)}\end{aligned}$$

Последнее равенство верно, так как все слагаемые, кроме $\widehat{cov}(x, \varepsilon)$, являются неслучайными и, следовательно, могут быть вынесены за знак математического ожидания. Однако $E(\widehat{cov}(x, \varepsilon)) = 0$. Действительно:

$$\begin{aligned}E(\widehat{cov}(x, \varepsilon)) &= E\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})\right) = \\ &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(E\varepsilon_i - E\bar{\varepsilon}) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(0 - 0) = 0\end{aligned}$$

Поэтому

$$E\hat{\beta}_2 = \beta_2 + \beta_3 \frac{\widehat{cov}(x, w)}{\widehat{var}(x)}$$

Из этого равенства легко видеть, что оценка коэффициента при интересующей нас переменной, вообще говоря, смещена. Например, если увеличение стажа работы приводит к увеличению дохода $\beta_3 > 0$ и более образованные работники в среднем имеют более высокий стаж работы $\widehat{cov}(x, w) > 0$, то $\beta_3 \frac{\widehat{cov}(x, w)}{\widehat{var}(x)} > 0$ и, следовательно, $E\hat{\beta}_2 > \beta_2$. В этом случае оценку коэффициента называют завышенной.

Если, напротив, образование отрицательно коррелировано со стажем ($\widehat{cov}(x, w) < 0$), то $\beta_3 \frac{\widehat{cov}(x, w)}{\widehat{var}(x)} < 0$ и $E\hat{\beta}_2 < \beta_2$. В этом случае оценка коэффициента называется заниженной. Отметим, что в нашем примере этот случай более вероятен, так как обычно продолжение обучения связано с отказом от немедленного выхода на рынок труда. Таким образом, в нашем примере, оценив парную регрессию, мы будем получать заниженную оценку коэффициента при переменной x , то есть будем, как правило, недооценивать вклад образования в доходы работника.

Рассмотренный пример показывает, что использование парной регрессии вместо множественной может привести к неверным выводам.

Описанная ситуация называется **смещением из-за пропуска существенной переменной (omitted variable bias)**. Для того чтобы избежать этого смещения, необходимо учитывать в вашей регрессии все существенные факторы (то есть все коррелированные с интересующей вас переменной факторы, коэффициенты при которых в истинной модели регрессии отличны от нуля). Это приводит нас к необходимости анализа модели множественной регрессии.

Прежде чем мы перейдем к этому анализу, подчеркнем, что смещение возникает только в том случае, если пропущенная переменная коррелирована с переменной, коэффициент при которой нас интересует. Действительно, если в нашем примере образование и стаж не связаны между собой $\widehat{cov}(x, w) = 0$, то $E\widehat{\beta}_2 = \beta_2$, и смещение отсутствует. Поэтому если нас интересует эффект от уровня образования, то в регрессию следует включать переменные, которые коррелированы с уровнем образования, а прочие факторы можно игнорировать.

3.2. Классическая линейная модель множественной регрессии

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

При записи уравнения множественной регрессии мы будем использовать следующие обозначения:

$$y_i = \beta_1 + \beta_2 x_i^{(2)} + \beta_3 x_i^{(3)} + \dots + \beta_k x_i^{(k)} + \varepsilon_i, i = 1, \dots, n$$

y_i — зависимая (объясняемая) переменная;

$x_i^{(m)}$ — объясняющие переменные (регрессоры), $m = 2, \dots, k$;

ε_i — случайные ошибки;

k — число коэффициентов в модели;

n — по-прежнему число наблюдений.

Чтобы подчеркнуть, что константа — тоже своеобразный регрессор, это уравнение иногда удобно записывать вот так:

$$y_i = \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \beta_3 x_i^{(3)} + \dots + \beta_k x_i^{(k)} + \varepsilon_i, i = 1, \dots, n, \text{ где } x_i^{(1)} = 1 \text{ для всех наблюдений.}$$

Предпосылки классической линейной модели множественной регрессии во многом схожи с предпосылками аналогичной модели для регрессии парной.

Предпосылки классической линейной модели множественной регрессии (КЛММР)

1. Модель линейна по параметрам и корректно специфицирована

$$y_i = \beta_1 + \beta_2 x_i^{(2)} + \beta_3 x_i^{(3)} + \dots + \beta_k x_i^{(k)} + \varepsilon_i, i = 1, 2, \dots, n.$$

2. Объясняющие переменные $x_i^{(m)}$, $m = 1, 2, \dots, k$ являются детерминированными и линейно независимыми.
3. Математическое ожидание случайных ошибок равно нулю $E\varepsilon_i = 0$.
4. Дисперсия случайной ошибки одинакова для всех наблюдений $\text{var}(\varepsilon_i) = \sigma^2$.
5. Случайные ошибки, относящиеся к разным наблюдениям, взаимно независимы.
6. Случайные ошибки имеют нормальное распределение $\varepsilon_i \sim N(0, \sigma^2)$.

Легко видеть, что отличия от парной регрессии касаются только первых двух предпосылок. В первой предпосылке теперь фигурирует уравнение, в котором не 2 коэффициента, а целых k штук.

Вторая предпосылка теперь требует, чтобы все регрессоры были линейно независимыми. Иными словами, не должно возникать ситуации, когда один регрессор линейно выражается через другие. Скажем, ситуация, когда для каждого наблюдения верно равенство $x^{(2)} = 6x^{(3)} + 5x^{(4)}$ будет означать нарушение этой предпосылки. Такая ситуация представляет собой пример так называемой мультиколлинеарности. Мы подробно обсудим эту проблему в главе 4.

Для КЛММР также может быть сформулирована теорема Гаусса — Маркова.

Теорема Гаусса — Маркова для модели множественной регрессии

Если выполнены предпосылки 1-5 классической линейной модели множественной регрессии, то МНК-оценки коэффициентов модели будут:

- (а) несмещенными,
- (б) эффективными в классе всех несмещенных и линейных по y оценок.

3.3. Векторно-матричная форма записи и некоторые доказательства

Введем следующие обозначения. Вектор значений зависимой переменной:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_{n-1} \\ y_n \end{pmatrix}$$

Вектор случайных ошибок:

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \dots \\ \varepsilon_{n-1} \\ \varepsilon_n \end{pmatrix}$$

Вектор коэффициентов модели:

$$\beta = \begin{pmatrix} \beta_1 \\ \dots \\ \beta_k \end{pmatrix}$$

Вектор МНК-оценок коэффициентов модели:

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \dots \\ \hat{\beta}_k \end{pmatrix}$$

Матрица регрессоров:

$$X = \begin{pmatrix} x_1^{(1)} & \dots & x_1^{(k)} \\ \vdots & \ddots & \vdots \\ x_n^{(1)} & \dots & x_n^{(k)} \end{pmatrix}$$

Таким образом, матрица регрессоров представляет собой таблицу, где число столбцов равно числу регрессоров, и в каждом столбце записаны данные об одном из них. Число строк в ней, соответственно, равно числу наблюдений. По умолчанию мы по-прежнему будем предполагать, что рассматривается регрессия с константой и, следовательно, $x_i^{(1)} = 1$ для всех наблюдений. То есть первый столбец в матрице регрессоров заполнен исключительно единицами. Это предположение не критично для вывода формул МНК-оценок и исследования их свойств, однако полезно для некоторых других целей. Например, чтобы коэффициент R-квадрат в нашей модели гарантированно лежал на отрезке от 0 до 1.

Выведем МНК-оценки коэффициентов в модели множественной регрессии. Напомним, что метод наименьших квадратов состоит в подборе оценок коэффициентов, которые минимизируют сумму квадратов остатков модели:

$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \hat{\beta}_1 x_i^{(1)} - \dots - \hat{\beta}_k x_i^{(k)})^2$$

Эту сумму квадратов мы будем минимизировать по $\hat{\beta}_1, \dots, \hat{\beta}_k$. Так как рассматриваемая квадратичная форма является положительно определенной, то необходимое условие экстремума будет и достаточным условием минимума. Поэтому, чтобы найти нужные нам оценки, следует просто взять производные по $\hat{\beta}_1, \dots, \hat{\beta}_k$ и приравнять каждую из них к нулю. Получим систему из k уравнений:

$$\begin{aligned} -2 \sum x_i^{(1)} (y_i - \hat{\beta}_1 x_i^{(1)} - \dots - \hat{\beta}_k x_i^{(k)}) &= 0 \\ -2 \sum x_i^{(2)} (y_i - \hat{\beta}_1 x_i^{(1)} - \dots - \hat{\beta}_k x_i^{(k)}) &= 0 \\ &\dots \\ -2 \sum x_i^{(k)} (y_i - \hat{\beta}_1 x_i^{(1)} - \dots - \hat{\beta}_k x_i^{(k)}) &= 0 \end{aligned}$$

Эту систему можно переписать так:

$$\begin{aligned}\sum x_i^{(1)} y_i &= \sum x_i^{(1)} (\hat{\beta}_1 x_i^{(1)} + \dots + \hat{\beta}_k x_i^{(k)}) \\ \sum x_i^{(2)} y_i &= \sum x_i^{(2)} (\hat{\beta}_1 x_i^{(1)} + \dots + \hat{\beta}_k x_i^{(k)}) \\ &\dots \\ \sum x_i^{(k)} y_i &= \sum x_i^{(k)} (\hat{\beta}_1 x_i^{(1)} + \dots + \hat{\beta}_k x_i^{(k)})\end{aligned}$$

Или, что то же самое:

$$\begin{pmatrix} \sum x_i^{(1)} y_i \\ \dots \\ \sum x_i^{(k)} y_i \end{pmatrix} = \begin{pmatrix} \sum x_i^{(1)} (\hat{\beta}_1 x_i^{(1)} + \dots + \hat{\beta}_k x_i^{(k)}) \\ \dots \\ \sum x_i^{(k)} (\hat{\beta}_1 x_i^{(1)} + \dots + \hat{\beta}_k x_i^{(k)}) \end{pmatrix}$$

С учетом введенных нами обозначений данную систему можно гораздо более лаконично записать в матричной форме:

$$X'y = X'X\hat{\beta} \quad (3.1)$$

Действительно, для левой части равенства (3.1) верны преобразования:

$$\begin{aligned}X'y &= \begin{pmatrix} x_1^{(1)} & x_2^{(1)} & \dots & x_n^{(1)} \\ \vdots & & \ddots & \vdots \\ x_1^{(k)} & x_2^{(k)} & \dots & x_n^{(k)} \end{pmatrix} * \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} = \\ &= \begin{pmatrix} x_1^{(1)} y_1 + \dots + x_n^{(1)} y_n \\ \dots \\ x_1^{(k)} y_1 + \dots + x_n^{(k)} y_n \end{pmatrix} = \begin{pmatrix} \sum x_i^{(1)} y_i \\ \dots \\ \sum x_i^{(k)} y_i \end{pmatrix}\end{aligned}$$

А для правой части равенства (3.1) можно записать:

$$\begin{aligned}X'X\hat{\beta} &= \begin{pmatrix} \sum_{i=1}^n (x_i^{(1)})^2 & \dots & \sum_{i=1}^n x_i^{(1)} x_i^{(k)} \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_i^{(k)} x_i^{(1)} & \dots & \sum_{i=1}^n (x_i^{(k)})^2 \end{pmatrix} * \begin{pmatrix} \hat{\beta}_1 \\ \dots \\ \hat{\beta}_k \end{pmatrix} = \\ &= \begin{pmatrix} \hat{\beta}_1 \sum_{i=1}^n (x_i^{(1)})^2 + \dots + \hat{\beta}_k \sum_{i=1}^n x_i^{(1)} x_i^{(k)} \\ \dots \\ \hat{\beta}_1 \sum_{i=1}^n x_i^{(k)} x_i^{(1)} + \dots + \hat{\beta}_k \sum_{i=1}^n (x_i^{(k)})^2 \end{pmatrix} = \\ &= \begin{pmatrix} \sum x_i^{(1)} (\hat{\beta}_1 x_i^{(1)} + \dots + \hat{\beta}_k x_i^{(k)}) \\ \dots \\ \sum x_i^{(k)} (\hat{\beta}_1 x_i^{(1)} + \dots + \hat{\beta}_k x_i^{(k)}) \end{pmatrix}\end{aligned}$$

Решая уравнение (3.1) относительно $\hat{\beta}$, находим формулу для расчета вектора МНК-оценок коэффициентов:

$$\hat{\beta} = (X'X)^{-1} X'y.$$

Для исследования свойств этих коэффициентов нам потребуется ввести еще один объект: ковариационную матрицу вектора случайных ошибок. Это такая таблица размера n на n , в которой на пересечении i -ой строки и j -того столбца стоит коэффициент ковариации между случайными ошибками, относящимися к i -му и к j -тому наблюдениям: $Cov(\varepsilon_i, \varepsilon_j)$. Ковариационную матрицу для вектора ε будем обозначать $V(\varepsilon)$. Таким образом, ковариационная матрица вектора оценок коэффициентов имеет вид:

$$V(\varepsilon) = \begin{pmatrix} Cov(\varepsilon_1, \varepsilon_1) & \dots & Cov(\varepsilon_1, \varepsilon_n) \\ \vdots & \ddots & \vdots \\ Cov(\varepsilon_n, \varepsilon_1) & \dots & Cov(\varepsilon_n, \varepsilon_n) \end{pmatrix}$$

Напомним несколько полезных нам свойств ковариационной матрицы.

Свойство 1. Ковариационная матрица симметрична и положительно определена.

Свойство 2. На главной диагонали ковариационной матрицы расположены дисперсии соответствующих элементов случайного вектора.

Свойство 3. Добавление вектора констант не меняет ковариационную матрицу:

$$V(\varepsilon + b) = V(\varepsilon)$$

Свойство 4. Умножение на матрицу констант меняет ковариационную матрицу следующим образом:

$$V(C * \varepsilon) = C * V(\varepsilon) * C'$$

Отметим, что указанные свойства верны для произвольной ковариационной матрицы, а не только для ковариационной матрицы вектора случайных ошибок.

С учетом сформулированных определений можно переписать предпосылки КЛММР в матричной форме. Сравните их с предпосылками, сформулированными в предыдущем параграфе, и убедитесь, что это в точности одно и то же.

Предпосылки классической линейной модели множественной регрессии в матричной форме

1. $y = X\beta + \varepsilon$

2. Матрица X — детерминированная матрица, имеющая максимальный ранг k .

3. $E\varepsilon = 0_n$

$$4-5. V(\varepsilon) = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix} = \sigma^2 I_n$$

6*. $\varepsilon \sim N(0_n, \sigma^2 I_n)$

Здесь I_n — единичная матрица размера n на n , а 0_n — нулевой вектор-столбец длины n .

Исследуем свойства МНК-оценок в случае выполнения указанных предпосылок. Для начала отметим, что вторая предпосылка гарантирует существование вектора МНК-оценок $\hat{\beta} = (X'X)^{-1}X'y$. Действительно, если матрица X имеет максимальный ранг k , то матрица $X'X$ имеет ранг k . Следовательно, она является невырожденной, поэтому существует $(X'X)^{-1}$.

Теперь вычислим математическое ожидание вектора МНК-оценок:

$$\begin{aligned} E\hat{\beta} &= E\left((X'X)^{-1}X'y\right) = E\left((X'X)^{-1}X'(X\beta + \varepsilon)\right) = \\ &= E\left((X'X)^{-1}X'X\beta + (X'X)^{-1}X'\varepsilon\right) = \\ &= E\left(\beta + (X'X)^{-1}X'\varepsilon\right) = \\ &= \beta + (X'X)^{-1}X'E(\varepsilon) = \beta + (X'X)^{-1}X'0_n = \beta \end{aligned}$$

Таким образом, мы показали, что МНК-оценка не смещена, то есть доказали первую часть теоремы Гаусса — Маркова для множественной регрессии. Здесь мы использовали предпосылку №1 (когда подставляли выражение $X\beta + \varepsilon$ вместо вектора y), предпосылку №2 о том, что матрица регрессоров детерминированная (когда выносили её за знак математического ожидания) и предпосылку №3 о том, что математическое ожидание вектора случайных ошибок равно нулю.

Предпосылки 4-6 для несмещенности МНК-оценок не нужны. Однако они понадобятся нам для вычисления ковариационной матрицы вектора оценок коэффициентов $V(\hat{\beta})$, то есть матрицы размера k на k , где на пересечении i -той строки и j -го столбца стоит ковариация $cov(\hat{\beta}_i, \hat{\beta}_j)$.

Эта матрица пригодится нам для тестирования гипотез относительно коэффициентов модели. Например, на главной диагонали такой матрицы стоят дисперсии оценок коэффициентов $cov(\hat{\beta}_i, \hat{\beta}_i) = var(\hat{\beta}_i)$. Нам потребуется оценить их для тестирования незначимости соответствующих коэффициентов.

$$\begin{aligned} V(\hat{\beta}) &= V\left((X'X)^{-1}X'y\right) = V\left((X'X)^{-1}X'(X\beta + \varepsilon)\right) = \\ &= V\left((X'X)^{-1}X'X\beta + (X'X)^{-1}X'\varepsilon\right) = \\ &= V\left(\beta + (X'X)^{-1}X'\varepsilon\right) = \end{aligned}$$

= {воспользуемся 3-им свойством ковариационной матрицы} =

$$V((X'X)^{-1}X'\varepsilon) =$$

= {воспользуемся 4-ым свойством ковариационной матрицы} =

$$(X'X)^{-1}X'V(\varepsilon)((X'X)^{-1}X')' =$$

= {воспользуемся предположениями №№4-5 КЛММР} =

$$(X'X)^{-1}X'(I_n * \sigma^2)X(X'X)^{-1} = \\ (X'X)^{-1}X'X(X'X)^{-1}\sigma^2 = (X'X)^{-1}\sigma^2$$

Таким образом, ковариационная матрица вектора оценок коэффициентов имеет вид:

$$V(\hat{\beta}) = (X'X)^{-1}\sigma^2$$

Обратите внимание: чтобы это равенство было корректно, требуется выполнение всех предположений КЛММР с первой по пятую. Например, при нарушении предположения о постоянстве дисперсии случайной ошибки ковариационная матрица вектора оценок коэффициентов будет иметь другой вид (мы обсудим этот вариант в главе 5).

Непосредственно эту матрицу на практике вычислить мы не можем, так как не знаем величину дисперсии случайной ошибки σ^2 . Однако мы можем получить несмещенную оценку этой матрицы, если заменим величину σ^2 на её несмещенную оценку

$$S^2 = \frac{1}{n-k} * \sum_{i=1}^n e_i^2.$$

В этом случае мы получим **оценку** ковариационной матрицы вектора оценок коэффициентов:

$$\hat{V}(\hat{\beta}) = (X'X)^{-1}S^2$$

Это таблица размера k на k , где на пересечении i -той строки и j -го столбца стоит несмещенная оценка коэффициента ковариации между $\hat{\beta}_i$ и $\hat{\beta}_j$. А на главной диагонали этой матрицы в j -м столбце стоит несмещенная оценка дисперсии коэффициента $\hat{\beta}_j$ — $\widehat{var}(\hat{\beta}_j)$. Корень из этой дисперсии — это стандартная ошибка оценки коэффициента

$$se(\hat{\beta}_j) = \sqrt{\widehat{var}(\hat{\beta}_j)}.$$

Далее мы увидим, что эта стандартная ошибка может использоваться, например, для тестирования незначимости соответствующего коэффициента.

Чтобы лучше разобраться во взаимосвязях между введенными нами векторами и матрицами, рассмотрим числовой пример.

Пример 3.1. Оценка параметров в модели множественной регрессии

Рассматривается классическая линейная модель множественной регрессии $y_i = \beta_1 + \beta_2 x_i^{(2)} + \beta_3 x_i^{(3)} + \varepsilon_i$. Имеются следующие данные о тысяче наблюдений соответствующих переменных:

$$\sum_{i=1}^{1000} x_i^{(2)} = 1000, \sum_{i=1}^{1000} x_i^{(3)} = 1000, \sum_{i=1}^{1000} x_i^{(2)} x_i^{(3)} = 1000, \\ \sum_{i=1}^{1000} (x_i^{(2)})^2 = 3000, \sum_{i=1}^{1000} (x_i^{(3)})^2 = 2000, \\ \sum_{i=1}^{1000} x_i^{(2)} y_i = 1000, \sum_{i=1}^{1000} x_i^{(3)} y_i = 2000, \sum_{i=1}^{1000} y_i = 0.$$

(а) Вычислите МНК-оценки коэффициентов модели.

(б) Пусть также известно, что сумма квадратов остатков в оцененной регрессии оказалась равна 997 000. Вычислите оценку ковариационной матрицы вектора оценок коэффициентов. Укажите, чему равна, например, оценка коэффициента ковариации между $\hat{\beta}_2$ и $\hat{\beta}_3$.

(в) Выпишите оцененное уравнение регрессии в стандартной форме, указав в скобках под оценками коэффициентов соответствующие стандартные ошибки.

Решение:

$$\begin{aligned}
 \text{(a) } \hat{\beta} = (X'X)^{-1}X'y = & \begin{pmatrix} n & \sum_{i=1}^{1000} x_i^{(2)} & \sum_{i=1}^{1000} x_i^{(3)} \\ \sum_{i=1}^{1000} x_i^{(2)} & \sum_{i=1}^{1000} (x_i^{(2)})^2 & \sum_{i=1}^{1000} x_i^{(2)} x_i^{(3)} \\ \sum_{i=1}^{1000} x_i^{(3)} & \sum_{i=1}^{1000} x_i^{(2)} x_i^{(3)} & \sum_{i=1}^{1000} (x_i^{(3)})^2 \end{pmatrix}^{-1} * \begin{pmatrix} \sum_{i=1}^{1000} y_i \\ \sum_{i=1}^{1000} x_i^{(2)} y_i \\ \sum_{i=1}^{1000} x_i^{(3)} y_i \end{pmatrix} = \\
 & \begin{pmatrix} 1000 & 1000 & 1000 \\ 1000 & 3000 & 1000 \\ 1000 & 1000 & 2000 \end{pmatrix}^{-1} * \begin{pmatrix} 0 \\ 1000 \\ 2000 \end{pmatrix} = \\
 & \frac{1}{1000} \begin{pmatrix} 2,5 & -0,5 & -1 \\ -0,5 & 0,5 & 0 \\ -1 & 0 & 1 \end{pmatrix} * \begin{pmatrix} 0 \\ 1000 \\ 2000 \end{pmatrix} = \begin{pmatrix} -2,5 \\ 0,5 \\ 2 \end{pmatrix}
 \end{aligned}$$

$$\text{(б) } S^2 = \frac{\sum e_i^2}{n-k} = \frac{997000}{1000-3} = 1000.$$

Оценка ковариационной матрицы вектора оценок коэффициентов:

$$\hat{V}(\hat{\beta}) = (X'X)^{-1} * S^2 = \begin{pmatrix} 2,5 & -0,5 & -1 \\ -0,5 & 0,5 & 0 \\ -1 & 0 & 1 \end{pmatrix}$$

Оценка коэффициента ковариации между $\hat{\beta}_2$ и $\hat{\beta}_3$ равна нулю, так как именно это число стоит на пересечении второй строки и третьего столбца в нашей матрице.

$$\begin{aligned}
 \text{(в) } \text{Стандартные ошибки оценок коэффициентов составляют } se(\hat{\beta}_1) &= \sqrt{2,5} = 1,58, se(\hat{\beta}_2) = \sqrt{0,5} = 0,71, \\
 se(\hat{\beta}_3) &= \sqrt{1} = 1.
 \end{aligned}$$

Теперь можно записать оцененное уравнение в стандартной форме:

$$\hat{y}_i = \underset{(1,58)}{-2,50} + \underset{(0,71)}{0,50}x_i^{(2)} + \underset{(1,00)}{2,00}x_i^{(3)}$$

Этими результатами мы воспользуемся в примере 3.5 в параграфе, посвященном тестированию гипотез.

3.4. Степень соответствия модели данным

Для множественной регрессии формула несмещенной оценки дисперсии случайной ошибки имеет вид

$$\widehat{\sigma^2} = S^2 = \frac{1}{n-k} * \sum_{i=1}^n e_i^2$$

Она почти такая же, как для парной регрессии за тем исключением, что в знаменателе вместо выражения $(n-2)$ стоит $(n-k)$. Если извлечь корень из этой величины, то можно получить стандартную ошибку регрессии

$$SEE = \sqrt{S^2} = \sqrt{\frac{1}{n-k} * \sum_{i=1}^n e_i^2}$$

Расчет стандартной ошибки регрессии — это один из способов оценить точность вашей модели в целом. То есть оценить, насколько хорошо она соответствует данным. Чем меньше стандартная ошибка регрессии, тем лучше ваша модель соответствует доступным вам наблюдениям.

Следующая характеристика качества подгонки — это коэффициент детерминации R^2 .

Для множественной регрессии с константой так же, как и для парной, верно, что общая сумма квадратов может быть представлена как сумма квадратов остатков и объясненная сумма квадратов:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n e_i^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

Поэтому и R^2 может быть рассчитан в точности таким же образом, как и для модели парной регрессии:

$$R^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\widehat{Var}(\hat{y})}{\widehat{Var}(y)}$$

И точно так же, как и в случае парной регрессии, он будет лежать между нулем и единицей. Если ваша модель хорошо соответствует данным, то R^2 будет близок к единице, если нет, то к нулю. Ещё раз подчеркнем, что условие

$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n e_i^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ выполняется только тогда, когда в модели есть константа. Если же ее нет, то указанное равенство, вообще говоря, неверно, и R^2 не обязан лежать между нулем и единицей, и интерпретировать стандартным образом его нельзя.

Некоторые эконометристы старой школы придают важное значение величине коэффициента R^2 . Действительно, если он близок к единице, то это, как правило, приятная новость. Однако не стоит переоценивать эту характеристику качества модели потому, что у коэффициента R^2 есть существенные ограничения:

1. Высокий R^2 характеризует наличие множественной корреляции между регрессорами и зависимой переменной, но **ничего не говорит о наличии или отсутствии причинно-следственной связи** между анализируемыми переменными. Вспомните примеры из первой главы, где мы обсуждали, что высокая корреляция не гарантирует причинно-следственной связи.
2. R^2 не может быть использован для принятия решения о том, стоит ли добавлять в модель новые переменные или нет. Дело в том, что, когда вы добавляете новые переменные в ваше уравнение, качество подгонки данных не может стать хуже, следовательно, и сумма квадратов остатков не может увеличиться. В теории она может остаться неизменной, но на практике она всегда будет уменьшаться. А в этом случае, как видно из расчетной формулы, R^2 будет увеличиваться. Получается, что какие бы дурацкие новые переменные вы ни добавляли в модель, коэффициент R^2 будет увеличиваться (или, в крайнем случае, оставаться неизменным).

Последний из указанных недостатков легко можно преодолеть. Для этого есть усовершенствованная версия R^2 , которую называют скорректированным (или нормированным) коэффициентом R^2 (R^2 adjusted):

$$R_{adj}^2 = R^2 - \frac{k-1}{n-k} * (1 - R^2)$$

R_{adj}^2 меньше, чем обычный R^2 , на величину $\frac{k-1}{n-k} * (1 - R^2)$, которая представляет собой штраф за добавление избыточных переменных. Обратите внимание, что при прочих равных этот штраф растет по мере увеличения параметра k , характеризующего число коэффициентов в вашей модели. Если вы будете добавлять в модель много регрессоров, которые не вносят существенного вклада в объяснение зависимой переменной, то R_{adj}^2 будет снижаться.

Поэтому, если вы хотите сравнить между собой модели с разным числом объясняющих переменных, то лучше использовать R_{adj}^2 , чем обычный R^2 . А ещё лучше обращать внимание не только на этот коэффициент, но и на прочие характеристики адекватности вашей модели, которые мы обсудим в этой книге.

Чтобы понять, откуда берется формула для скорректированного R -квадрата, запишем обычный R -квадрат следующим образом:

$$R^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\frac{\sum_{i=1}^n e_i^2}{n}}{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}}$$

В числителе дроби стоит выборочная дисперсия остатков, а в знаменателе — выборочная дисперсия зависимой переменной. Если и ту, и другую дисперсии заменить их несмещенными аналогами, то получим следующее выражение:

$$1 - \frac{S^2}{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}} = 1 - \frac{\frac{\sum_{i=1}^n e_i^2}{n-k}}{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}}.$$

Легко проверить, что это и есть скорректированный R-квадрат:

$$1 - \frac{\frac{\sum_{i=1}^n e_i^2}{n-k}}{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}} = 1 - \frac{n-1}{n-k} \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{n-1}{n-k} (1 - R^2) =$$

$$R^2 - \frac{k-1}{n-k} * (1 - R^2) = R_{adj}^2.$$

3.5. Тестирование гипотез и построение доверительных интервалов

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

Начнем с тестирования гипотез для отдельных коэффициентов. Пусть вы имеете дело с классической линейной моделью:

$$y_i = \beta_1 + \beta_2 x_i^{(2)} + \beta_3 x_i^{(3)} + \dots + \beta_k x_i^{(k)} + \varepsilon_i.$$

Если вас интересует, влияет ли регрессор $x^{(j)}$ на зависимую переменную, то для этого вам нужно осуществить тест на незначимость соответствующего коэффициента.

Процедура тестирования незначимости коэффициента в модели множественной регрессии:

1. Формулируем тестируемую гипотезу $H_0 : \beta_j = 0$ («переменная $x^{(j)}$ не влияет на переменную y ») и альтернативную гипотезу $H_1 : \beta_j \neq 0$ («переменная $x^{(j)}$ влияет на переменную y »)
2. Находим расчетное значение тестовой статистики по формуле $\frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$.
3. Выбираем уровень значимости α .
4. Из таблиц распределения Стьюдента находим критическое значение тестовой статистики t_{n-k}^α для выбранного уровня значимости и так называемого числа степеней свободы, которое в нашем случае равно $(n - k)$.
5. Если $\left| \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \right| > t_{n-k}^\alpha$, то есть $\hat{\beta}_j$ достаточно велик по абсолютной величине, следует отвергнуть гипотезу $H_0 : \beta_j = 0$ и сделать вывод в пользу альтернативной гипотезы, то есть заключить, что переменная $x^{(j)}$ влияет на переменную y . В этом случае переменную $x^{(j)}$ называют статистически значимой при уровне значимости α . В противном случае, соответственно, гипотеза H_0 не может быть отвергнута, и переменную $x^{(j)}$ называют статистически незначимой при уровне значимости α .

Легко заметить, что описанная процедура в целом такая же, как и для парной регрессии. Отличие состоит в том, что число степеней свободы теперь равно $(n - k)$. И в том, что оценки коэффициентов и их стандартные ошибки рассчитываются по формулам для множественной регрессии. Как именно их рассчитать, мы обсудили в параграфе 3.3, но на практике вы можете доверить эту скучную работу компьютеру.

Аналогичным образом можно тестировать гипотезу $H_0 : \beta_j = c$ (против альтернативной гипотезы $H_1 : \beta_j \neq c$), где c — это некоторая константа. В этом случае процедура тестирования остается такой же с одним исключением: расчетное значение тестовой статистики будет иметь вид $\frac{\hat{\beta}_j - c}{se(\hat{\beta}_j)}$.

Такая же аналогия с парной регрессией работает и при построении доверительных интервалов. Например, 95-процентный доверительный интервал для коэффициента β_j имеет вид:

$$\left(\hat{\beta}_j - se(\hat{\beta}_j) * t_{n-k}^{0,05}; \hat{\beta}_j + se(\hat{\beta}_j) * t_{n-k}^{0,05} \right)$$

Здесь $t_{n-k}^{0,05}$ — критическое значение распределения Стьюдента для уровня значимости 5% и $(n - k)$ степеней свободы.

Это распределение пригодится и для тестирования гипотез по поводу линейных комбинаций коэффициентов. Например, гипотезы следующего вида:

$$H_0 : a * \beta_1 + b * \beta_2 = c$$

В этом случае процедура тестирования снова остается такой же с одним исключением: расчетное значение тестовой статистики будет иметь вид

$$\frac{a * \hat{\beta}_1 + b * \hat{\beta}_2 - c}{\widehat{se}(a * \hat{\beta}_1 + b * \hat{\beta}_2)} = \frac{a * \hat{\beta}_1 + b * \hat{\beta}_2 - c}{\sqrt{\widehat{var}(a * \hat{\beta}_1 + b * \hat{\beta}_2)}} = \frac{a * \hat{\beta}_1 + b * \hat{\beta}_2 - c}{\sqrt{a^2 * \widehat{var}(\hat{\beta}_1) + b^2 * \widehat{var}(\hat{\beta}_2) + 2ab * \widehat{cov}(\hat{\beta}_1, \hat{\beta}_2)}}$$

Пример 3.2. Тестирование гипотез в модели множественной регрессии (продолжение примера 3.1)

Рассматривается классическая линейная модель множественной регрессии $y_i = \beta_1 + \beta_2 x_i^{(2)} + \beta_3 x_i^{(3)} + \varepsilon_i$. В ходе оценивания модели на основе данных по тысяче наблюдений при помощи МНК были получены следующие результаты:

$$\hat{y}_i = \underset{(1,58)}{-2,50} + \underset{(0,71)}{0,50} x_i^{(2)} + \underset{(1,00)}{2,00} x_i^{(3)}$$

Кроме того, вычислена оценка коэффициента ковариации между $\hat{\beta}_2$ и $\hat{\beta}_3$, которая составляет $\widehat{cov}(\hat{\beta}_2, \hat{\beta}_3) = 0$.

(а) Значима ли переменная $x^{(3)}$ (при уровне значимости 5%)?

(б) Проверьте гипотезу $\beta_2 + 2\beta_3 = 5$.

Решение:

(а) Критическое значение тестовой статистики из таблиц распределения Стьюдента при уровне значимости 5% и

$(1000 - 3) = 997$ степеней свободы составляет 1,96. Расчетное значение тестовой статистики: $\frac{\hat{\beta}_3}{se}$. Следует сделать вывод о том, что переменная $x^{(3)}$ статистически значима при уровне значимости 5%.

(б) $H_0 : 1*\beta_2 + 2*\beta_3 = 5$:

$$t_{расч} = \frac{1*\hat{\beta}_2 + 2*\hat{\beta}_3 - 5}{\widehat{se}(1*\hat{\beta}_2 + 2*\hat{\beta}_3)} = \frac{1*\hat{\beta}_2 + 2*\hat{\beta}_3 - 5}{\sqrt{\widehat{var}(1*\hat{\beta}_2 + 2*\hat{\beta}_3)}}$$

$$\frac{0,5 + 2*2 - 5}{\sqrt{\widehat{var}(\hat{\beta}_2) + 4*\widehat{var}(\hat{\beta}_3) + 4*\widehat{cov}(\hat{\beta}_2, \hat{\beta}_3)}} =$$

$$\frac{0,5 + 2*2 - 5}{\sqrt{(se(\hat{\beta}_2))^2 + 4*(se(\hat{\beta}_3))^2 + 4*\widehat{cov}(\hat{\beta}_2, \hat{\beta}_3)}} =$$

$$\frac{-0,5}{\sqrt{0,5 + 4*1 + 4*0}} = -0,236$$

Число $-0,236$ по модулю меньше, чем 1,96. Следовательно, тестируемая гипотеза **не** отвергается.

Во всех рассмотренных выше случаях мы тестировали гипотезу по поводу выполнения единственного линейного ограничения (например, $\beta_3 = 0$ или $\beta_2 + 2\beta_3 = 5$). Однако на практике часто возникает необходимость тестировать одновременное выполнение сразу нескольких ограничений.

Представим, например, что вы сначала оценили параметры регрессии, в которой есть m коэффициентов:

$$y_i = \beta_1 + \beta_2 * x_i^{(2)} + \dots + \beta_m * x_i^{(m)} + \varepsilon_i$$

После этого вы задались вопросом по поводу того, стоит ли добавить в эту модель ещё q новых переменных. То есть о том, стоит ли переходить вот к такому уравнению:

$$y_i = \beta_1 + \beta_2 * x_i^{(2)} + \dots + \beta_m * x_i^{(m)} +$$

$$+ \beta_{m+1} * x_i^{(m+1)} + \dots + \beta_{m+q} * x_i^{(m+q)} + \varepsilon_i$$

Для удобства назовем первую из двух моделей «короткой» (так как в ней меньше переменных), а вторую — «длинной» (так как в ней переменных больше). Как сделать выбор между этими моделями?

«Короткая» регрессия будет предпочтительной, если ни одна из добавленных в «длинную» регрессию переменных не является значимой. Иными словами, следует выбрать «короткую» регрессию, если верна следующая гипотеза:

$$H_0 : \beta_{m+1} = \dots = \beta_{m+q} = 0$$

То есть в том случае, если одновременно выполнено q ограничений.

Для этого можно использовать так называемый **F{-тест}**.

Процедура теста на сравнение «короткой» и «длинной» регрессий:

Оцените «короткую» регрессию, получите коэффициент R-квадрат из этой регрессии. Обозначим его R_R^2 (R^2 restricted, то есть R-квадрат в регрессии, для которой выполнено ограничение restriction).

Оцените «длинную» регрессию, получите коэффициент R-квадрат из этой регрессии. Обозначим его R_{UR}^2 (R^2 unrestricted).

Вычислите расчетное значение тестовой статистики:

$$F_{расч} = \frac{\frac{R_{UR}^2 - R_R^2}{1 - R_{UR}^2} * (n - k)}{q}$$

Здесь $k = m + q$ — количество коэффициентов в «длинной» регрессии.

Если верна нулевая гипотеза, то расчетное значение тестовой статистики имеет распределение Фишера с q и $(n - k)$ степенями свободы. Поэтому, если расчетное значение больше критического значения из таблиц распределения Фишера

$F_{расч} > F^{\alpha}(q, n - k)$, то тестируемая гипотеза отвергается при уровне значимости α , то есть следует сделать выбор в пользу «длинной регрессии».

Если же $F_{расч} \leq F^{\alpha}(q, n - k)$, то тестируемая гипотеза **не** отвергается при уровне значимости α , то есть следует сделать выбор в пользу «короткой» регрессии.

Таблицы распределения для осуществления всех тестов из этой главы содержатся в Приложении 3.А.

Важным частным случаем F-теста является ситуация, когда «короткая» регрессия включает в себя только константу $y_i = \beta_1 + \varepsilon_i$, а «длинная» регрессия по-прежнему содержит много переменных:

$$y_i = \beta_1 + \beta_2 x_i^{(2)} + \dots + \beta_k x_i^{(k)} + \varepsilon_i.$$

В этой ситуации, сравнение «короткой» и «длинной» регрессии состоит в проверке гипотезы $\beta_2 = \dots = \beta_k = 0$, то есть гипотезы о том, что **ни один** из регрессоров не влияет на зависимую переменную. Разумеется, если эта гипотеза не отвергается, то стоит заключить, что факторы для вашей модели мы выбрали скверные (раз уж ни один из них не помогает объяснить зависимую переменную). В таком случае уравнение называют в целом незначимым. А саму процедуру проверки гипотезы $\beta_2 = \dots = \beta_k = 0$ называют **тестом на незначимость уравнения в целом**.

Так как в уравнении, содержащем только константу, R-квадрат всегда равен нулю (см. задачу 9 из главы 2), то можно упростить формулу расчётного значения тестовой статистики. Подставим в неё 0 вместо величины R_R^2 , а величину R_{UR}^2 обозначим просто R^2 . Кроме того, не забудем, что в нашем случае $q = k - 1$. Получим:

$$F_{расч} = \frac{R^2}{1 - R^2} * \frac{n - k}{k - 1}.$$

Это и есть расчётное значение тестовой статистики для теста на незначимость уравнения в целом. Сравнивать его нужно с критическим значением из таблиц распределения Фишера $F^{\alpha}(k - 1, n - k)$.

Тест на сравнение «короткой» и «длинной» регрессий можно обобщить на случай сравнения **невложенных моделей**. Представим, например, что вам нужно сделать выбор между двумя такими моделями:

$$y_i = \beta_1 + \beta_2 x_i^{(2)} + \beta_3 x_i^{(3)} + \varepsilon_i (A)$$

и

$$y_i = \alpha_1 + \alpha_2 z_i^{(2)} + \alpha_3 z_i^{(3)} + \varepsilon_i (B)$$

Они называются **невложенными** (nonnested), так как по крайней мере некоторые переменные из модели **A** не входят в модель **B** и, наоборот, по крайней мере некоторые переменные из модели **B** не входят в модель **A**. Указанные модели не получится сопоставить, используя тест на сравнение «короткой» и «длинной» регрессий, так как ни одна из моделей не вложена в другую (то есть не является её частным случаем, «короткой» версией). Чтобы преодолеть это ограничение, можно прибегнуть к следующему трюку: рассмотреть новую модель, которая обобщает две предыдущих:

$$y_i = \beta_1 + \beta_2 x_i^{(2)} + \beta_3 x_i^{(3)} + \alpha_2 z_i^{(2)} + \alpha_3 z_i^{(3)} + \varepsilon_i$$

Для этой модели нужно провести два теста на сравнение «короткой» и «длинной» регрессий:

- Сначала проверить гипотезу о том, что незначимыми являются все переменные, которые входят в модель **A**, но не входят в модель **B**.
- Затем проверить гипотезу о том, что, наоборот, незначимыми являются все переменные, которые входят в модель **B**, но не входят в модель **A**.

Вывод на основе этого теста следует делать так:

- Если первая гипотеза будет отвергнута, а вторая — нет, то следует сделать выбор в пользу модели **A**.
- Если же, наоборот, вторая гипотеза будет отвергнута, а первая — нет, то следует сделать выбор в пользу модели **B**.

Конечно, возможна ситуация, в которой обе гипотезы будут отвергнуты. В этом случае следует сделать выбор в пользу наиболее общей объединенной модели. Если же не удастся отвергнуть ни одну из гипотез, то, по всей видимости, ни одна из моделей не является удовлетворительной.

Описанная процедура называется **тестом на сравнение невложенных моделей**.

Пример 3.3. Разные тесты для модели множественной регрессии

На основе 20 наблюдений была оценена следующая модель регрессии (в скобках указаны стандартные ошибки оценок коэффициентов):

$$\hat{y}_i = 2,4 + 6,9x_i + 5,1w_i$$

(0,6) (0,3) (9,8)

Кроме того, известно, что общая сумма квадратов равна 2000, а сумма квадратов остатков равна 200.

(а) Вычислите значение коэффициента R^2 , значение скорректированного коэффициента R_{adj}^2 и стандартную ошибку регрессии.

(б) Проверьте значимость уравнения в целом: сформулируйте и проверьте гипотезу о том, что все коэффициенты при переменных уравнения одновременно равны нулю.

(в) Значим ли коэффициент при переменной x ? Сформулируйте и проверьте соответствующую гипотезу.

(г) Проверьте гипотезу о том, что коэффициент при переменной x равен 7.

(д) Постройте 99-процентный доверительный интервал для коэффициента при переменной x .

(е) После того, как исследователь добавил в уравнение еще две переменные (назовём их p и s), R^2 в этой модели увеличился до 0,95. Осуществив соответствующий тест, определите, стоило ли добавлять в модель эти переменные?

Примечание: все гипотезы в этой задаче проверяйте при уровне значимости 1%.

Решение:

$$(a) R^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{200}{2000} = 0,9$$

$$R^2_{adj} = R^2 - \frac{k-1}{n-k} (1 - R^2) = 0,9 - \frac{3-1}{20-3} * (1 - 0,9) = 0,89$$

$$\text{Стандартная ошибка регрессии: } \sqrt{\frac{200}{20-3}} = 3,43$$

(б) Если обозначить коэффициенты в рассматриваемой модели стандартным образом: $y_i = \beta_1 + \beta_2 x_i + \beta_3 w_i + \varepsilon_i$, то тестируемая гипотеза может быть записана так:

$$H_0 : \beta_2 = \beta_3 = 0.$$

$$\text{Расчетное значение: } F = \frac{\frac{R^2}{1-R^2} * n - k}{k-1} = \frac{\frac{0,9}{0,1} * 17}{2} = 76,5.$$

Критическое значение при уровне значимости 1% $F(2, 17) = 6,11$.

$76,5 > 6,11$, поэтому тестируемая гипотеза отвергается. Следует сделать вывод о том, что уравнение в целом значимо.

(в) Проверяемая гипотеза: $H_0 : \beta_2 = 0$. Расчетное значение $6,9/0,3 = 23$. Критическое значение при уровне значимости 1% составляет $t(20-3) = 2,898$.

$23 > 2,898$, поэтому тестируемая гипотеза отвергается. Следует сделать вывод о том, что переменная значима.

(г) Проверяемая гипотеза: $H_0 : \beta_2 = 7$. Расчетное значение

$$\frac{6,9 - 7}{0,3} = -0,33.$$

Критическое значение при уровне значимости 1% составляет $t(20-3) = 2,898$. $0,3 < 2,898$ вывод: мы не можем отклонить гипотезу о том, что коэффициент β_2 равен 7.

(д) С вероятностью 99% $\beta_2 \in (6,9 - 0,3 * 2,898; 6,9 + 0,3 * 2,898)$

$$\beta_2 \in (6,03; 7,77).$$

(е) Если обозначить коэффициенты в новой модели стандартным образом: $y_i = \beta_1 + \beta_2 x_i + \beta_3 w_i + \beta_4 p_i + \beta_5 s_i + \varepsilon_i$, то тестируемая гипотеза может быть записана так:

$$H_0 : \beta_4 = \beta_5 = 0.$$

$$\text{Расчетное значение: } F = \frac{\frac{R^2_{UR} - R^2_R}{1 - R^2_{UR}} * n - k}{q} = \frac{\frac{0,95 - 0,9}{1 - 0,95} * 20 - 5}{2} = 7,5.$$

Критическое значение при уровне значимости 1% $F(2, 15) = 6,36$.

$7,5 > 6,36$, поэтому тестируемая гипотеза отвергается. «Длинная» регрессия значимо лучше, чем «короткая». То есть переменные добавлять стоило. Хотя, конечно, в реальных исследованиях лучше не оценивать уравнение с четырьмя переменными всего по 20 точкам.

3.6. Тест на линейное ограничение общего вида

F-тест из предыдущего параграфа можно использовать для тестирования гипотез не только о том, что группа коэффициентов одновременно равна нулю. Его можно обобщить для случая тестирования произвольного набора линейных ограничений на параметры модели.

Чтобы это сделать, удобно использовать матричную форму записи из параграфа 3.3.

Пусть тестируемая гипотеза имеет вид $H\beta = r$. Здесь β — по-прежнему обозначает вектор коэффициентов модели. H — матрица размера q на k , r — вектор-столбец длины q . q — количество тестируемых ограничений, то есть количество уравнений в системе ограничений, k — число коэффициентов в модели.

Например, если вы анализируете модель с тремя ($k = 3$) коэффициентами $y_i = \beta_1 + \beta_2 * x_i^{(2)} + \beta_3 * x_i^{(3)} + \varepsilon_i$ и хотите проверить одновременное выполнение двух ($q = 2$) вот таких ограничений:

$H_0 : \beta_2 = 9\beta_3$ и $\beta_1 + \beta_2 + \beta_3 = 5$, то матрица H и вектор r будут иметь следующий вид:

$$H = \begin{pmatrix} 0 & 1 & -9 \\ 1 & 1 & 1 \end{pmatrix}, r = \begin{pmatrix} 0 \\ 5 \end{pmatrix}$$

Расчетное значение тестовой статистики для тестирования гипотезы $H\beta = r$ имеет вид:

$$F_{расч} = \frac{(H\hat{\beta} - r)' (H(X'X)^{-1}H')^{-1} (H\hat{\beta} - r)/q}{\sum_{i=1}^n e_i^2 / (n - k)}$$

Если верна тестируемая гипотеза, то эта величина снова имеет распределение Фишера с q и $(n - k)$ степенями свободы.

3.7. Обобщающий пример

В заключение этой главы рассмотрим пример, позволяющий обобщить всё, что мы в ней выяснили.

Пример 3.4. Размер класса и эффективность обучения

Одним из дискуссионных вопросов в организации школьного образования является вопрос о том, действительно ли более эффективно учить школьников в маленьких классах (например, в классах по 10-15 человек) по сравнению с большими классами (например, в классах на 20-25 человек)? И если в маленьких классах школьники правда учатся лучше, то насколько велико это улучшение качества?

Вопрос важен в том числе и с экономической точки зрения. Действительно, если мы хотим улучшить качество школьного образования за счет уменьшения численности учеников в каждом классе, то нам придется нанять больше учителей, что потребует более значительных расходов на оплату их труда. Кроме того, мы столкнемся с издержками на организацию большего количества помещений, подходящих для обучения. Поэтому количественной оценке воздействия размера класса на эффективность обучения посвящен ряд исследований¹. В представленном примере вам также предлагается проанализировать массив данных, посвященный этому вопросу. В файле Students вам доступны следующие данные о двух сотнях школьников: CLASS — размер класса, в котором обучается школьник. Воздействие именно этой переменной на качество обучения будет интересоваться нас в этом примере; EXPN — средние расходы на одного школьника в школе, где он учится, измеренные в тысячах долларов в год; INCOME — средний доход на одного члена семьи в семье школьника, измеренный в тысячах долларов в год; TEST — результат итогового стандартизированного теста, который писали все школьники в конце учебного года. Эта переменная будет выступать в нашей регрессии в качестве зависимой переменной, так как она характеризует качество обучения (конечно, результаты тестов не являются совершенным измерителем качества обучения, однако в условиях отсутствия иных данных придется использовать их)

(а) Оцените параметры модели №1:

$$TEST_i = \beta_1 + \beta_2 CLASS_i + \varepsilon_i$$

Является переменная CLASS значимой? Интерпретируйте полученные результаты.

(б) Оцените параметры модели №2:

$$TEST_i = \beta_1 + \beta_2 CLASS_i + \beta_3 EXPN_i + \varepsilon_i$$

Является ли уравнение в целом статистически значимым?

Как изменилась оценка коэффициента при переменной CLASS. Чем можно объяснить такое изменение?

(в) Оцените параметры модели №3:

$$TEST_i = \beta_1 + \beta_2 CLASS_i + \beta_3 EXPN_i + \beta_4 INCOME_i + \varepsilon_i$$

Используя тест, на сравнение «короткой» и «длинной» регрессий сравните модели №1 и №3. Оправдано ли включение в модель новых переменных. Остается ли в новой модели переменная CLASS значимой? Как теперь можно интерпретировать коэффициент при этой переменной?

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

[Скрипт в R для этого примера](#)

[Файл с данными для этого примера.](#)

Решение:

(а) Результаты оценивания параметров модели №1 представлены ниже:

Модель 1: МНК, использованы наблюдения 1-200
Зависимая переменная: TEST

	Коэффициент	Ст. ошибка	t-статистика	P-значение	
const	83,6554	6,27855	13,32	2,47e-029	***
CLASS	-1,30921	0,280996	-4,659	5,82e-06	***
Среднее зав. перемен	55,18000	Ст. откл. зав. перемен	21,36774		
Сумма кв. остатков	81882,25	Ст. ошибка модели	20,33585		
R-квадрат	0,098804	Испр. R-квадрат	0,094252		
F(1, 198)	21,70799	P-значение (F)	5,82e-06		
Лог. правдоподобие	-885,2597	Крит. Акаике	1774,519		
Крит. Шварца	1781,116	Крит. Хеннана-Куинна	1777,189		

Мы видим, что переменная CLASS статистически значима при уровне значимости 1%, так как соответствующее P-значение меньше одной сотой². Коэффициент при этой переменной равен -1,3, что можно интерпретировать так: увеличение размера класса на одного ученика в среднем приводит к снижению результата школьника, который в этом классе учится, на 1,3 балла.

(б) Оценим теперь модель №2:

Модель 2: МНК, использованы наблюдения 1-200
Зависимая переменная: TEST

Коэффициент Ст. ошибка t-статистика P-значение

const	60,1560	7,97117	7,547	1,62e-012	***
CLASS	-0,908844	0,282915	-3,212	0,0015	***
EXPN	2,53712	0,566948	4,475	1,29e-05	***
Среднее зав. перемен	55,18000	Ст. откл. зав. перемен	21,36774		
Сумма кв. остатков	74326,57	Ст. ошибка модели	19,42401		
R-квадрат	0,181962	Испр. R-квадрат	0,173657		
F(2, 197)	21,91000	P-значение (F)	2,56e-09		
Лог. правдоподобие	-875,5783	Крит. Акаике	1757,157		
Крит. Шварца	1767,052	Крит. Хеннана-Куинна	1761,161		

Обратите внимание, что в этой таблице приведено расчетное значение F-статистики для проверки значимости уравнения в целом. И соответствующее P-значение:

F(2, 197) 21,91000 P-значение (F) 2,56e-09

Так как это P-значение меньше одной сотой, можно заключить, что уравнение в целом является значимым.

Заметим, что коэффициент при переменной CLASS по-прежнему значимый и отрицательный, однако по абсолютной величине он стал меньше. Теперь увеличение размера класса на единицу снижает результаты школьника за тест в среднем (при неизменных расходах на одного школьника) всего на 0,9 балла вместо прежних 1,3.

Такой результат легко объяснить, если вспомнить самое начало нашей главы. По всей видимости, переменная EXPN, которую мы добавили в модель, является существенной: она тоже значимо влияет на успехи школьника. Кроме того, она коррелирована с переменной CLASS. Если вычислить соответствующий выборочный коэффициент корреляции, то он окажется равен -0,31. В модели №1 эта переменная была пропущена, поэтому в модели №1 оценка коэффициента β_2 была смещена из-за пропуска существенной переменной.

Теперь мы добавили пропущенную переменную в модель и устранили указанное смещение. Поэтому оценка интересующего нас коэффициента изменилась. Следовательно, новая оценка (-0,9) вызывает больше доверия, чем предыдущая оценка (-1,3). Однако, возможно, мы по-прежнему упускаем что-то важное. Поэтому добавим в модель ещё один регрессор.

(в) Добавление в модель переменной INCOME позволяет получить следующие результаты:

Модель 3: МНК, использованы наблюдения 1-200
Зависимая переменная: TEST

	Коэффициент	Ст. ошибка	t-статистика	P-значение	
const	29,1786	4,84720	6,020	8,46e-09	***
CLASS	-1,07250	0,163173	-6,573	4,36e-010	***
EXPN	2,05022	0,327487	6,260	2,37e-09	***
INCOME	1,04899	0,0525994	19,94	4,69e-049	***
Среднее зав. перемен	55,18000	Ст. откл. зав. перемен	21,36774		
Сумма кв. остатков	24536,68	Ст. ошибка модели	11,18871		
R-квадрат	0,729949	Испр. R-квадрат	0,725816		
F(3, 196)	176,5965	P-значение (F)	1,84e-55		
Лог. правдоподобие	-764,7484	Крит. Акаике	1537,497		
Крит. Шварца	1550,690	Крит. Хеннана-Куинна	1542,836		

Третья модель также позволяет заключить, что размер класса значимо влияет на эффективность обучения. Теперь коэффициент при этой переменной можно интерпретировать так: увеличение размера класса на одного ученика в среднем приводит при прочих равных условиях к снижению результата школьника, который в этом классе учится, на 1,1 балла. Формулировка «при прочих равных» важна. В данном случае она означает, что мы сравниваем успехи школьников при прочих равных значениях двух других регрессоров: расходах на одного ученика и уровне благосостояния семьи школьника. Преимущество множественной регрессии как раз и состоит в том, что можно давать количественные оценки изменений при фиксированных значениях прочих важных факторов.

В таблице 3.1 содержатся сводные результаты оценки трех моделей. Обратите внимание, что при представлении результатов моделирования хорошим тоном является использование именно таких сводных таблиц, содержащих только необходимую информацию, а не необработанных таблиц, выданных эконометрическим пакетом, которые мы в учебных целях приводили выше.

Таблица 3.1. Влияние размера класса на качество обучения

	Модель 1	Модель 2	Модель 3
Константа	83,655***	60,156***	29,179***
	(6,279)	(7,971)	(4,847)
CLASS	-1,309***	-0,909***	-1,073***
	(0,281)	(0,283)	(0,163)
EXPN	—	2,537***	2,050***
		(0,567)	(0,327)
INCOME	—	—	1,049***
			(0,053)

Число наблюдений	200	200	200
R^2	0,099	0,182	0,730
Исправленный R^2	0,094	0,174	0,726

Примечания: зависимая переменная — балл за итоговый тест. В скобках под оценками коэффициентов указаны стандартные ошибки. ***, ** и * обозначают значимость на одно-, пяти- и десятипроцентном уровнях соответственно.

Сопоставим первую и третью модели, используя тест на сравнение «короткой» и «длинной» регрессии.

Нулевая гипотеза: $\beta_3 = \beta_4 = 0$.

Расчетное значение тестовой статистики составит:

$$F = \frac{\frac{R_{UR}^2 - R_R^2}{1 - R_{UR}^2} * n - k}{q} = \frac{\frac{0,730 - 0,099}{1 - 0,730} * 200 - 4}{2} = 229,0$$

Критическое значение из таблицы распределения Фишера для уровня значимости 1% и 2 и 196 степеней свободы примерно равно 4,6. Так как расчетное значение больше критического, мы отвергаем нулевую гипотезу и делаем вывод в пользу «длинной» регрессии. Таким образом, включение дополнительных переменных оправдано.

Аналогичный результат может быть получен в результате автоматического проведения теста в эконометрическом пакете:

Нулевая гипотеза: параметры регрессии нулевые

EXPN, INCOME

Тестовая статистика: $F(2, 196) = 229,039$, Р-значение 5,12105e-052

Легко видеть, что соответствующее Р-значение меньше одной сотой, следовательно, тестируемая гипотеза действительно должна быть отвергнута.

С точки зрения исправленного R-квадрата третья модель также лучше всех, так как там этот коэффициент принимает самое большое значение (напомним, что сравнивать модели с разным числом регрессоров корректно при помощи именно исправленного, а не обычного коэффициента R-квадрат).

1. См., например, Kreuger (1999) Experimental Estimates of Education Production Functions // The Quarterly Journal of Economics. [↗](#)
2. Запись «5,82e-06», используемая в таблице означает **$5,82 \cdot 10^{-6}$** , что заметно меньше, чем 0,01. [↗](#)

Приложение 3.А. таблицы распределения Стьюдента и Фишера
t-распределение: критические значения, двусторонний тест

Число степеней свободы	Уровень значимости	
	5%	1%
1	12,706	63,657
2	4,303	9,925
3	3,182	5,841
4	2,776	4,604
5	2,571	4,032
6	2,447	3,707
7	2,365	3,499
8	2,306	3,355
9	2,262	3,250
10	2,228	3,169
11	2,201	3,106
12	2,179	3,055
13	2,160	3,012
14	2,145	2,977
15	2,131	2,947
16	2,120	2,921
17	2,110	2,898
18	2,101	2,878
19	2,093	2,861
20	2,086	2,845
21	2,080	2,831
22	2,074	2,819
23	2,069	2,807
24	2,064	2,797
25	2,060	2,787
26	2,056	2,779
27	2,052	2,771
28	2,048	2,763
29	2,045	2,756
30	2,042	2,750
31	2,040	2,744
32	2,037	2,738
33	2,035	2,733

34	2,032	2,728
35	2,030	2,724
36	2,028	2,719
37	2,026	2,715
38	2,024	2,712
39	2,023	2,708
40	2,021	2,704
41	2,020	2,701
42	2,018	2,698
43	2,017	2,695
44	2,015	2,692
45	2,014	2,690
46	2,013	2,687
47	2,012	2,685
48	2,011	2,682
49	2,010	2,680
50	2,009	2,678
60	2,000	2,660
70	1,994	2,648
80	1,990	2,639
90	1,987	2,632
100	1,984	2,626
120	1,980	2,617
∞	1,960	2,576

F-распределение: критические значения F с v_1 и v_2 степенями свободы при уровне значимости 5%

$v_2 \setminus v_1$	1	2	3	4	5	6	7	8	9	10
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85

12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99
120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83

F-распределение: критические значения F с v_1 и v_2 степенями свободы при уровне значимости 1%

$v_2 \backslash v_1$	1	2	3	4	5	6	7	8	9	10
1	4052	5000	5403	5625	5764	5859	5928	5981	6022	6056
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30

13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47
∞	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32

Глава 4. Множественная регрессия: мультиколлинеарность, фиктивные переменные, нелинейные модели

В этой главе обсуждаются важные с прикладной точки зрения аспекты, касающиеся множественной регрессии. Речь будет идти о мультиколлинеарности, фиктивных переменных и нелинейных моделях. Кроме того, в заключительной части главы мы обсудим правила хорошего тона, которым стоит следовать при оформлении результатов ваших собственных эконометрических работ.

[4.1. Мультиколлинеарность](#)

[4.2. Фиктивные переменные](#)

[4.3. Нелинейные модели](#)

[4.4. Обобщающий пример](#)

[4.5. Рекомендации по оформлению результатов эконометрических расчетов](#)

[Задания для самостоятельного решения](#)

4.1. Мультиколлинеарность

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

Выделяют два вида мультиколлинеарности:

- строгая мультиколлинеарность (её также называют полной или точной),
- нестрогая мультиколлинеарность (её также называют частичной).

Под строгой мультиколлинеарностью понимается ситуация, когда между регрессорами в модели есть точная линейная связь, т. е. когда одна объясняющая переменная точным образом линейно выражается через другие.

Представим, например, что при анализе макроэкономической модели в качестве переменных в неё включили экспорт, импорт и чистый экспорт. Чистый экспорт равен разности между экспортом и импортом и, следовательно, при включении в модель этих трёх переменных окажется, что регрессоры модели линейно выражаются друг через друга.

В терминах матричной записи точная мультиколлинеарность предполагает линейную зависимость столбцов матрицы регрессоров, откуда следует неполный ранг матрицы регрессоров. Это означает, что при полной мультиколлинеарности невозможно вычислить МНК-оценки коэффициентов, потому что матрица $\mathbf{X}'\mathbf{X}$ является вырожденной, и матрица $(\mathbf{X}'\mathbf{X})^{-1}$ не определена.

Из определения и из приведенного выше примера легко догадаться, как можно решить проблему строгой мультиколлинеарности. Для этого следует исключить лишнюю переменную. Например, если в модели уже учтены экспорт и импорт, то понятно, что включение ещё и чистого экспорта не принесет никакой дополнительной информации, и этой третьей переменной можно безболезненно пожертвовать.

Современные эконометрические пакеты при возникновении чистой мультиколлинеарности сами избавляются от одной из линейно зависимых переменных, чтобы вычисление МНК-оценок стало технически возможным.

Частичная мультиколлинеарность — это ситуация, когда между объясняющими переменными нет точной линейной связи, но эти переменные сильно коррелируют между собой. Иными словами, они не линейно зависимы, но «почти» линейно зависимы. При частичной мультиколлинеарности вычислить МНК-оценки можно, однако стандартные ошибки оценок коэффициентов оказываются высокими, а точность оценок коэффициентов — низкой. Так происходит потому, что при сильной корреляции двух регрессоров в выборке они, как правило, меняются одновременно, и оказывается трудно отличить влияние одного регрессора на зависимую переменную от влияния другого. Таким образом, основным негативным последствием мультиколлинеарности является снижение точности оценки отдельных коэффициентов.

Частичная мультиколлинеарность не нарушает ни одну из предпосылок классической линейной модели множественной регрессии и поэтому не приводит к смещению оценок коэффициентов модели.

Это особенно хорошая новость потому, что на практике почти в любой множественной регрессии объясняющие переменные в той или иной степени коррелированы. Поэтому частичная мультиколлинеарность в данных наблюдается очень часто. Представьте, например, что вы моделируете выпуск фирмы некоторой отрасли в зависимости от количества используемых фирмой труда и физического капитала (то есть, как сказали бы экономисты, моделируете производственную функцию). Скорее всего, в вашей выборке будут большие фирмы и маленькие, причем большие фирмы в среднем будут использовать относительно много каждого из факторов производства, а маленькие, напротив, относительно мало. В результате переменные, характеризующие количество труда и количество капитала, будут положительно коррелированы друг с другом.

Есть несколько способов выявить мультиколлинеарность на этапе предварительного анализа данных (то есть ещё до оценки параметров уравнения регрессии). О наличии существенной частичной мультиколлинеарности в модели говорят:

1. большие по абсолютной величине (больше 0,9) парные коэффициенты корреляции между регрессорами;
2. близость к нулю определителя матрицы $\mathbf{X}'\mathbf{X}$;
3. большие (больше 10) значения коэффициентов VIF.

Коэффициенты VIF (variance inflation factor) показывают, насколько сильно связаны друг с другом регрессоры модели. Чтобы определить коэффициент VIF, соответствующий регрессору $\mathbf{x}^{(j)}$, нужно оценить вспомогательную регрессию, в которой слева стоит $\mathbf{x}^{(j)}$, а справа — все остальные объясняющие переменные исходной модели. После этого нужно вычислить коэффициент VIF по формуле:

$$VIF = \frac{1}{1-R^2},$$

где R^2 — это коэффициент детерминации из оцененной вспомогательной регрессии.

Если коэффициенты VIF для всех регрессоров оказались меньше 10, это значит, что существенной мультиколлинеарности в модели не наблюдается. В противном случае стоит сделать вывод о том, что в модели есть мультиколлинеарность.

Пример 4.1. Мультиколлинеарность.

Эконометрист исследует модель:

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 w_i + \beta_4 z_i + \varepsilon_i$$

На этапе предварительного анализа данных он оценил следующие вспомогательные уравнения:

$$\hat{x}_i = 10,1 + 1,9w_i + 2,3z_i, R^2 = 0,95$$

$$\hat{w}_i = 18,7 + 0,8x_i + 4,8z_i, R^2 = 0,99$$

$$\hat{z}_i = -5,0 + 0,1w_i + 0,3x_i, R^2 = 0,20$$

Что можно сказать о наличии мультиколлинеарности в исходной модели?

Решение:

Коэффициенты VIF для переменных x, w, z равны, соответственно: $\frac{1}{1-0,95} = 20$, $\frac{1}{1-0,99} = 100$ и $\frac{1}{1-0,2} = 1,25$. Так как некоторые из коэффициентов больше 10, можно заключить, что в модели присутствует существенная мультиколлинеарность.

Некоторые признаки мультиколлинеарности можно увидеть уже после оценки параметров модели. Перечислим их:

— **Неустойчивость результатов.** Небольшое изменение исходных данных приводит к существенному изменению оценок коэффициентов. Например, если после оценки уравнения по 200 наблюдениям вы исключили из выборки несколько точек, оценили модель заново и обнаружили сильное изменение результатов.

— **Незначимость большинства переменных.** Каждая переменная в отдельности является незначимой, а уравнение в целом является значимым и характеризуется близким к единице коэффициентом R^2 .

— **Неправдоподобность результатов.** Оценки коэффициентов имеют неправильные с точки зрения экономической теории знаки или неоправданно большие значения. Стоит отметить, что причиной возникновения такой ситуации может быть не только потеря точности оценивания из-за мультиколлинеарности, но и гораздо более серьезные проблемы, например, смещение из-за пропуска существенной переменной (не говоря уж о том, что не все экономические теории прошлого проходят испытание современными данными).

Что можно предпринять, если вы столкнулись с негативными последствиями мультиколлинеарности в вашей модели? Существует несколько путей решения этой проблемы.

Если есть возможность увеличить количество наблюдений, то это отличный вариант, так как больший размер выборки увеличит точность результатов, компенсировав её потерю из-за мультиколлинеарности.

Мультиколлинеарность будет устранена, если вы исключите из уравнения тот регрессор, который сильно коррелирован с остальными объясняющими переменными модели. Однако следует помнить, что применение этого способа не всегда целесообразно, так как может привести к гораздо более серьезным последствиям: смещению оценок в результате пропуска существенной переменной. Скажем, в нашем примере про производственную функцию ни труд, ни капитал из уравнения исключать не хотелось бы, так как ясно, что выпуск фирмы зависит от каждого из этих факторов производства.

Решением проблемы может быть использование вместо отдельных переменных их линейных комбинаций. Возвращаясь к нашему примеру с экспортом и импортом, заметим, что эти переменные, включенные в модель по отдельности, могут быть причиной мультиколлинеарности, так как обычно коррелированы друг с другом. Однако, заменив их чистым экспортом (который как раз и представляет собой их линейную комбинацию), вы сможете избежать этой проблемы. Другой пример: представим, что вы оцениваете зависимость успеваемости студента физического факультета от баллов за ЕГЭ по математике и по физике, которые этот студент получил, будучи школьником. Так как два этих регрессора наверняка коррелированы, то вместо включения в модель каждого из них по отдельности вы могли бы оставить в уравнении одну переменную — средний балл ЕГЭ по двум этим предметам.

Использование альтернативных (нелинейных) форм зависимостей в некоторых случаях также может снизить остроту проблемы мультиколлинеарности. Оценивание такого рода моделей мы обсудим в конце данной главы.

В заключение ещё раз подчеркнем, что мультиколлинеарность сама по себе не вызывает смещения оценок коэффициентов. Поэтому бороться с ней нужно только в том случае, если она приводит к существенным проблемам (например, к огромным стандартным ошибкам оценок коэффициентов или заведомой неадекватности полученных результатов). Во всех остальных случаях данную проблему можно игнорировать¹.

1. Некоторые специфические инструменты, которые иногда тоже могут быть полезны в борьбе с мультиколлинеарностью, обсуждаются в рамках курсов машинного обучения и многомерного статистического анализа. См.: метод главных компонент, LASSO и ridge-регрессии, метод эластичной сети. [↪](#)

4.2. Фиктивные переменные

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

Иногда в процессе эконометрического моделирования у исследователя возникает потребность учитывать в качестве объясняющих факторов не только количественные, но и качественные характеристики. Например, на цену квартиры могут влиять не только её жилая площадь и расстояние до ближайшего метро (количественные переменные), но и материал, из которого изготовлен дом, или наличие в этой квартире балкона (качественные переменные). На величину заработной платы работника могут влиять не только его стаж работы (количественный признак), но и факт наличия у него высшего образования или пол (качественные признаки). Во всех этих случаях удобно использовать так называемые фиктивные переменные.

Фиктивные переменные — это такие переменные, которые принимают одно из двух значений — 0 или 1. Их также называют бинарными или дамми-переменными (dummy variable).

Представим, например, что заработная плата описывается следующим уравнением, для которого выполнены все предпосылки классической линейной модели множественной регрессии:

$$Y_i = \beta_1 + \beta_2 * X_i + \beta_3 * D_i + \varepsilon_i,$$

где Y_i — зарплата i -го работника в долларах в час, X_i — стаж работы i -го работника в годах, D_i — фиктивная переменная, которая равна единице, если i -ый работник — женщина, и равна нулю, если мужчина. Исследователь включает в модель эту переменную потому, что подозревает наличие на рассматриваемом рынке труда дискриминации по гендерному признаку.

В результате МНК-оценивания параметров модели на основе данных о 1000 работниках исследователь получил следующее уравнение:

$$\hat{Y}_i = \underset{(0,3)}{4,2} + \underset{(0,1)}{2,1}X_i - \underset{(0,2)}{3,5}D_i$$

Результаты построения модели с фиктивной переменной удобно интерпретировать, если записать её для двух случаев: когда фиктивная переменная равна 0 и когда она равна 1. В нашем примере это приведёт к двум вот таким уравнениям:

$$\text{мужчины } (D_i = 0): Y_i = 4,2 + 2,1 * X_i,$$

$$\text{женщины } (D_i = 1): Y_i = 4,2 + 2,1 * X_i - 3,5 = 0,7 + 2,1 * X_i.$$

Отсюда видно, что при прочих равных условиях (при равном стаже работы) женщины получают на 3,5 доллара меньше, чем мужчины. Подчеркнем, что оценка такой модели гораздо лучше, чем просто сравнение средней по выборке заработной платы мужчин со средней по выборке заработной платой женщин, так как гипотетически различие между этими средними могло бы объясняться не гендерной дискриминацией, а разным стажем работы у мужчин и женщин. Мы же в нашем примере контролируем это различие, включая стаж работы в модель. В реальном исследовании, разумеется, было бы целесообразно включить в модель и прочие факторы, которые могут влиять на заработную плату (скажем, образование), однако нам для целей объяснения идеи фиктивных переменных пока хватит этого упрощенного примера.

Графически полученные уравнения представлены на рисунке 4.1. Мы видим, что наша фиктивная переменная отражает сдвиг линии, характеризующей зависимость заработной платы от стажа работы. Поэтому фиктивные переменные такого сорта иногда называют фиктивными переменными сдвига.

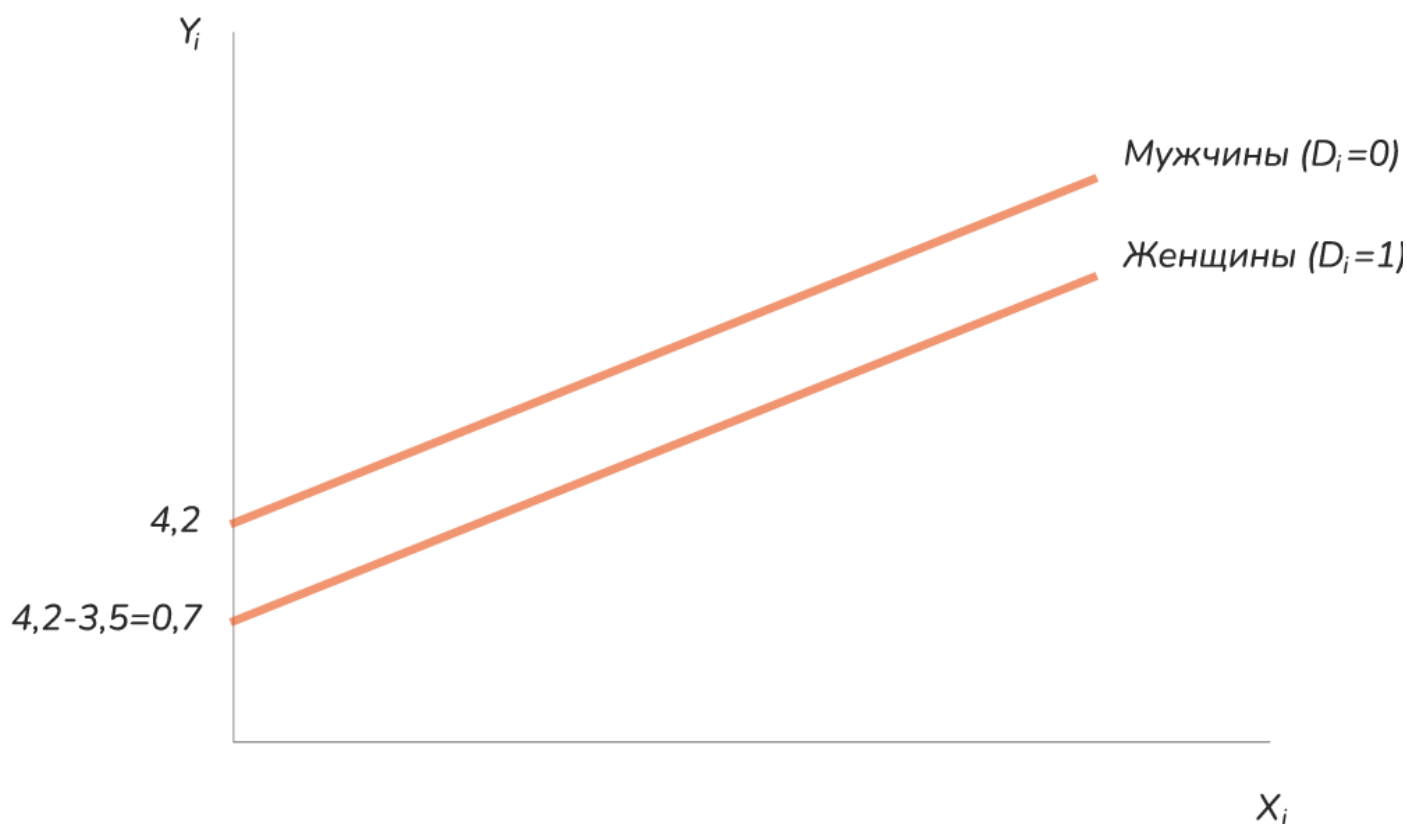


Рисунок 4.1. Зависимость между стажем работы и заработной платой для мужчин и для женщин в случае использования фиктивной переменной сдвига.

У модели, которую оценил наш исследователь, есть важное ограничение. Мы видим, что женщины получают «штраф» к зарплате в размере 3,5 долларов. Причем этот «штраф» фиксирован и не зависит от стажа работы. В действительности, возможна ситуация, когда с ростом опыта работы зарплата у мужчин растет быстрее, чем у женщин. Иными словами, разрыв между зарплатами мужчин и женщин может становиться больше по мере увеличения стажа работы. Чтобы выявить подобную тенденцию, нам потребуется новый вид фиктивных переменных — фиктивные переменные наклона. Это произведение переменных X_i и D_i .

Включив такое произведение в модель, исследователь получит следующее уравнение:

$$Y_i = \beta_1 + \beta_2 * X_i + \beta_3 * D_i + \beta_4 * X_i D_i + \epsilon_i$$

И снова, чтобы понять, как его интерпретировать, удобно переписать уравнение отдельно для женщин и мужчин:

мужчины ($D_i = 0$): $Y_i = \beta_1 + \beta_2 * X_i + \epsilon_i$,

женщины ($D_i = 1$): $Y_i = (\beta_1 + \beta_3) + (\beta_2 + \beta_4) * X_i + \epsilon_i$.

Получается, что в «мужской» модели коэффициент перед переменной X равен β_2 , то есть каждый дополнительный год стажа увеличивает зарплату мужчины на β_2 . А для женщин коэффициент при X равен $\beta_2 + \beta_4$, то есть каждый дополнительный год стажа работы увеличивает зарплату на $\beta_2 + \beta_4$. И если, например, $\beta_4 < 0$, то это означает, что наклон линии регрессии для женщин более пологий, чем для мужчин (см. рис. 4.2), то есть каждый дополнительный год опыта работы дает женщинам меньшую прибавку к зарплате по сравнению с мужчинами.

β_1

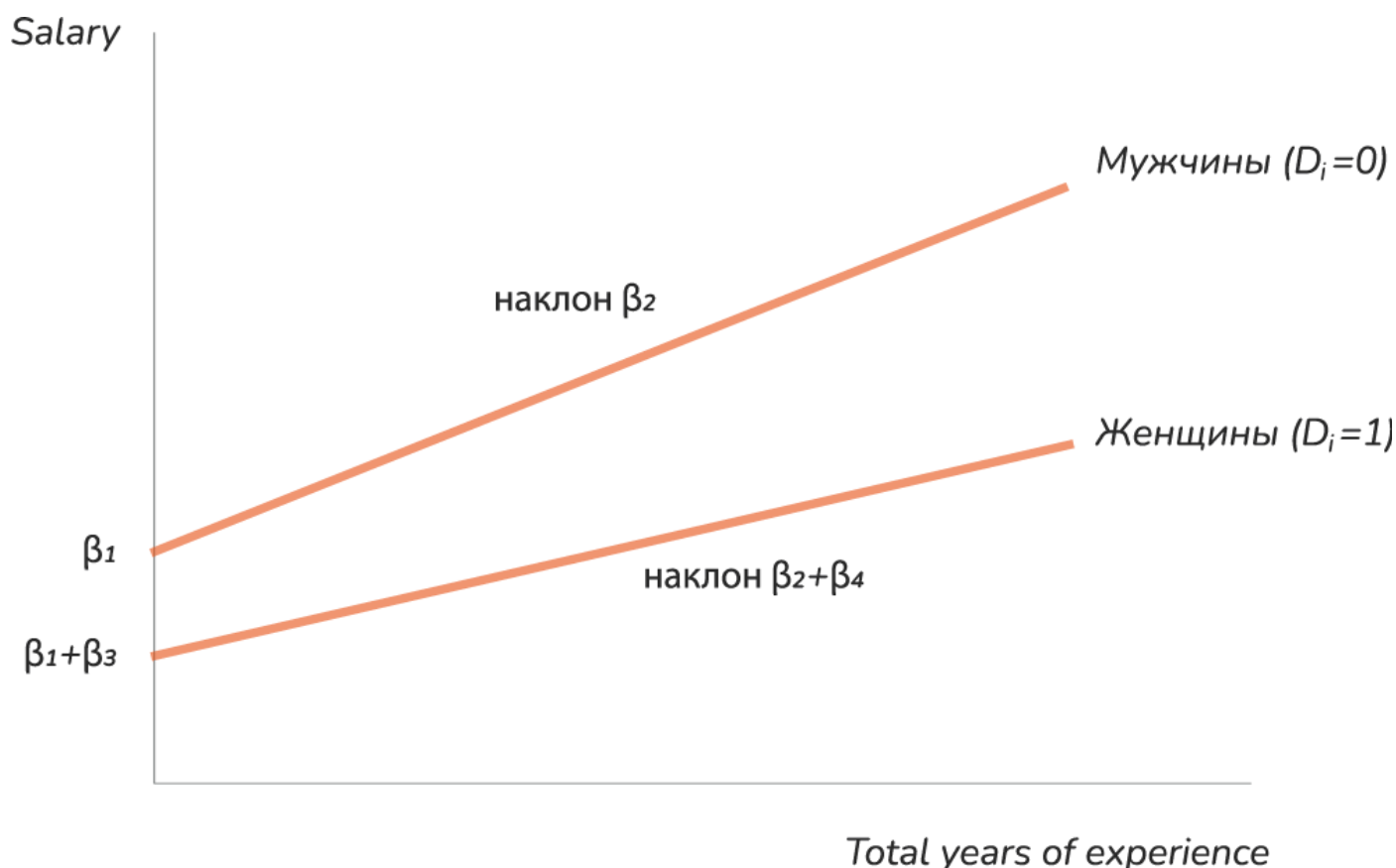


Рисунок 4.2. Зависимость между стажем работы и заработной платой для мужчин и для женщин в случае использования фиктивных переменных сдвига и наклона ($\beta_3 < 0, \beta_4 < 0$).

Фиктивные переменные могут помочь выявить структурные различия в моделях для разных подвыборок. В нашем примере мы можем проверить наличие или отсутствие структурных различий в моделях заработной платы для мужчин и женщин. Для этого достаточно проверить гипотезу

$$\beta_3 = \beta_4 = 0.$$

Действительно, легко видеть, что если эта гипотеза верна, то уравнения заработной платы для мужчин и для женщин являются одинаковыми. Чтобы тестировать эту гипотезу, следует осуществить уже знакомый нам тест для сравнения «короткой» и «длинной» регрессий. Применительно к фиктивным переменным этот тест иногда называют тестом Чоу или тестом на структурный сдвиг. Он устроен следующим образом: необходимо добавить в модель фиктивную переменную сдвига и все соответствующие фиктивные переменные наклона, а затем тестировать гипотезу о том, что коэффициенты при этой фиктивной переменной сдвига и всех фиктивных переменных наклона одновременно равны нулю.

Пример 4.2. Тест на структурный сдвиг.

Опираясь на одну и ту же выборку из 1000 работников, исследователь оценил параметры двух моделей:

$$Y_i = \beta_1 + \beta_2 * X_i + \varepsilon_i$$

$$Y_i = \beta_1 + \beta_2 * X_i + \beta_3 * D_i + \beta_4 * X_i D_i + \varepsilon_i$$

В первой модели R-квадрат оказался равен 0,6, а во второй — 0,8. Осуществите тест на структурный сдвиг и интерпретируйте его результаты.

Решение:

Нужно тестировать гипотезу $\beta_3 = \beta_4 = 0$ против альтернативной гипотезы о том, что хотя бы один из двух указанных коэффициентов отличен от нуля.

Расчетное значение тестовой статистики может быть определено по формуле:

$$F_{расч} = \frac{\frac{R_{UR}^2 - R_R^2}{1 - R_{UR}^2} * n - k}{q} = \frac{\frac{0,8 - 0,6}{1 - 0,8} * 1000 - 4}{2} = 498$$

Это больше, чем критическое значение соответствующей тестовой статистики при любом разумном уровне значимости. Например, при уровне значимости 1% $F(q, n - k) = F(2, 996) = 4,61$. Поэтому нулевая гипотеза должна быть отвергнута, и следует сделать вывод о наличии структурного сдвига между моделями заработной платы для мужчин и для женщин. Иными словами, сделать вывод о том, что в рассматриваемой отрасли присутствует дискриминация по гендерному признаку.

В рассмотренном нами примере качественный признак может принимать два возможных значения: работник является либо мужчиной, либо женщиной. При помощи фиктивных переменных можно анализировать и случаи большего количества возможных

значений.

Представим, что мы в качестве моделируемого признака рассматриваем университет, который закончил работник, и что в выборке есть выпускники ровно трёх университетов: А, В и С (и нет работников, которые не закончили никакого университета). Ясно, что одной бинарной переменной нам уже не хватит, и этот качественный признак нужно закодировать каким-то другим образом. Оказывается, это просто сделать, добавив в модель не одну, а две фиктивные переменные. Тогда уравнение будет выглядеть вот так:

$$Y_i = \beta_1 + \beta_2 * X_i + \beta_3 * A_i + \beta_4 * B_i + \varepsilon_i,$$

где A_i — фиктивная переменная, которая равна 1, если i -ый респондент является выпускником вуза А, и равна 0 в противном случае, B_i — аналогичная переменная для вуза В.

Конечно, есть некоторый соблазн добавить три фиктивные переменные: по одной для каждого университета. Действительно, почему бы не добавить в модель фиктивную переменную C_i , равную единице для выпускников университета С и нулю для выпускников остальных университетов? Ответ на этот вопрос дает нам первая часть данной главы.

Дело в том, что если мы добавим 3-ю фиктивную переменную, то мы столкнемся с чистой мультиколлинеарностью. Если i -ый работник закончил университет В, тогда для него $A_i = 0, B_i = 1, C_i = 0$ и, следовательно

$$A_i + B_i + C_i = 1.$$

Аналогично для работника, закончившего любой университет (то есть для каждого работника в нашей выборке), сумма трех указанных переменных будет равна единице. Тем самым наблюдается строгая линейная связь между переменными модели, что соответствует определению строгой мультиколлинеарности. Поэтому оценка модели, включающей константу и три этих переменных, невозможна. Ситуация возникновения чистой мультиколлинеарности из-за добавления в модель избыточного количества фиктивных переменных называется ловушкой фиктивных переменных. Избежать этой ловушки легко: нужно добавлять в модель на одну переменную меньше, чем есть значений признака. То есть, если моделируемый признак принимает m возможных значений, то для его описания в уравнение следует добавить $m - 1$ фиктивную переменную¹.

Представим, что мы в нашем примере ограничились двумя фиктивными переменными, собрали данные о трех тысячах выпускников и, проведя необходимые расчеты, получили следующие оценки параметров:

$$\hat{Y}_i = 5,2 + 1,1X_i + 2,0A_i + 3,0B_i$$

(0,5) (0,2) (0,1) (0,2)

Как интерпретировать полученные оценки коэффициентов? Снова запишем модель для каждого типа выпускников отдельно:

$$\text{вуз А } (A_i = 1, B_i = 0): \hat{Y}_i = 5,2 + 1,1 * X_i + 2,0,$$

$$\text{вуз В } (A_i = 0, B_i = 1): \hat{Y}_i = 5,2 + 1,1 * X_i + 3,0,$$

$$\text{вуз С } (A_i = 0, B_i = 0): \hat{Y}_i = 5,2 + 1,1 * X_i.$$

Коэффициент при фиктивной переменной А, оценка которого равна 2, означает, что при прочих равных условиях выпускник вуза А зарабатывает на 2 доллара в час больше, чем выпускник вуза С. Важно помнить, что, когда мы интерпретируем коэффициент, мы должны не просто говорить, что кто-то зарабатывает больше, а указывать, по сравнению с кем больше. В данном случае фраза «выпускник вуза А в среднем получает на 2 доллара в час больше, чем выпускник вуза С» — это корректная фраза. А фраза «выпускник вуза А получает на 2 доллара больше, чем выпускники других вузов» — это некорректная фраза, так как в модели видно, что выпускник вуза А по сравнению с выпускниками вуза В получает не больше, а меньше.

Обычно в качестве базы для сравнения (или так называемой эталонной категории) выступает та категория, для которой мы не стали добавлять фиктивную переменную. В нашем примере эталонным университетом выступает вуз С (эталонным не в том смысле, что он самый хороший, а в том смысле, что с ним все сравнивается).

Еще раз подчеркнем, что подобная содержательная интерпретация коэффициентов осмыслена только в том случае, если эти коэффициенты статистически значимы. Если же они статистически не значимы, то у нас нет уверенности в том, что они отличаются от нуля, и это должно отразиться на наших выводах. Например, если бы в уравнении выше стандартная ошибка оценки коэффициента при переменной A_i была бы равна не 0,1, а 10,0, то расчетное значение соответствующей t -статистики оказалось бы равно $2,0/10,0=0,2$, что меньше критического значения при любом разумном уровне значимости. Следовательно, мы не могли бы отвергнуть гипотезу о том, что $\beta_3 = 0$, и должны были бы заключить, что различия в заработных платах между выпускниками вузов А и С отсутствуют.

1. Или добавить все m фиктивных переменных, но тогда не добавлять константу. Этот вариант менее удобен для содержательной интерпретации результатов, поэтому используется сравнительно редко. [↗](#)

4.3. Нелинейные модели

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

До этого момента мы концентрировались на линейных зависимостях между переменными. Однако мир многообразен, и в процессе эконометрического моделирования часто приходится сталкиваться с нелинейными взаимосвязями. Обычно их использование мотивируется одной из двух причин:

1. **Графическим анализом данных.** Например, если на этапе предварительного исследования данных вы построили график, характеризующий взаимосвязь между вашими переменными, и увидели нечто подобное рисунку 4.3, то, скорее всего, вы придете к выводу, что связь нелинейна.
2. **Содержательными теоретическими соображениями.** К примеру, если вы анализируете зависимость между количеством используемых фирмой факторов производства и её объемом выпуска, вы наверняка захотите проверить, не описывается ли эта взаимосвязь производственной функцией Кобба — Дугласа, которая имеет следующий (нелинейный) вид: $Y_i = AK_i^\alpha L_i^\beta$.

Рассмотрим несколько нелинейных моделей, которые часто встречаются в эконометрическом анализе.

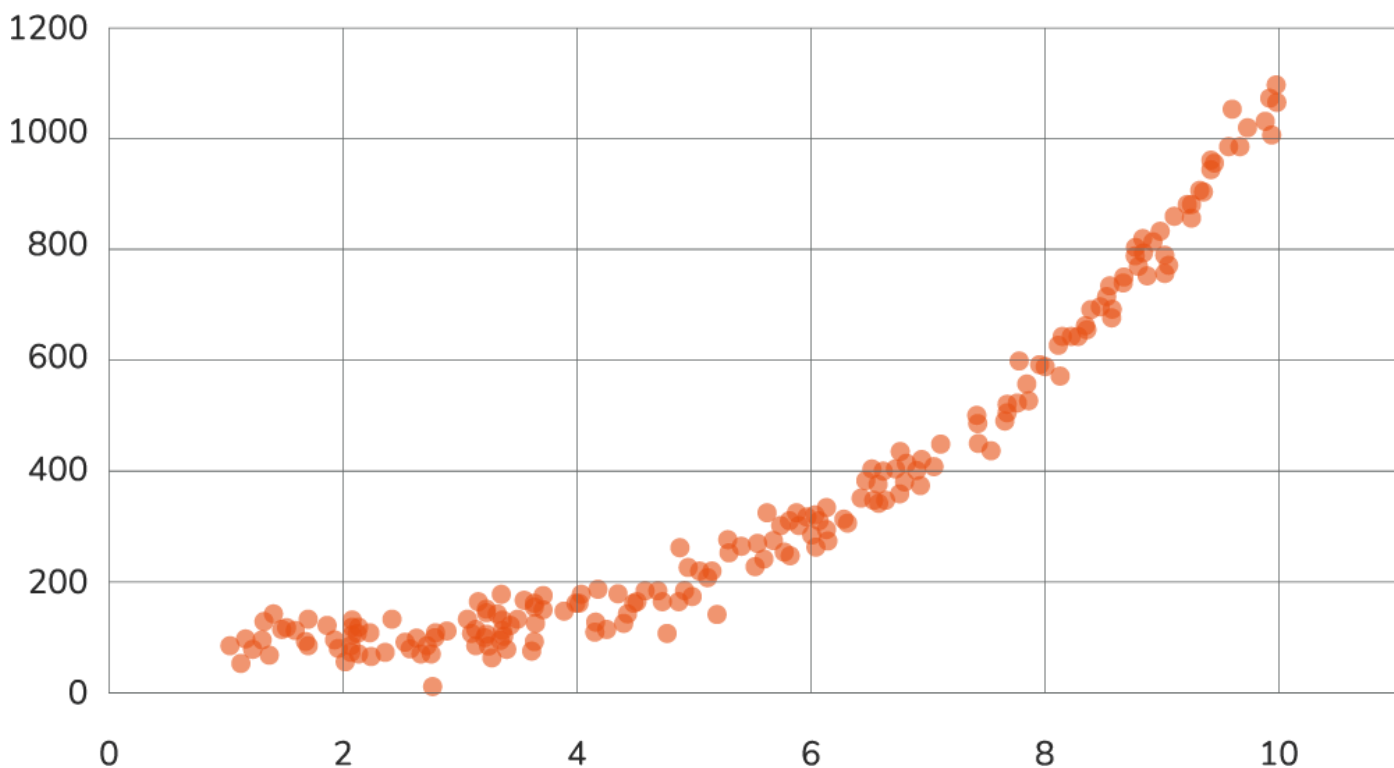


Рисунок 4.3. Пример нелинейной связи между переменными

Логарифмическая модель

Во многих случаях зависимость между переменными в экономике носит степенной характер:

$$y_i = Ax_i^a$$

Прологарифмируем правую и левую части этого равенства:

$$\ln y_i = \ln A + a \ln x_i$$

Переобозначим переменные более привычным нам образом: $\beta_1 = \ln A$ и $\beta_2 = a$:

$$\ln y_i = \beta_1 + \beta_2 \ln x_i$$

Наконец, чтобы сделать модель пригодной для эконометрического моделирования, добавим в неё случайные ошибки:

$$\ln y_i = \beta_1 + \beta_2 \ln x_i + \varepsilon_i$$

Поскольку в новом уравнении в правой и в левой частях стоят логарифмы исходных переменных, эту модель называют логарифмической (или двойной логарифмической).

Смысл проведенного преобразования состоит в том, что относительно параметров (β_1 и β_2) новая модель является линейной. Поэтому к ней можно применять все стандартные методы оценивания, которые обсуждались в предшествующих главах. Точно так же можно вычислять МНК-оценки коэффициентов, рассчитывать их стандартные ошибки, тестировать гипотезы и так далее. Таким образом, стандартный трюк, который используют эконометристы, заключается в переходе от нелинейной по параметрам модели ($y_i = Ax_i^a$) к линейной по параметрам¹.

Единственное существенное отличие будет возникать в интерпретации полученного результата. Действительно, в линейной модели

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

коэффициент при переменной x_i может быть интерпретирован так: увеличение переменной x_i на одну единицу приводит к увеличению переменной y_i на β_2 единиц.

Однако в логарифмической модели интерпретация коэффициента β_2 будет отличаться. Чтобы понять, как она устроена, перепишем логарифмическую модель следующим образом

$$y_i = e^{\beta_1} x_i^{\beta_2} e^{\varepsilon_i}$$

и вычислим производную зависимой переменной по объясняющей переменной:

$$\frac{dy_i}{dx_i} = \beta_2 e^{\beta_1} x_i^{\beta_2-1} e^{\varepsilon_i} = \frac{\beta_2 e^{\beta_1} x_i^{\beta_2} e^{\varepsilon_i}}{x_i} = \beta_2 \frac{y_i}{x_i}.$$

Преобразуем это выражение, выразим темп прироста зависимой переменной:

$$\frac{dy_i}{y_i} = \beta_2 \frac{dx_i}{x_i}$$

Или, что приближенно то же самое:

$$\frac{\Delta y_i}{y_i} * 100\% \approx \beta_2 \frac{\Delta x_i}{x_i} * 100\%$$

Таким образом, мы получаем следующую интерпретацию коэффициента при регрессоре: **увеличение объясняющей переменной на один процент приводит к увеличению зависимой переменной в среднем на β_2 процентов**. Иными словами, коэффициент β_2 характеризует *эластичность* зависимой переменной по объясняющей переменной ².

Линейно-логарифмическая модель

Аналогичным образом можно разобраться с тем, как интерпретировать коэффициент при переменной в линейно-логарифмической модели, то есть в модели следующего вида:

$$y_i = \beta_1 + \beta_2 \ln x_i + \varepsilon_i$$

Возьмем производную зависимой переменной по объясняющей переменной:

$$\frac{dy_i}{dx_i} = \beta_2 \frac{1}{x_i}$$

$$dy_i = \beta_2 \frac{dx_i}{x_i}$$

$$\Delta y_i \approx \beta_2 \frac{\Delta x_i}{x_i}$$

Отсюда легко видеть, что если объясняющая переменная увеличивается на один процент, то есть $\frac{\Delta x_i}{x_i} = \frac{1}{100}$, то зависимая переменная увеличивается на величину $\Delta y_i = \frac{\beta_2 * 1}{100}$. Таким образом, коэффициент при переменной в линейно-логарифмической модели может быть интерпретирован следующим образом: **увеличение объясняющей переменной на один процент приводит к увеличению зависимой переменной в среднем на $\frac{\beta_2}{100}$ единиц**.

Логарифмически-линейная модель

Для полноты картины осталось рассмотреть экспоненциальную зависимость:

$$y_i = e^{\beta_1 + \beta_2 x_i + \varepsilon_i}$$

Такая модель так же легко может быть приспособлена к оцениванию путем перехода к логарифмам (так как в этом случае она снова становится линейной по параметрам):

$$\ln y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

Так как в левой части этого уравнения стоит логарифм исходной зависимой переменной, а справа регрессор входит в уравнение линейно, такие модели называются логарифмически-линейными.

Чтобы понять, как можно интерпретировать результаты моделирования в этом случае, посмотрим, насколько меняется зависимая переменная при изменении объясняющей переменной на Δx_i :

$$\Delta y_i = e^{\beta_1 + \beta_2(x_i + \Delta x_i) + \varepsilon_i} - e^{\beta_1 + \beta_2 x_i + \varepsilon_i}$$

$$\frac{\Delta y_i}{y_i} = \frac{e^{\beta_1 + \beta_2(x_i + \Delta x_i) + \varepsilon_i} - e^{\beta_1 + \beta_2 x_i + \varepsilon_i}}{e^{\beta_1 + \beta_2 x_i + \varepsilon_i}}$$

$$\frac{\Delta y_i}{y_i} = e^{\beta_2 \Delta x_i} - 1$$

$$\frac{\Delta y_i}{y_i} * 100\% = (e^{\beta_2 \Delta x_i} - 1) * 100\%$$

$$\text{Если } \Delta x_i = 1, \text{ то } \frac{\Delta y_i}{y_i} * 100\% = (e^{\beta_2} - 1) * 100\%$$

Поэтому интерпретировать коэффициент в модели в этом случае можно так: **увеличение регрессора на единицу приводит к увеличению зависимой переменной на $(e^{\beta_2} - 1) * 100\%$** .

Если коэффициент β_2 близок к нулю, то $e^{\beta_2} \approx 1 + \beta_2$. В этом случае $(e^{\beta_2} - 1) * 100\% \approx \beta_2 * 100\%$ и можно интерпретировать соответствующий коэффициент так: **увеличение регрессора на единицу приводит к увеличению зависимой переменной на**

$\beta_2 * 100\%$.

На практике приближение является вполне удовлетворительным при $|\beta_2| < 0,1$. В этом случае погрешность меньше одного процента. При больших по абсолютной величине значениях коэффициента β_2 лучше использовать точную формулу.

Все полученные нами выводы об интерпретации коэффициентов в линейной, логарифмической, логарифмически-линейной и линейно-логарифмической моделях обобщены в таблице 4.1.

Таблица 4.1. Интерпретация коэффициентов в разных моделях

Зависимость	Интерпретация
Линейная $y = \beta_1 + \beta_2 x$	Увеличение x на единицу приводит к увеличению y на β_2 единиц
Логарифмическая $\ln y = \beta_1 + \beta_2 \ln x$	Увеличение x на один процент приводит к увеличению y на β_2 процентов
Линейно-логарифмическая $y = \beta_1 + \beta_2 \ln x$	Увеличение x на один процент приводит к увеличению y на $\beta_2/100$ единиц
Логарифмически-линейная $\ln y = \beta_1 + \beta_2 x$	Увеличение x на единицу приводит к увеличению y на $\beta_2 * 100$ процентов

Примечания: (1) Разумеется, если коэффициент β_2 отрицательный, то увеличение регрессора приводит не к увеличению, а, наоборот, к уменьшению зависимой переменной. (2) Следует помнить, что указанные интерпретации получены на основе приближенных формул. В последнем случае (для логарифмически-линейной модели) приближением можно пользоваться только в том случае, если коэффициент β_2 не слишком велик (см. пояснения в тексте параграфа).

Пример 4.3. Интерпретация коэффициента в логарифмически-линейной модели

Исследователь анализирует, как меняется ВВП некоторой страны во времени. Изучив график ВВП, исследователь заключил, что он растет экспоненциально, следовательно, для моделирования его динамики хорошо подойдет логарифмически-линейная модель:

$$\ln GDP_t = \beta_1 + \beta_2 * t + \varepsilon_t$$

Оценка параметров модели на ежегодных данных за 50 лет приводит к следующей линии регрессии:

$$\widehat{\ln GDP_t} = \underset{(0,301)}{10,2} + \underset{(0,001)}{0,02} * t$$

Интерпретируйте полученные результаты.

Решение:

Для начала отметим, что коэффициент при переменной t является статистически значимым, так как гипотеза $H_0 : \beta_2 = 0$ уверенно отвергается при уровне значимости 1%. (соответствующее расчетное значение тестовой статистики равно $0,02/0,001=20$, что больше критического значения, которое составляет 2,68). Поэтому можно заключить, что со временем ВВП в рассматриваемой экономике, действительно, в среднем растет.

После этого можно перейти к интерпретации. В соответствии с полученным нами правилом интерпретации можно сказать, что увеличение переменной t на единицу приводит к увеличению переменной GDP_t на $0,02*100\%$. Иными словами, в среднем ВВП в рассматриваемой экономике увеличивается на 2% в год.

Полиномиальные модели

В некоторых ситуациях зависимость между регрессором и объясняемой переменной носит немонотонный характер. см., например, рисунок 4.4. В таких случаях оправдано использование полиномиальных зависимостей. Например, квадратичных:

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \varepsilon_i.$$

В это уравнение все коэффициенты снова входят линейно, а значит, параметры соответствующей модели также легко могут быть оценены обычным МНК.

$$\frac{dy_i}{dx_i} = \beta_2 + 2\beta_3 x_i$$

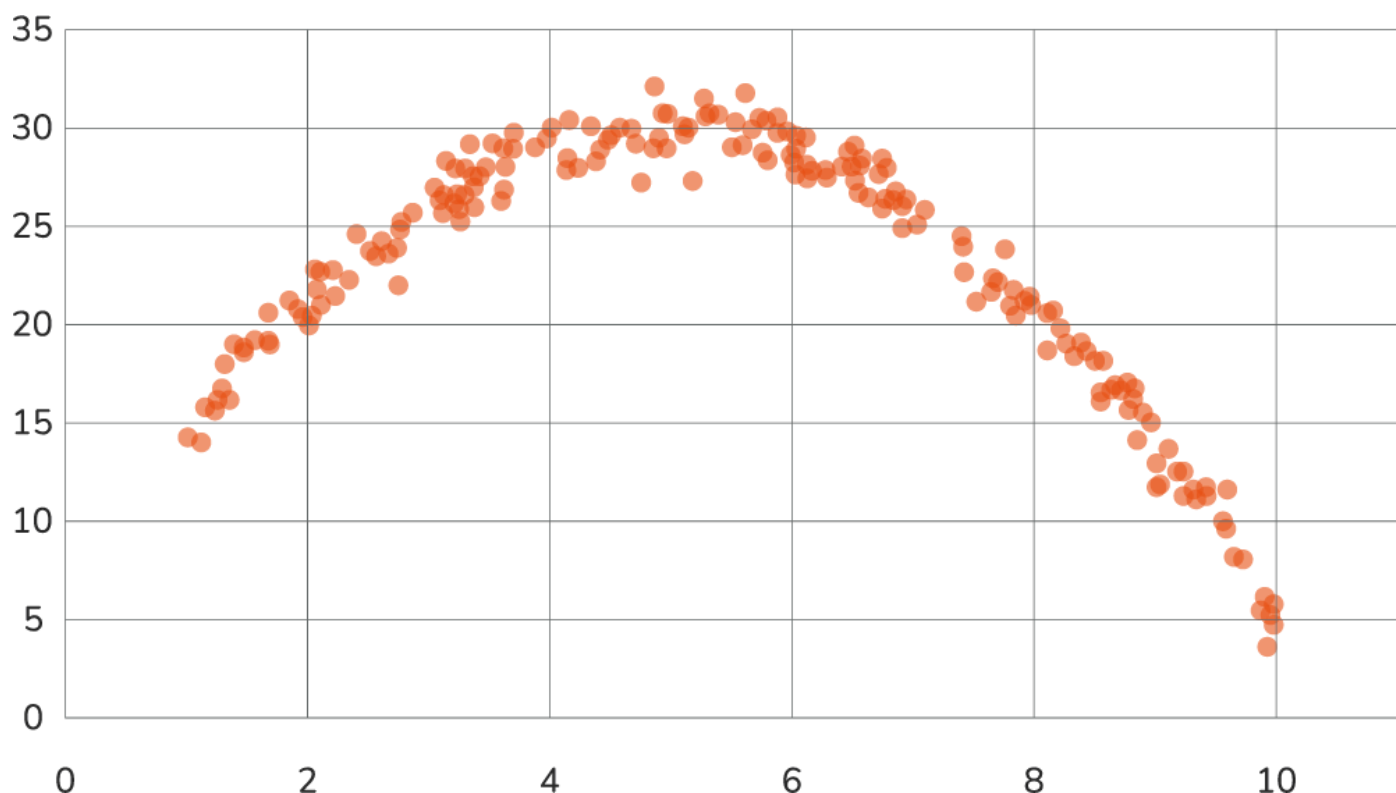


Рис. 4.4. Немонотонный характер зависимости между переменными

Чувствительность зависимой переменной к изменению регрессора будет существенно зависеть от его величины:

$$\frac{dy_i}{dx_i} = \beta_2 + 2\beta_3 x_i$$

Поэтому при интерпретации коэффициента указанную чувствительность обычно вычисляют в конкретной точке. Например, в точке среднего по выборке значения регрессора.

Иногда полиномиальные модели могут быть удобным инструментом. Особенно если немонотонный характер зависимости следует из содержательных соображений. Однако не следует увлекаться оценкой полиномов высоких степеней просто в угоду получению большого коэффициента R-квадрат. Конечно, технически через l точек всегда можно провести кривую, описываемую полиномом степени $n - 1$, и R-квадрат при этом будет равен единице. Однако содержательно интерпретировать подобную зависимость будет невозможно, результаты оценивания будут крайне неустойчивыми, а точность прогнозирования вне выборки — низкой. Поэтому на практике, как правило, ограничиваются квадратичными функциями.

Обратите внимание, что во всех рассмотренных ситуациях путем простых преобразований мы приводили модель к виду, в котором параметры входят в уравнение линейно. Конечно, можно привести пример функции, которая не сводится к линейной по параметрам:

$$y_i = \beta_1 + \beta_2 w_i e^{\beta_3 x_i} + \ln(\beta_4 z_i + \varepsilon_i).$$

В этом случае для оценивания параметров приходится использовать альтернативные методы (например, нелинейный МНК или метод максимального правдоподобия), обсуждение которых выходит за рамки этой главы. К счастью, в прикладных исследованиях часто можно обойтись теми моделями, которые мы разобрали выше.

Естественный вопрос, возникающий после рассмотрения разнообразных нелинейных моделей, состоит в том, как выбрать подходящий вид зависимости для ваших данных: использовать ли линейную модель, или логарифмическую, или ещё какую-то? Тут следует принимать во внимание следующие соображения:

- Графический анализ исходных данных.** Например, ясно, что если диаграммы рассеяния для ваших данных выглядят как рисунки 4.3 или 4.4, то использовать линейную модель будет не слишком хорошей идеей.
- Графический анализ остатков.** После построения модели отсортируйте наблюдения по возрастанию одного из регрессоров и постройте график остатков. Если остатки равномерно колеблются вокруг нуля (то есть, если их поведение не противоречит предпосылке о том, что они являются независимыми случайными величинами), это аргумент в пользу корректной спецификации. Подобный пример приведен на рисунке 4.5.а. Если же остатки имеют некоторый регулярный вид, например, как на рисунке 4.5.б, то, скорее всего, спецификация выбрана неверно³.
- Экономическая теория.** Как было сказано в самом начале данного параграфа, если в основе ваших эмпирических расчетов лежит теоретическая модель, то это может быть хорошим подспорьем в выборе корректной формы функциональной зависимости. Например, если вы моделируете кривую Лаффера⁴, то естественно использовать немонотонную функцию.
- Формальные статистические критерии.** Речь о них пойдет в заключительной части этого параграфа.

Одним из распространенных вариантов тестировать корректность выбранной функциональной формы модели является тест Рамсея (Ramsey test). Иногда его также называют RESET (Regression Equation Specification Error Test). Нулевая гипотеза в этом тесте состоит в том, что спецификация уравнения регрессии верна.

Процедура его проведения такова:

Оцениваем исходное уравнение, корректность спецификации которого как раз и хотим проверить. Например, такое:

$$y_i = \beta_1 + \beta_2 x_i^{(2)} + \dots + \beta_k x_i^{(k)} + \varepsilon_i$$

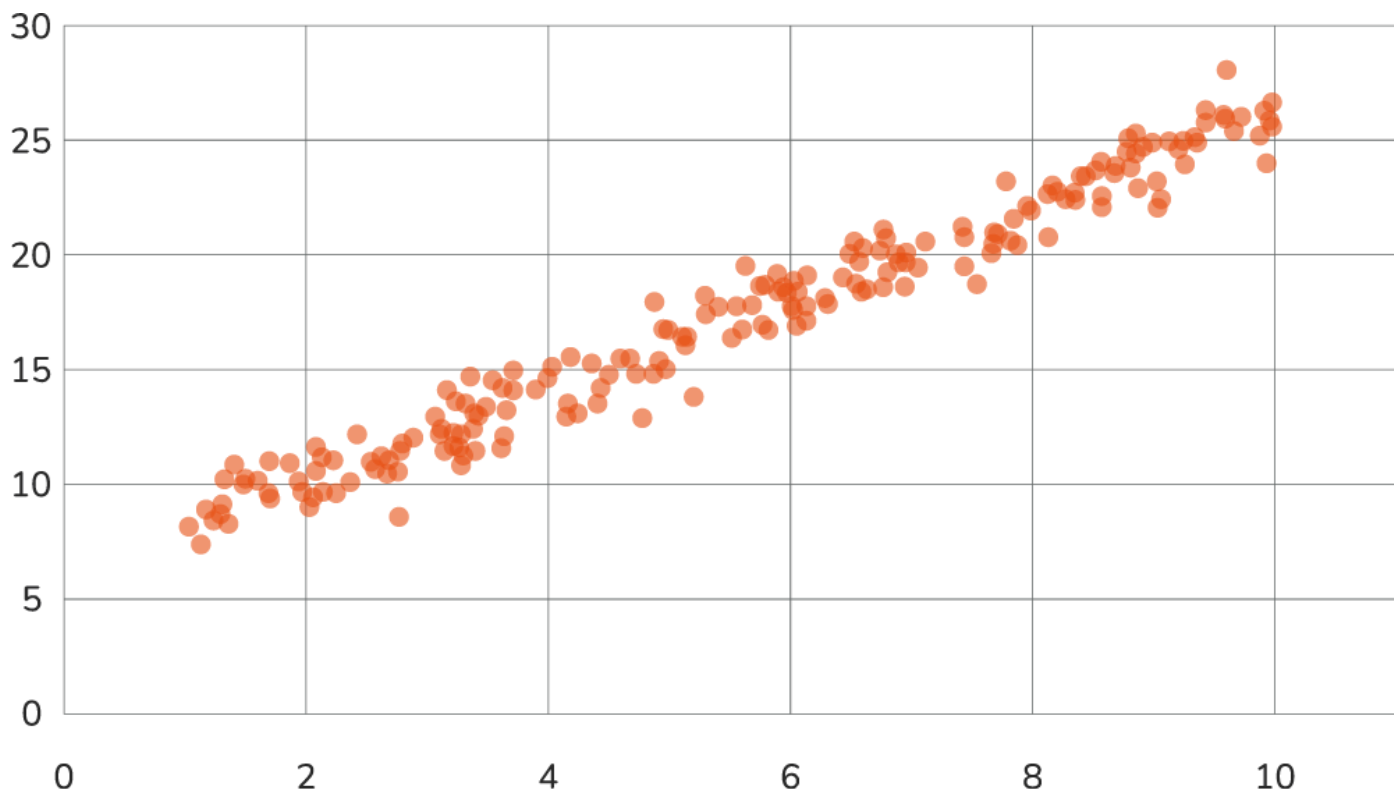
Извлекаем предсказанные значения объясняемой переменной $\hat{y}_i$.

Оцениваем вспомогательное уравнение – в него включены все исходные переменные и дополнительно $\hat{y}_i^2$ и $\hat{y}_i^3$:

$$y_i = \beta_1 + \beta_2 x_i^{(2)} + \dots + \beta_k x_i^{(k)} + \alpha_2 \hat{y}_i^2 + \alpha_3 \hat{y}_i^3 + \varepsilon_i$$

Теперь для имеющихся двух уравнений (исходного и вспомогательного) осуществляем тест на сравнение «короткой» и «длинной» регрессий, чтобы проверить гипотезу $H_0 : \alpha_2 = \alpha_3 = 0$. Если эта гипотеза отвергается, то следует отвергнуть исходную гипотезу теста Рамсея и заключить, что спецификация уравнения неверна. Если же гипотеза $H_0 : \alpha_2 = \alpha_3 = 0$ не отвергается, то следует сделать вывод, что спецификация исходного уравнения верна.

Иногда, если в выборке доступно мало наблюдений, вместо двух слагаемых $\alpha_2 \hat{y}_i^2$ и $\alpha_3 \hat{y}_i^3$ в уравнение добавляют только одно из них.



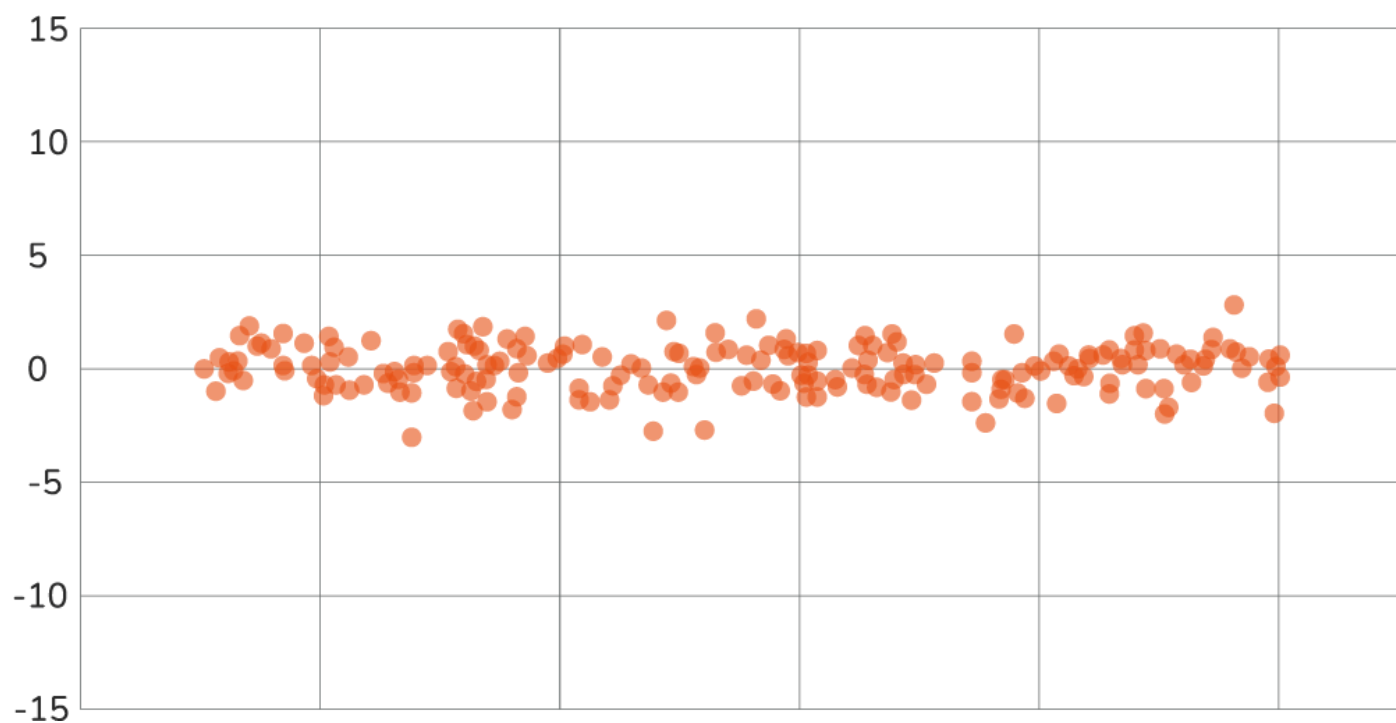
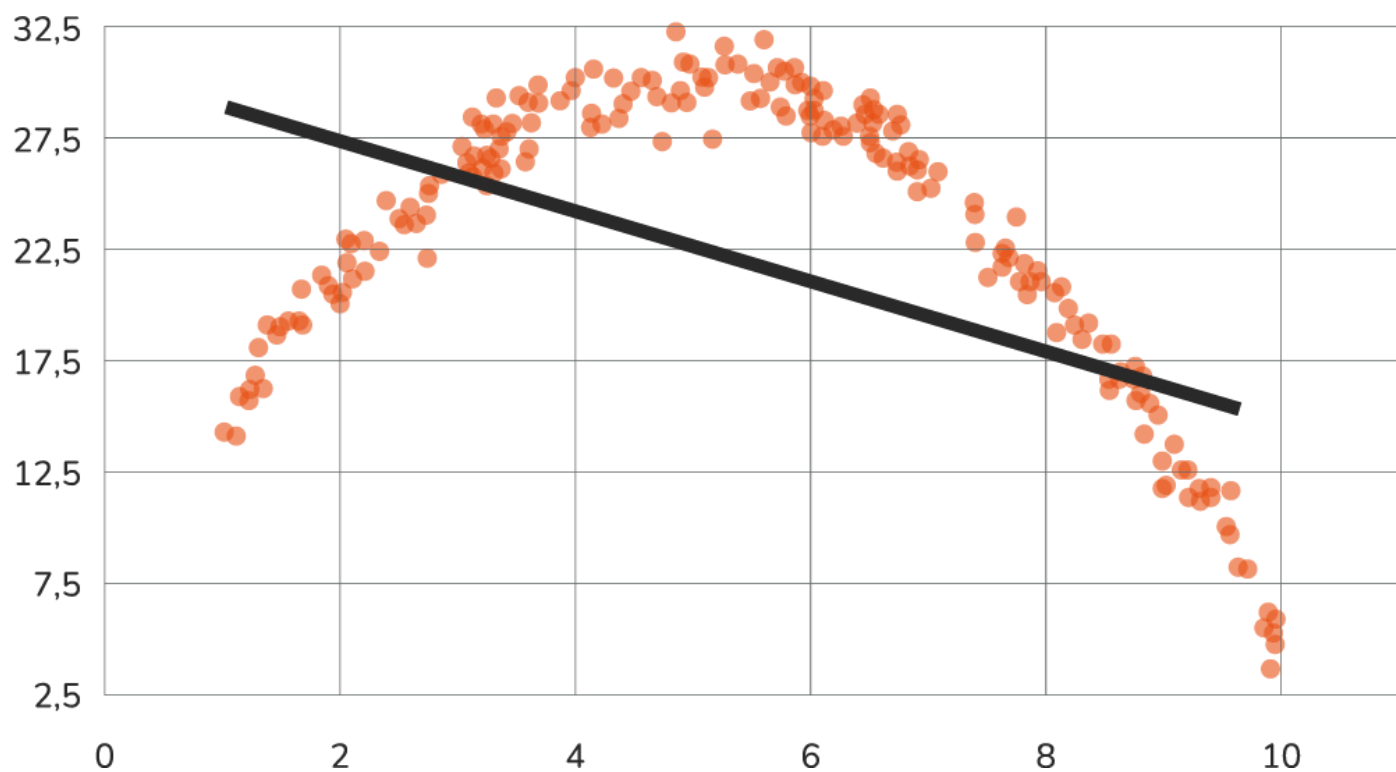


Рис. 4.5а. Линия регрессии (верхний график) и соответствующий ей график остатков (нижний график) в случае, когда линейная функциональная форма связи между переменными является корректным предположением



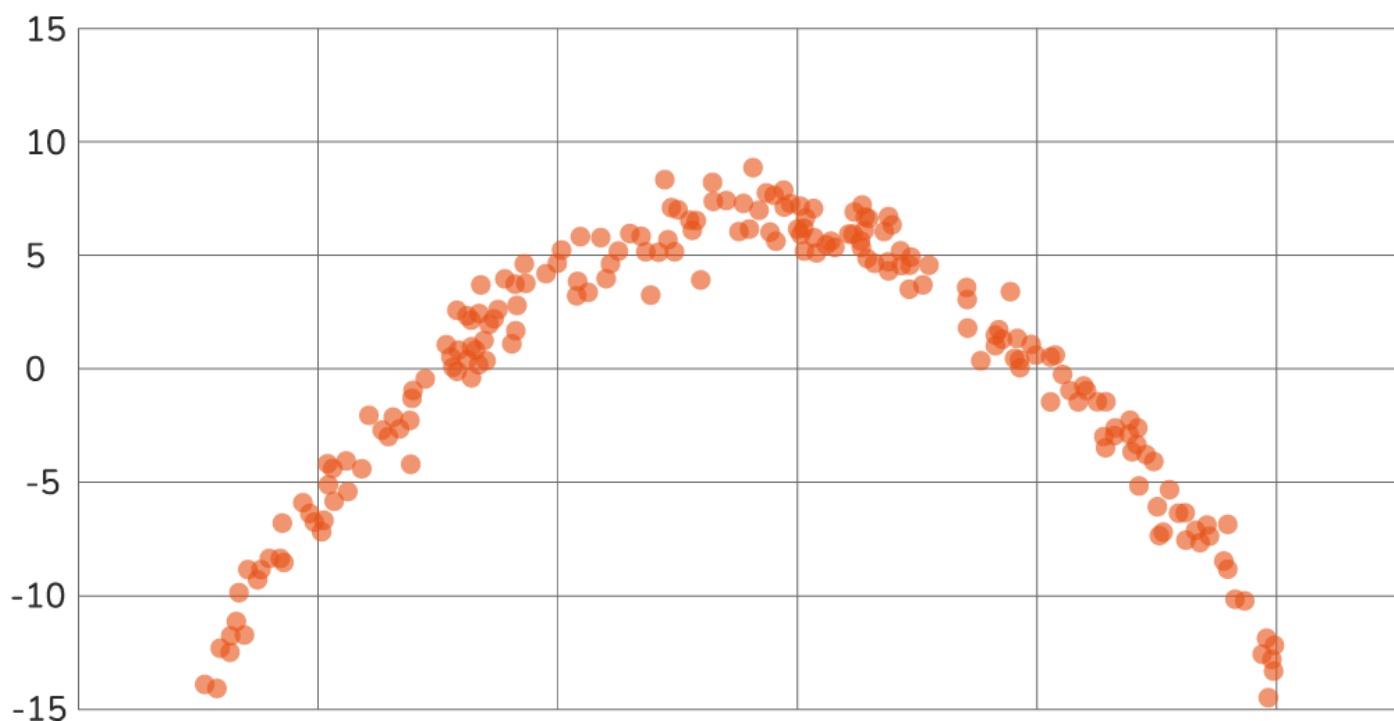


Рис. 4.56. Линия регрессии (верхний график) и соответствующий ей график остатков (нижний график) в случае, когда линейная функциональная форма связи между переменными является некорректным предположением, а в действительности зависимость носит нелинейный характер

Тест Рамсея может отвергать нулевую гипотезу по двум причинам: либо функциональная форма для уравнения выбрана ошибочно, либо в уравнении пропущены важные переменные. К сожалению, этот тест не дает четкого указания, что именно надо сделать для исправления ситуации, так что тут исследователю придется принимать решение самому. Однако в любом случае отвержение нулевой гипотезы тестом Рамсея — это повод задуматься о том, чтобы усовершенствовать спецификацию вашей модели⁵.

Иногда выбор функциональной формы осуществляют и на основе качества соответствия модели данным. Например, линейную модель $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ и линейно-логарифмическую модель $y_i = \beta_1 + \beta_2 \ln x_i + \varepsilon_i$ можно сравнить при помощи коэффициента R-квадрат или суммы квадратов остатков. Такое сравнение будет корректным, так как у рассматриваемых типов моделей одинаковая зависимая переменная (y_i). В то же время сравнивать линейную модель и логарифмическую модель подобным образом не следует, так как у логарифмической модели другая зависимая переменная ($\ln y_i$). Поэтому сумма квадратов и общая сумма квадратов в таких моделях будут существенно отличаться просто из-за того, что измеряются в разных шкалах. В любом случае, как мы обсуждали выше, R-квадрат имеет существенные ограничения, поэтому в современных исследованиях при выборе функциональной формы гораздо чаще опираются на соображения экономической теории и спецификационные тесты.

1. Для того, чтобы случайные ошибки в нашем преобразовании не брались «ниоткуда», можно предположить, что исходная модель тоже их включает, то есть имеет вот такой вид: $y_i = Ax_i^a e^{\varepsilon_i}$. [↪](#)

2. Для читателя, знакомого с микроэкономикой, этот результат, скорее всего, является вполне ожидаемым: действительно, показатель степени в степенной зависимости численно равен соответствующей эластичности. [↪](#)

3. Дополнительные полезные соображения о том, какую информацию можно извлечь из анализа остатков уравнения регрессии, содержатся в следующей главе, посвященной гетероскедастичности. [↪](#)

4. Зависимость между ставкой налога и суммарными налоговыми поступлениями в государственный бюджет. [↪](#)

5. В более ранних исследованиях для выбора функциональной формы также использовались тесты Зарембки и Бокса — Кокса. См., например, Доугерти, 2009. [↪](#)

4.4. Обобщающий пример

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

[Скрипт в R для этого примера](#)

[Файл с данными для этого примера.](#)

В заключение этой главы рассмотрим пример, позволяющий обобщить всё, что мы в ней выяснили.

Пример 4.4. Цена коттеджа

В файле Cottage.xlsx имеются следующие данные о двух сотнях коттеджей:

living_area — жилая площадь коттеджа, м²;

total_area — общая площадь коттеджа, м²;

land — площадь участка, на котором расположен коттедж, сотки;

dist — расстояние от города до участка с коттеджем, км;

lake — фиктивная переменная, равная единице для коттеджей, расположенных на берегу естественного водоема: реки или озера.

price — цена коттеджа, млн руб.

Мы не будем переходить сразу к построению моделей с этой переменной, а начнем с предварительного анализа данных. В таблицах 4.2 и 4.3 представлены описательные статистики для анализируемых переменных.

Из таблицы 4.2 видно, например, что общая площадь коттеджей в нашей выборке изменяется от 54 до 175 квадратных метров, а средняя цена коттеджа с участком, на котором он расположен, равна 16,6 миллионов рублей. Среднее значение переменной *lake*, равное 0,055, говорит о том, что 5,5% коттеджей в выборке расположены непосредственно рядом с водоемом.

Таблица 4.3 позволяет заключить, что знаки коэффициентов корреляции между ценой коттеджа и прочими факторами соответствуют нашим ожиданиям: цена коттеджа положительно коррелирована с его площадью (как жилой, так и общей), с площадью участка и с близостью к водоему. Между ценой коттеджа и расстоянием от него до города, напротив, наблюдается отрицательная корреляция.

Таблица 4.2. Описательные статистики для всех переменных

Переменная	Среднее	Медиана	S.D.	Min	Max
<i>living_area</i>	89,8	92,0	28,7	41,0	140,0
<i>total_area</i>	113,0	115,0	29,7	54,0	175,0
<i>land</i>	24,6	24,0	8,96	10,0	40,0
<i>dist</i>	60,9	58,5	27,9	10,0	109,0
<i>lake</i>	0,055	0,000	0,229	0,000	1,000
<i>price</i>	16,6	13,0	14,2	2,00	99,0

Таблица 4.3. Матрица парных коэффициентов корреляции

<i>living_area</i>	<i>total_area</i>	<i>land</i>	<i>dist</i>	<i>lake</i>	<i>price</i>	
1,000	0,963	-0,069	-0,192	-0,006	0,379	<i>living_area</i>
	1,000	-0,083	-0,187	-0,018	0,380	<i>total_area</i>
		1,000	-0,061	0,003	0,076	<i>land</i>
			1,000	0,090	-0,293	<i>dist</i>
				1,000	0,471	<i>lake</i>
					1,000	<i>price</i>

Ещё одно важное наблюдение, которое можно сделать из таблицы 4.3: жилая площадь и общая площадь коттеджа сильно коррелированы, что может говорить о мультиколлинеарности, которая, скорее всего, возникнет в модели с обеими этими переменными. Убедимся в этом, построив модель для цены, включающую две эти переменные, площадь участка и расстояние до города.

Результаты обработки эконометрическим пакетом представлены в таблице 4.4. Мы видим, что ряд переменных, которые в соответствии с соображениями здравого смысла должны влиять на цену коттеджа, оказались незначимыми. Проверим наше

предположение о мультиколлинеарности, сформулированное абзацем выше, вычислив коэффициенты VIF для представленной модели. Значения коэффициентов VIF для переменных living_area, total_area, land и dist составляют, соответственно, 13,616, 13,820, 1,015, 1,045. Первые два из них больше 10, что является дополнительным аргументом в пользу сильной мультиколлинеарности. Попробуем решить эту проблему, оставив в нашей модели только одну из двух площадей, например, общую. Новая модель представлена в таблице 4.5. Для неё все коэффициенты VIF меньше 10, так что существенной мультиколлинеарности в ней нет. Отметим, что это хорошо сказалось на точности оценивания коэффициента при переменной total_area: его стандартная ошибка уменьшилась, и он оказался значимым на однопроцентном уровне (о чем свидетельствует Р-значение, которое существенно меньше 0,01). Обратите внимание, что по критерию исправленного R-квадрата новая модель также стала немного лучше: этот коэффициент увеличился с 0,188 до 0,191.

Таблица 4.4. Результаты оценивания модели №1 (линейной)

Модель 1: МНК, использованы наблюдения 1-200

Зависимая переменная: price

	Коэффициент	Ст. ошибка	t-статистика	Р-значение	
const	2,25967	5,65823	0,3994	0,6901	
living_area	0,0652513	0,117412	0,5557	0,5790	
total_area	0,104345	0,113172	0,9220	0,3577	
land	0,141501	0,101732	1,391	0,1658	
dist	-0,112346	0,0331773	-3,386	0,0009	***
Сумма кв. остатков	31797,57		Ст. ошибка модели	12,76967	
R-квадрат	0,204604		Испр. R-квадрат	0,188288	
F(4, 195)	12,54020		Р-значение (F)	4,25e-09	

Таблица 4.5. Результаты оценивания модели №2 (линейной)

Модель 2: МНК, использованы наблюдения 1-200

Зависимая переменная: price

	Коэффициент	Ст. ошибка	t-статистика	Р-значение	
const	1,26480	5,35810	0,2361	0,8136	
total_area	0,164814	0,0310774	5,303	<0,0001	***
land	0,143690	0,101477	1,416	0,1584	
dist	-0,113168	0,0330858	-3,420	0,0008	***
Сумма кв. остатков	31847,94		Ст. ошибка модели	12,74714	
R-квадрат	0,203344		Испр. R-квадрат	0,191150	
F(3, 196)	16,67612		Р-значение (F)	1,09e-09	

Запишем полученную модель в виде уравнения:

$$\widehat{price}_i = 1,26 + 0,16total_{area_i} + 0,14land_i - 0,11dist_i$$

(5,36) (0,03) (0,10) (0,03)

Как можно интерпретировать полученные результаты? Увеличение площади коттеджа на один квадратный метр увеличивает его цену в среднем при прочих равных условиях на 0,16 млн рублей. Каждый дополнительный километр расстояния до города в среднем при прочих равных условиях снижает цену коттеджа на 0,11 млн рублей.

Коэффициент при переменной land в данном случае является статистически незначимым, так что его интерпретировать смысла нет (так как нет уверенности в том, что он отличен от нуля). Как объяснить подобное наблюдение? Возможно, конечно, дело в том, что площадь участка не слишком важна для цены коттеджа, а возможно, спецификация модели пока не совершенна. Попробуем проанализировать нелинейную спецификацию. В таблице 4.6 представлены результаты оценки логарифмической модели с тем же самым набором переменных.

Таблица 4.6. Результаты оценивания модели №3 (логарифмической)

Модель 3: МНК, использованы наблюдения 1-200

Зависимая переменная: l_price

	Коэффициент	Ст. ошибка	t-статистика	P-значение	
const	-1,66295	0,945301	-1,759	0,0801	*
l_total_area	1,04244	0,157187	6,632	<0,0001	***
l_land	0,265954	0,109011	2,440	0,0156	**
l_dist	-0,383089	0,0750266	-5,106	<0,0001	***
Сумма кв. остатков	72,16836		Ст. ошибка модели	0,606800	
R-квадрат	0,315417		Испр. R-квадрат	0,304939	
F(3, 196)	30,10192		P-значение (F)	4,74e-16	

Отметим, что в новой модели коэффициент при переменной land является статистически значимым при уровне значимости 5%, так как соответствующее P-значение меньше пяти сотых. Коэффициенты при двух остальных регрессорах значимы на однопроцентном уровне.

Запишем полученную модель в виде уравнения:

$$\widehat{\ln price_i} = -1,66 + 1,04 \ln total_{area_i} + 0,27 \ln land_i - 0,38 \ln dist_i$$

(0,95) (0,16) (0,11) (0,08)

Поскольку теперь модель является логарифмической, то интерпретировать результаты можно так: увеличение общей площади коттеджа на 1% увеличивает её цену в среднем при прочих равных тоже примерно на 1%. Однопроцентное увеличение площади участка, в свою очередь, соответствует увеличению цены примерно на 0,3%. Наконец, увеличение расстояния от коттеджа до города на 1% снижает цену коттеджа на 0,4%.

Обратимся теперь к влиянию на цену коттеджа его близости к водоему. Для начала воспользуемся тестом Чоу, чтобы проверить гипотезу о том, что близость к водоему не приводит к структурному сдвигу в модели для цены. В нашем случае для этого требуется добавить в рассматриваемую модель одну фиктивную переменную сдвига (переменную lake) и три фиктивных переменных наклона (три произведения lake*total_area, lake*land, lake*dist), а затем проверить гипотезу о том, что коэффициенты при всех этих переменных одновременно равны нулю (для этого будем использовать обычный тест на сравнение «короткой» и «длинной» регрессий). Это приводит к следующим результатам:

Тест Чоу для структурных изменений в точке lake

F(4, 192) = 11,487; p-значение 0,000

Поскольку p-значение теста Чоу меньше одной сотой, при уровне значимости 1% можно заключить, что структурный сдвиг в данных присутствует, и близость к водоему влияет на цену коттеджа. Рассмотрим несколько спецификаций, учитывающих это влияние. Начнем с модели, включающей переменную сдвига (см. таблицу 4.7).

Таблица 4.7. Результаты оценивания модели №4

Модель 4: МНК, использованы наблюдения 1-200

Зависимая переменная: l_price

	Коэффициент	Ст. ошибка	t-статистика	P-значение	
const	-1,54847	0,863024	-1,794	0,0743	*
l_total_area	1,03798	0,143476	7,235	<0,0001	***
l_land	0,260242	0,0995056	2,615	0,0096	***
l_dist	-0,417305	0,0686934	-6,075	<0,0001	***
lake	1,09338	0,172327	6,345	<0,0001	***
Сумма кв. остатков	59,81914		Ст. ошибка модели	0,553864	
R-квадрат	0,432561		Испр. R-квадрат	0,420921	
F(4, 195)	37,16229		P-значение (F)	4,39e-23	

Отметим, что в новой модели все переменные статистически значимы при уровне значимости 1%. В том числе добавленная нами фиктивная переменная lake. Так как эта переменная входит в уравнение линейно, а зависимая переменная — под логарифмом, то для интерпретации результатов нам потребуется воспользоваться формулой для логарифмически-линейной модели, которую мы получили в предыдущем параграфе. В соответствии с ней можно заключить, что в среднем при прочих равных условиях коттеджи, расположенные рядом с водоемом, дороже остальных коттеджей на

$$(e^{1,09} - 1) * 100\% = 197\%.$$

Получается, что в нашей выборке коттеджи рядом с водоемом почти в три раза дороже коттеджей, расположенных не рядом с ним (при условии равенства всех прочих характеристик). Альтернативная (и, возможно, более реалистичная гипотеза) состоит в том, что прибавка к цене коттеджа за близость к водоему не является фиксированной, а зависит от его площади. Это приводит нас к спецификации с фиктивной переменной наклона, представленной в таблице 4.8.

Таблица 4.8. Результаты оценивания модели №5

Модель 5: МНК, использованы наблюдения 1-200

Зависимая переменная: $\ln price$

	Коэффициент	Ст. ошибка	t-статистика	P-значение	
const	-1,49135	0,863459	-1,727	0,0857	*
$\ln total_area$	1,02785	0,143526	7,161	<0,0001	***
$\ln land$	0,257921	0,0995325	2,591	0,0103	**
$\ln dist$	-0,417855	0,0687166	-6,081	<0,0001	***
$lake * \ln total_area$	0,232968	0,0367676	6,336	<0,0001	***
Сумма кв. остатков	59,84676		Ст. ошибка модели	0,553991	
R-квадрат	0,432299		Испр. R-квадрат	0,420654	
F(4, 195)	37,12264		P-значение (F)	4,59e-23	

Записав эту модель в виде уравнения, получим следующий результат:

$$\widehat{\ln price}_i = -1,49 + 1,03 \ln total_area_i + 0,26 \ln land_i - 0,42 \ln dist_i + 0,23 lake_i * \ln total_area_i$$

(0,86) (0,14) (0,10) (0,07) (0,04)

Чтобы понять, как близость к водоему влияет на цену коттеджа в данном случае, воспользуемся уже знакомым нам приемом: выпишем уравнение отдельно для коттеджей, расположенных не рядом с водоемом (то есть для тех наблюдений, где переменная lake равна нулю), и для коттеджей, расположенных рядом с водоемом (переменная lake равна единице).

Случай $lake_i = 0$:

$$\widehat{\ln price}_i = -1,49 + 1,03 \ln total_area_i + 0,26 \ln land_i - 0,42 \ln dist_i$$

Случай $lake_i = 1$:

$$\widehat{\ln price}_i = -1,49 + 1,26 \ln total_area_i + 0,26 \ln land_i - 0,42 \ln dist_i$$

Таким образом, для коттеджей, расположенных не рядом с водоемом, один дополнительный процент площади увеличивает цену на 1,03%, а для коттеджей, расположенных рядом с водоемом, — аж на 1,26%. Иными словами, для коттеджей рядом с водой эластичность цены по площади больше на 0,23.

Проверим корректность спецификации последней рассмотренной модели при помощи теста Рамсея. Соответствующие результаты представлены в таблице 4.9. Обратите внимание, что в соответствии с процедурой теста в уравнение для модели 5 добавлены квадраты и кубы зависимой переменной. Гипотеза о том, что коэффициенты при этих переменных равны нулю, не отвергается (это видно из того, что указанное внизу таблицы р-значение=0,568>0,05). Поэтому гипотеза о том, что предложенная спецификация корректна, не отвергается.

Таблица 4.9. Результаты теста Рамсея для модели №5

Вспомогательная регрессия для теста Рамсея

МНК, использованы наблюдения 1-200

Зависимая переменная: $\ln price$

	Коэффициент	Ст. ошибка	t-статистика	P-значение
const	-4,53589	7,39461	-0,6134	0,5403
$\ln total_area$	2,43723	3,24327	0,7515	0,4533
$\ln land$	0,606548	0,815676	0,7436	0,4580
$\ln dist$	-0,975025	1,33506	-0,7303	0,4661
$lake * \ln total_area$	0,522222	0,737288	0,7083	0,4796
$\hat{y}^2$	-0,624421	1,17952	-0,5294	0,5971
$\hat{y}^3$	0,0906579	0,143396	0,6322	0,5280

Тестовая статистика: $F = 0,567765$,
 р-значение = $P(F(2,193) > 0,567765) = 0,568$

4.5. Рекомендации по оформлению результатов эконометрических расчетов

Еще раз подчеркнем, что последние несколько таблиц, которые мы привели в примере выше — это практически необработанные технические таблицы, выдаваемые эконометрическим пакетом при оценивании соответствующих моделей. Мы сознательно приводим их на страницах данной книги в учебных целях, однако в реальном академическом исследовании представлять результаты подобным образом — дурной тон, потому что это неудобно для читателя.

Гораздо более удачным вариантом представления результатов являются сводные таблицы, позволяющие сопоставить несколько анализируемых моделей и содержащие только необходимую информацию. В качестве примера хорошего оформления результатов эконометрического моделирования обратимся к таблице 4.10, где представлены сводные результаты оценивания трех моделей из последнего примера. Помимо прочего, обратите внимание на следующие аспекты:

- Количество знаков после запятой. Оно не должно быть слишком большим (двух-трех знаков после запятой чаще всего достаточно), так как в действительности вы редко можете оценить коэффициенты в модели с точностью даже до одного знака после запятой, так что подобное количество цифр будет только вводить читателя в заблуждение.
- Понятные названия переменных. Вместо технических сокращений, которые вы используете в эконометрическом пакете, лучше добавить в таблицу понятные названия переменных. Если речь идет о русскоязычной статье или, например, дипломе на русском языке, то и названия лучше писать по-русски. Если же вы готовите публикацию на английском, то и названия должны быть записаны на нём.
- Лучше сопровождать таблицу примечаниями, в которых пояснять все необходимые детали. Скажем, если в таблице приведены результаты каких-то специфических тестов или разные модели в одной таблице оценены на разных подвыборках (или относятся к разным периодам времени), то всё это стоит прокомментировать под ней.

Как проверить себя? Если можно открыть вашу таблицу, и, не читая остальную часть работы, разобраться во всем, что в этой таблице происходит, значит, вы всё сделали правильно.

Таблица 4.10. Влияние различных факторов на цену коттеджа

	Модель 3	Модель 4	Модель 5
Константа	-1,66*	-1,55*	-1,49*
	(0,95)	(0,86)	(0,86)
Логарифм общей площади дома	1,04***	1,04***	1,03***
	(0,16)	(0,14)	(0,14)
Логарифм площади участка	0,27**	0,26***	0,26**
	(0,11)	(0,10)	(0,10)
Логарифм расстояния до города	-0,38***	-0,42***	-0,42***
	(0,08)	(0,07)	(0,07)
Близость к водоему	—	1,09***	—
		(0,17)	
Произведение близости к водоёму и общей площади дома	—	—	0,23***
			(0,04)
Число наблюдений	200	200	200
Исправленный R^2	0,30	0,42	0,42

Примечания: Все модели оценены при помощи обычного МНК. Зависимая переменная — логарифм цены коттеджа. В скобках указаны стандартные ошибки. * обозначает значимость на 10-процентном уровне; ** обозначает значимость на 5-процентном уровне; *** обозначает значимость на 1-процентном уровне

Глава 5. Гетероскедастичность, обобщенная линейная модель

В этой главе мы выясним, как меняются свойства эконометрических оценок, если отказаться от двух предпосылок классической линейной модели множественной регрессии: о постоянстве дисперсии случайных ошибок, а также (ближе к концу главы) об отсутствии корреляции между случайными ошибками, относящимися к разным наблюдениям.

Это важно сделать, так как в реальных прикладных исследованиях хотя бы одна из этих проблем возникает почти всегда. К счастью, существуют простые и действенные средства, позволяющие избежать основных негативных последствий нарушения указанных предпосылок. Их мы тоже обсудим.

[5.1. Гетероскедастичность: определение и последствия](#)

[5.2. Состоятельные в условиях гетероскедастичности стандартные ошибки](#)

[5.3. Взвешенный метод наименьших квадратов](#)

[5.4. Выявление гетероскедастичности](#)

[5.5. Обобщенная линейная модель и обобщенный МНК](#)

[Задания для самостоятельного решения](#)

5.1. Гетероскедастичность: определение и последствия

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

Вспомните предпосылки классической линейной модели парной регрессии (сформулированы в параграфе 2.3) и множественной регрессии (в параграфе 3.2). В обоих случаях предпосылка №4 состояла в том, что дисперсия случайной ошибки постоянна: $\text{var}(\varepsilon_i) = \sigma^2 = \text{const}$. (гомоскедастичность случайных ошибок).

Если же дисперсия случайной ошибки не является постоянной: $\text{var}(\varepsilon_i) = \sigma_i^2 \neq \text{const}$, то есть четвертая предпосылка классической модели не выполняется, такая ситуация называется **гетероскедастичностью**.

Напомним, что интуитивно понять различие между двумя этими ситуациями можно, взглянув на рисунки 2.3а и 2.3б, приведенные в параграфе 2.3. Первый из них соответствует ситуации, когда дисперсия случайной ошибки постоянна, и поэтому разброс точек вокруг линии регрессии более-менее равномерен. Во втором случае наблюдается частный случай гетероскедастичности, а именно ситуация, когда дисперсия случайной ошибки $\text{var}(\varepsilon_i)$ положительно зависит от x_i , и поэтому разброс точек вокруг линии регрессии увеличивается по мере увеличения значения этой переменной.

Когда можно ожидать, что в реальном исследовании в данных будет наблюдаться гетероскедастичность? Представим, например, что мы анализируем зависимость потребления индивида от его располагаемого дохода. Тогда располагаемый доход индивида является объясняющей переменной x . Понятно, что для групп индивидов с маленьким доходом, который измеряется десятками долларов в месяц, потребление будет разным, но оно, скорее всего, тоже будет измеряться десятками долларов в месяц. Соответственно, и разброс потребления (отклонение от линии регрессии) для этих индивидов также будет измеряться в десятках долларов. С другой стороны, если взять очень богатых индивидов, у которых доход измеряется десятками тысяч долларов, то и разброс потребления у них тоже будет составлять несколько тысяч долларов. Получается, что для бедных индивидов разброс потребления будет маленьким, а для богатых индивидов — большим. Это и есть ситуация гетероскедастичности.

Подчеркнем, что гетероскедастичность не обязательно имеет вид, приведенный на рисунке 2.3б, то есть дисперсия случайной ошибки не обязательно должна расти пропорционально какому-то регрессору. Зависимость дисперсии случайной ошибки от тех или иных переменных может иметь и более сложный характер.

Пусть выполнены все предпосылки классической линейной модели множественной регрессии за одним исключением: в данных наблюдается гетероскедастичность. Как это скажется на свойствах МНК-оценок коэффициентов? Перечислим основные последствия:

1. **МНК-оценки коэффициентов останутся несмещенными.** В этом легко убедиться, если вернуться к параграфу 2.4 и обратить внимание, что предпосылка №4 об отсутствии гетероскедастичности никак не используется при доказательстве несмещенности.
2. **МНК-оценки коэффициентов больше не являются эффективными.** Из того же параграфа 2.4 видно, что соответствующая предпосылка критична для доказательства эффективности.
3. **Стандартные ошибки оценок коэффициентов, рассчитанные по формуле для случая гомоскедастичности, оказываются смещенными и несостоятельными.** Следовательно, их использование для тестирования гипотез и построения доверительных интервалов может привести к некорректным выводам.

Первые два перечисленных последствия говорят о том, что МНК-оценки коэффициентов в условиях гетероскедастичности хотя и теряют в точности, однако остаются в среднем правильными. Третье же последствие весьма критично, так как увеличивает вероятность неверной интерпретации результатов моделирования. Поэтому в следующем параграфе мы сконцентрируемся на методе решения этой проблемы.

5.2. Состоятельные в условиях гетероскедастичности стандартные ошибки

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

Так как гетероскедастичность не приводит к смещению оценок коэффициентов, можно по-прежнему использовать МНК. Смещены и несостоятельны оказываются не сами оценки коэффициентов, а их стандартные ошибки, поэтому формула для расчета стандартных ошибок в условиях гомоскедастичности не подходит для случая гетероскедастичности.

Естественной идеей в этой ситуации является корректировка формулы расчета стандартных ошибок, чтобы она давала «правильный» (состоятельный) результат. Тогда можно снова будет корректно проводить тесты, проверяющие, например, незначимость коэффициентов, и строить доверительные интервалы. Соответствующие «правильные» стандартные ошибки называются **состоятельными в условиях гетероскедастичности стандартными ошибками** (heteroskedasticity consistent (heteroskedasticity robust) standard errors)¹. Первоначальная формула для их расчета была предложена Уайтом, поэтому иногда их также называют стандартными ошибками в форме Уайта (White standard errors). Предложенная Уайтом состоятельная оценка ковариационной матрицы вектора оценок коэффициентов имеет вид:

$$\hat{V}(\hat{\beta}) = n(X'X)^{-1} \left(\frac{1}{n} \sum_{s=1}^n e_s^2 x_s x_s' \right) (X'X)^{-1},$$

где x_s – это s-я строка матрицы регрессоров X. Легко видеть, что эта формула более громоздка, чем формула

$$\hat{V}(\hat{\beta}) = (X'X)^{-1} S^2,$$

которую мы вывели в третьей главе для случая гомоскедастичности. К счастью, на практике

соответствующие вычисления не представляют сложности, так как возможность автоматически рассчитывать стандартные ошибки в форме Уайта реализована во всех современных эконометрических пакетах. Общепринятое обозначение для этой версии стандартных ошибок: «HC0». В работах (MacKinnon, White, 1985) и (Davidson, MacKinnon, 2004) были предложены и альтернативные версии, которые обычно обозначаются в эконометрических пакетах «HC1», «HC2» и «HC3». Их расчетные формулы несколько отличаются, однако суть остается прежней: они позволяют состоятельно оценивать стандартные отклонения МНК-оценок коэффициентов в условиях гетероскедастичности.

Для случая парной регрессии состоятельная в условиях гетероскедастичности стандартная ошибка оценки коэффициента при регрессоре имеет вид:

$$se(\hat{\beta}_2) = \sqrt{\frac{1}{n} \frac{\sum_{i=1}^n (x_i - \bar{x})^2 e_i^2}{\widehat{var}(x)^2}}.$$

Формальное доказательство состоятельности будет приведено в следующей главе. Пока же обсудим пример, иллюстрирующий важность использования робастных стандартных ошибок.

Пример 5.1. Оценка эффективности использования удобрений

В файле *Agriculture* в материалах к этому учебнику содержатся следующие данные 2010 года об урожайности яровой и озимой пшеницы в Спасском районе Пензенской области:

PRODP – урожайность в денежном выражении, в **тысячах** рублей с 1 га,

SIZE – размер пахотного поля, га,

LABOUR – трудозатраты, руб. на 1 га,

FUNG1 – фунгициды, протравители семян, расходы на удобрение в руб. на 1 га,

FUNG2 – фунгициды, во время роста, расходы на удобрение в руб. на 1 га,

GIRB – гербициды, расходы на удобрение в руб. на 1 га,

INSEC – инсектициды, расходы на удобрение в руб. на 1 га,

YDOB1 – аммофос, во время сева, расходы на удобрение в руб. на 1 га,

YDOB2 – аммиачная селитра, во время роста, расходы на удобрение в руб. на 1 га.

Представим, что вас интересует ответ на вопрос: влияет ли использование фунгицидов на урожайность поля?

(а) Оцените зависимость урожайности в денежном выражении от константы и переменных FUNG1, FUNG2, YDOB1, YDOB2, GIRB, INSEC, LABOUR. Запишите уравнение регрессии в стандартной форме, указав коэффициент детерминации и (в скобках под соответствующими коэффициентами) стандартные ошибки **для случая гомоскедастичности**. Какие из переменных значимы на 5-процентном уровне значимости?

(б) Решите предыдущий пункт заново, используя теперь **состоятельные в условиях гетероскедастичности** стандартные ошибки. Сопоставьте выводы по поводу значимости (при пятипроцентном уровне) переменных, характеризующих использование фунгицидов.

Решение:

(а) Оценим требуемое уравнение:

Модель 1: МНК, использованы наблюдения 1-200

Зависимая переменная: PRODP

	Коэффициент	Ст. ошибка	t-статистика	P-значение	
--	-------------	------------	--------------	------------	--

const	-38,4019	7,5273	-5,1017	<0,00001	***
FUNG1	0,0445755	0,0487615	0,9142	0,36178	
FUNG2	0,103625	0,049254	2,1039	0,03669	**
GIRB	0,0776059	0,0523553	1,4823	0,13990	
INSEC	0,0782521	0,0484667	1,6146	0,10805	
LABOUR	0,0415064	0,00275277	15,0781	<0,00001	***
YDOB1	0,0492168	0,0233328	2,1093	0,03621	**
YDOB2	-0,0906824	0,025864	-3,5061	0,00057	***
Сумма кв. остатков	150575,6		Ст. ошибка модели	28,00443	
R-квадрат	0,801958		Испр. R-квадрат	0,794738	
F(7, 192)	111,0701		P-значение (F)	5,08e-64	

Переменные FUNG2, LABOUR, YDOB1 и YDOB2 значимы на пятипроцентном уровне значимости (причем LABOUR и YDOB2 — ещё и на однопроцентном).

Если представить те же самые результаты в форме уравнения, то получится вот так:

$$\widehat{PRODP}_i = -38,40 + 0,04 * FUNG1_i + 0,10 * FUNG2_i + 0,08 * GIRB_i + 0,08 * INSEC_i + 0,04 * LABOUR_i + 0,05 * YDOB1_i - 0,09 * YDOB2_i, R^2 = 0,802$$

(7,53) (0,05) (0,05) (0,05) (0,05) (0,003) (0,02) (0,03)

(6) При использовании альтернативных стандартных ошибок получим следующий результат:

Модель 2: МНК, использованы наблюдения 1-200

Зависимая переменная: PRODP

Робастные оценки стандартных ошибок (с поправкой на гетероскедастичность), вариант HC1

	Коэффициент	Ст. ошибка	t-статистика	P-значение	
const	-38,4019	7,40425	-5,1865	<0,00001	***
FUNG1	0,0445755	0,0629524	0,7081	0,47975	
FUNG2	0,103625	0,0624082	1,6604	0,09846	*
GIRB	0,0776059	0,0623777	1,2441	0,21497	
INSEC	0,0782521	0,0536527	1,4585	0,14634	
LABOUR	0,0415064	0,00300121	13,8299	<0,00001	***
YDOB1	0,0492168	0,0197491	2,4921	0,01355	**
YDOB2	-0,0906824	0,030999	-2,9253	0,00386	***
Сумма кв. остатков	150575,6		Ст. ошибка модели	28,00443	
R-квадрат	0,801958		Испр. R-квадрат	0,794738	
F(7, 192)	119,2263		P-значение (F)	2,16e-66	

Оценки коэффициентов по сравнению с пунктом (а) не поменялись, что естественно: мы ведь по-прежнему используем обычный МНК. Однако стандартные ошибки теперь немного другие. В некоторых случаях это меняет выводы тестов на незначимость.

Переменные LABOUR, YDOB1 и YDOB2 значимы на пятипроцентном уровне значимости (причем LABOUR и YDOB2 — ещё и на однопроцентном).

Переменная FUNG2 перестала быть значимой на пятипроцентном уровне. Таким образом, при использовании корректных стандартных ошибок следует сделать вывод о том, что соответствующий вид удобрений не важен для урожайности. Обратите внимание, что если бы мы использовали «обычные» стандартные ошибки, то мы пришли бы к противоположному заключению (см. пункт (а)).

* * *

Важно подчеркнуть, что в реальных пространственных данных гетероскедастичность в той или иной степени наблюдается практически всегда. А даже если её и нет, то состоятельные в условиях гетероскедастичности стандартные ошибки по-прежнему будут... состоятельными (и будут близки к «обычным» стандартным ошибкам, посчитанным по формулам из третьей главы). Поэтому в современных прикладных исследованиях при оценке уравнений по умолчанию используются именно робастные стандартные ошибки, а не стандартные ошибки для случая гомоскедастичности. Мы настоятельно рекомендуем читателю поступать так же ². В нашем учебнике с этого момента и во всех последующих главах, если прямо не оговорено иное, для МНК-оценок параметров **всегда** используются состоятельные в условиях гетероскедастичности стандартные ошибки.

1. Поскольку довольно утомительно каждый раз произносить это название полностью в англоязычном варианте их часто называют просто robust standard errors, что на русском языке эконометристов превратилось в «робастные стандартные ошибки». Кому-то подобный англицизм, конечно, режет слух, однако в устной речи он и правда куда удобней своей длинной альтернативы. [↪](#)

2. Просто не забывайте включать соответствующую опцию в своем эконометрическом пакете. [↪](#)

5.3. Взвешенный метод наименьших квадратов

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

Рассмотренные в предыдущем параграфе стандартные ошибки позволяют успешно тестировать гипотезы и строить доверительные интервалы в условиях гетероскедастичности, однако не устраняют другого её негативного последствия, упомянутого в начале главы: неэффективности МНК-оценок. Для получения эффективных оценок параметров можно воспользоваться так называемым **взвешенным МНК** (weighted least squares, WLS). Чтобы понять, как он работает, рассмотрим несколько важных случаев.

Случай 1. Дисперсия случайных ошибок $\text{var}(\varepsilon_i) = \sigma_i^2$ известна

Пусть рассматривается модель

$$y_i = \beta_1 + \beta_2 x_i^{(2)} + \dots + \beta_k x_i^{(k)} + \varepsilon_i$$

для которой выполнены все предпосылки классической линейной модели множественной регрессии за одним исключением: в данных наблюдается гетероскедастичность $\text{var}(\varepsilon_i) = \sigma_i^2$.

В этом случае можно разделить правую и левую часть уравнения регрессии на σ_i :

$$\frac{y_i}{\sigma_i} = \frac{\beta_1 + \beta_2 x_i^{(2)} + \dots + \beta_k x_i^{(k)} + \varepsilon_i}{\sigma_i}$$

После этого сделаем простую замену переменных:

$$\tilde{y}_i = \frac{y_i}{\sigma_i}; \tilde{x}_i^{(1)} = \frac{1}{\sigma_i}; \tilde{x}_i^{(2)} = \frac{x_i^{(2)}}{\sigma_i}; \dots; \tilde{x}_i^{(k)} = \frac{x_i^{(k)}}{\sigma_i}; \tilde{\varepsilon}_i = \frac{\varepsilon_i}{\sigma_i}$$

В результате замены переменных переходим к новой модели:

$$\tilde{y}_i = \beta_1 \tilde{x}_i^{(1)} + \beta_2 \tilde{x}_i^{(2)} + \dots + \beta_k \tilde{x}_i^{(k)} + \tilde{\varepsilon}_i$$

Новая модель полезна тем, что в ней гетероскедастичности нет, так как дисперсия случайной ошибки является константой:

$$\text{var}(\tilde{\varepsilon}_i) = \text{var}\left(\frac{\varepsilon_i}{\sigma_i}\right) = \frac{1}{\sigma_i^2} \text{var}(\varepsilon_i) = \frac{1}{\sigma_i^2} \sigma_i^2 = 1 = \text{const}$$

Следовательно, МНК, примененный к новой модели, будет давать не только несмещенный, но и эффективный результат. Таким образом, суть взвешенного МНК состоит в том, чтобы сделать правильную замену переменных так, чтобы применение к новой модели (с измененными переменными) обычного МНК приводило к получению эффективных оценок коэффициентов. После этого для интерпретации результатов можно вернуться к исходным переменным.

Чтобы понять, почему этот метод называется взвешенным МНК, сравним оптимизационные задачи в рамках обычного МНК и в рамках взвешенного МНК. В первом случае мы минимизируем сумму квадратов остатков:

$$\sum_{i=1}^n e_i^2 \rightarrow \min_{\hat{\beta}}$$

В случае взвешенного МНК мы минимизируем сумму квадратов остатков новой модели:

$$\sum_{i=1}^n (\tilde{e}_i)^2 = \sum_{i=1}^n \left(\frac{e_i}{\sigma_i}\right)^2 = \sum_{i=1}^n \frac{1}{\sigma_i^2} e_i^2 \rightarrow \min_{\hat{\beta}}$$

Получается, что мы минимизируем сумму квадратов остатков, но каждое слагаемое домножается на весовой коэффициент $1/\sigma_i^2$, то есть мы минимизируем сумму квадратов остатков с определенными весами. Чем меньше дисперсия для i -го наблюдения (то есть чем меньше фактор случайности для этого наблюдения), тем больший вес это наблюдение имеет в той сумме, которую мы минимизируем. Тем самым наибольший вес мы придаем наиболее «надежным» наблюдениям. Это и позволяет улучшить качество получаемых оценок.

Обратите внимание, что даже если в исходной модели был свободный член, в новой модели в результате замены переменных константа пропадает (вместо неё возникает переменная $\tilde{x}_i^{(1)} = \frac{1}{\sigma_i}$). Поэтому, в частности, привычный коэффициент R^2 для такой модели неприменим и не может интерпретироваться стандартным образом (хотя, если его все-таки посчитать, он часто оказывается выше по сравнению с аналогичным коэффициентом для обычного МНК).

Разумеется, в реальности дисперсия случайной ошибки обычно не известна исследователю, что приводит нас к необходимости рассмотрения более реалистичного случая.

Случай 2. Дисперсия случайных ошибок $\text{var}(\varepsilon_i) = \sigma_i^2$ не известна

В этой ситуации сначала следует получить оценки дисперсий σ_i^2 . Как правило, это делают так:

Сначала оценивают исходную модель при помощи обычного МНК и получают остатки регрессии e_i .

Затем оценивают вспомогательную модель для остатков следующего вида:

$$e_i^2 = \gamma_0 + \gamma_1 z_i^{(1)} + \dots + \gamma_p z_i^{(p)} + u_i \text{ или } \ln e_i^2 = \gamma_0 + \gamma_1 z_i^{(1)} + \dots + \gamma_p z_i^{(p)} + u_i.$$

Здесь $z_i^{(1)}, z_i^{(2)}, \dots, z_i^{(p)}$ — набор переменных, которые предположительно влияют на дисперсию случайной ошибки. Обычно в качестве таких переменных берутся регрессоры из исходной модели, а также их квадраты.

$\ln e_i^2$ иногда используется в левой части вспомогательного уравнения вместо e_i^2 для того, чтобы предсказанное значение квадрата остатков никогда не было отрицательным (что было бы нелогично).

Оценив вспомогательное уравнение, получаем предсказанные значения квадратов остатков $\hat{e}_i^2$. Их и используют, чтобы оценить дисперсию случайной ошибки:

$$\hat{\sigma}_i^2 = \hat{e}_i^2$$

После этого следует действовать аналогично первому случаю, только вместо дисперсии σ_i^2 брать для замены переменных её оценку $\hat{\sigma}_i^2$.

Пример 5.2. Оценка эффективности использования удобрений (продолжение)

Продолжите рассмотрение модели урожайности, которое мы начали в примере 5.1. Теперь оцените модель, используя взвешенный МНК. Сравните полученные результаты с результатами из примера 5.1.

Решение:

В качестве вспомогательного уравнения оценивалась следующая спецификация, включающая все регрессоры исходной модели, а также их квадраты:

$$\begin{aligned} \ln e_i^2 = & \gamma_0 + \gamma_1 \text{FUNG1}_i + \gamma_2 \text{FUNG2}_i + \gamma_3 \text{GIRB}_i + \\ & + \gamma_4 \text{INSEC}_i + \gamma_5 \text{LABOUR}_i + \gamma_6 \text{YDOB1}_i + \gamma_7 \text{YDOB2}_i + \\ & + \gamma_8 \text{FUNG1}_i^2 + \gamma_9 \text{FUNG2}_i^2 + \gamma_{10} \text{GIRB}_i^2 + \\ & + \gamma_{11} \text{INSEC}_i^2 + \gamma_{12} \text{LABOUR}_i^2 + \gamma_{13} \text{YDOB1}_i^2 + \gamma_{14} \text{YDOB2}_i^2 + u_i \end{aligned}$$

В результате оценки этого уравнения были получены расчетные значения $\widehat{\ln e_i^2}$, затем, $\widehat{e_i^2}$. После этого был осуществлен переход к взвешенной модели, как это описано выше, при помощи деления обеих частей уравнения из примера 5.1 на $\hat{\sigma}_i = \sqrt{\hat{e_i^2}}$. Наконец, при помощи обычного МНК были оценены параметры последней модели. Результаты оценивания представлены ниже.

Модель 3: С поправкой на гетероскедастичность, использованы наблюдения 1-200

Зависимая переменная: PRODP

	Коэффициент	Ст. ошибка	t-статистика	P-значение	
const	-36,6762	6,17857	-5,9360	<0,00001	***
FUNG1	0,0827277	0,0496946	1,6647	0,09760	*
FUNG2	0,114722	0,0519986	2,2063	0,02855	**
GIRB	0,0528989	0,0566521	0,9338	0,35161	
INSEC	0,0447185	0,0424588	1,0532	0,29356	
LABOUR	0,0411533	0,0026617	15,4613	<0,00001	***
YDOB1	0,0412047	0,0199716	2,0632	0,04044	**
YDOB2	-0,0817552	0,0232718	-3,5131	0,00055	***

Статистика, полученная по взвешенным данным:

Сумма кв. остатков	798,3622		Ст. ошибка модели	2,039151
F(7, 192)	169,9013		P-значение (F)	1,05e-78

То же самое в виде уравнения:

$$\begin{aligned} \widehat{PRODP}_i = & -36,68 + 0,08 \text{FUNG1}_i + 0,11 \text{FUNG2}_i + \\ & \quad (6,18) \quad (0,05) \quad (0,05) \\ & + 0,05 \text{GIRB}_i + 0,04 \text{INSEC}_i + 0,04 \text{LABOUR}_i + \\ & \quad (0,06) \quad (0,04) \quad (0,003) \\ & + 0,04 \text{YDOB1}_i - 0,08 \text{YDOB2}_i. \\ & \quad (0,02) \quad (0,02) \end{aligned}$$

При оценивании модели с коррекцией на гетероскедастичность изменились коэффициенты, а их стандартные ошибки в целом стали меньше.

Напомним, что обычная гетероскедастичность не приводит к смещению оценок коэффициентов, поэтому неудивительно, что полученные результаты совсем немного отличаются от оценок коэффициентов, вычисленных на основе обычного МНК (см. числа

в примере 5.1). Тем не менее, если мы корректно оценили дисперсию случайной ошибки, есть надежда, что новые результаты являются немного более точными.

* * *

В заключение данного параграфа рассмотрим еще один случай применения взвешенного МНК. Этот случай является более частным, чем случай 2, однако также может быть полезен в некоторых ситуациях.

Случай 3. Дисперсия случайной ошибки прямо пропорциональна квадрату единственной переменной: $\text{var}(\epsilon_i) = \sigma_i^2 = \sigma_0^2 z_i^2 > 0$.

Подразумевается, что величина σ_0^2 не известна, но это нам не мешает. Оказывается, столкнувшись с таким частным случаем гетероскедастичности, можно легко её устранить. Действительно, для этого достаточно поделить правую и левую части исходного уравнения на переменную z_i :

$$\frac{y_i}{z_i} = \frac{\beta_1 + \beta_2 x_i^{(2)} + \dots + \beta_k x_i^{(k)} + \epsilon_i}{z_i}$$

Делаем замену переменных:

$$y_i^* = \frac{y_i}{z_i}; \quad x_i^{(1)*} = \frac{1}{z_i}; \quad x_i^{(2)*} = \frac{x_i^{(2)}}{z_i}; \quad \epsilon_i^* = \frac{\epsilon_i}{z_i}$$

В результате замены переменных переходим к новой модели со звездочками:

$$y_i^* = \beta_1 x_i^{(1)*} + \beta_2 x_i^{(2)*} + \dots + \beta_k x_i^{(k)*} + \epsilon_i^*$$

В новой модели гетероскедастичности нет, так как дисперсия случайной ошибки является константой:

$$\text{var}(\epsilon_i) = \text{var}\left(\frac{\epsilon_i}{z_i}\right) = \frac{1}{z_i^2} \text{var}(\epsilon_i) = \frac{1}{z_i^2} \sigma_0^2 z_i^2 = \sigma_0^2 = \text{const}$$

Следовательно, для оценки параметров новой модели можно использовать обычный МНК, и оценки коэффициентов будут эффективными.

Какой же метод устранения негативных последствий гетероскедастичности лучше использовать: состоятельные в условиях гетероскедастичности стандартные ошибки или взвешенный МНК?

Казалось бы, взвешенный МНК предпочтительней, потому что он дает возможность получить эффективные оценки коэффициентов, то есть в ситуации гетероскедастичности он должен быть точнее обычного МНК. Однако это верно только при условии, что мы правильно специфицировали уравнение для дисперсии случайной ошибки (то есть правильно поняли, как именно устроена гетероскедастичность в анализируемой модели). К сожалению, на практике это может быть проблематично. Кроме того, при наличии достаточно большой выборки обычный МНК и так дает удовлетворительные результаты. Поэтому в прикладных эконометрических работах гораздо чаще применяется обычный МНК в сочетании с робастными стандартными ошибками.

5.4. Выявление гетероскедастичности

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

[Скрипт в R для этого примера](#)

[Данные для этого примера](#)

Гетероскедастичность — это типичная «болезнь» пространственных данных, поэтому лучше по умолчанию исходить из того, что она в вашей модели есть. Тем не менее, иногда бывает полезно уметь аккуратно проверить её наличие. Для этого можно использовать два традиционных подхода: графический анализ данных и формальные статистические тесты.

Один из способов выявления гетероскедастичности при помощи графического анализа состоит в том, чтобы построить диаграммы рассеяния, в каждой из которых по оси ординат стоит зависимая переменная, а по оси абсцисс — один из регрессоров. Если, разглядывая подобную диаграмму, вы видите нечто похожее на рисунок 2.3б, то у вас есть гетероскедастичность, связанная с соответствующим регрессором. Другой вариант состоит в анализе графика остатков регрессии. Отсортируйте остатки по возрастанию какой-либо объясняющей переменной и постройте их график. Если разброс остатков вокруг нуля равномерен (как, например, на рис. 5.1), то можно заключить, что гетероскедастичность, связанная с этим регрессором, в данных отсутствует. Если же на графике остатков явно видно, что их разброс вокруг нуля зависит от значения регрессора (как, например, на рис. 5.2), значит, гетероскедастичность есть.

Остатки

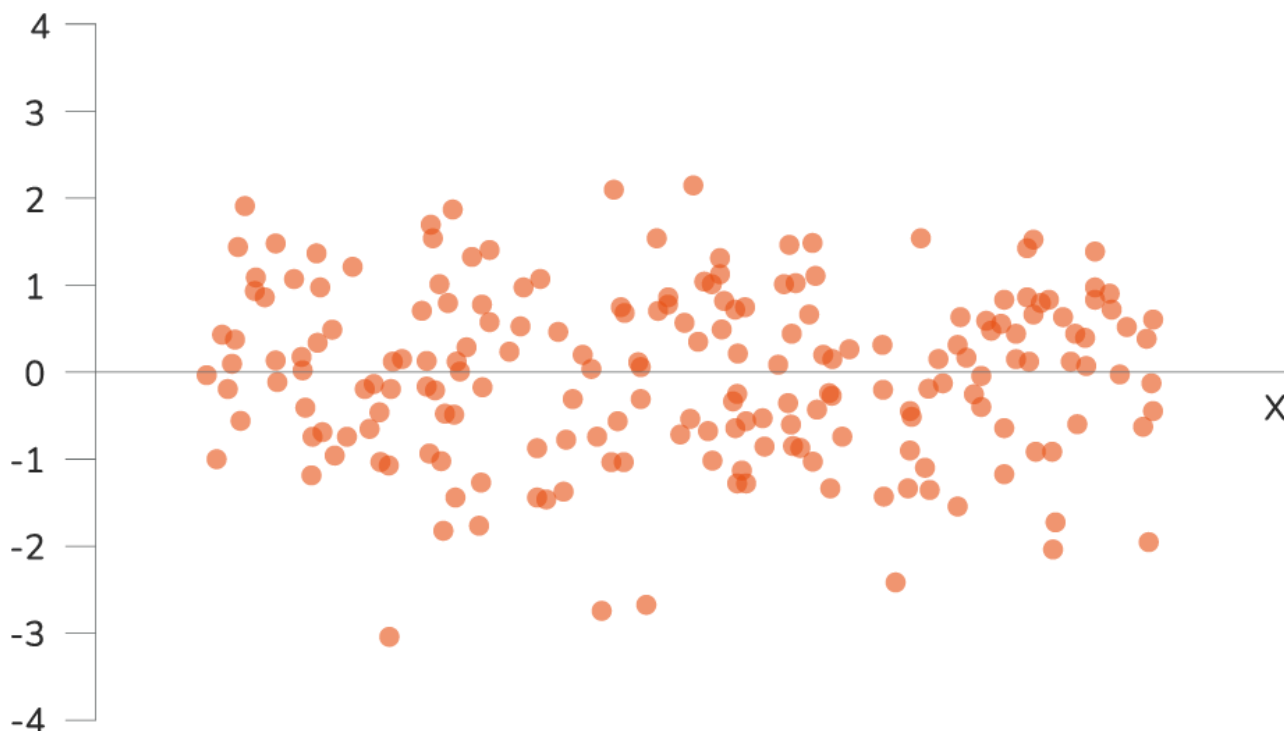


Рисунок 5.1. Поведение остатков регрессии говорит в пользу гомоскедастичности

Остатки

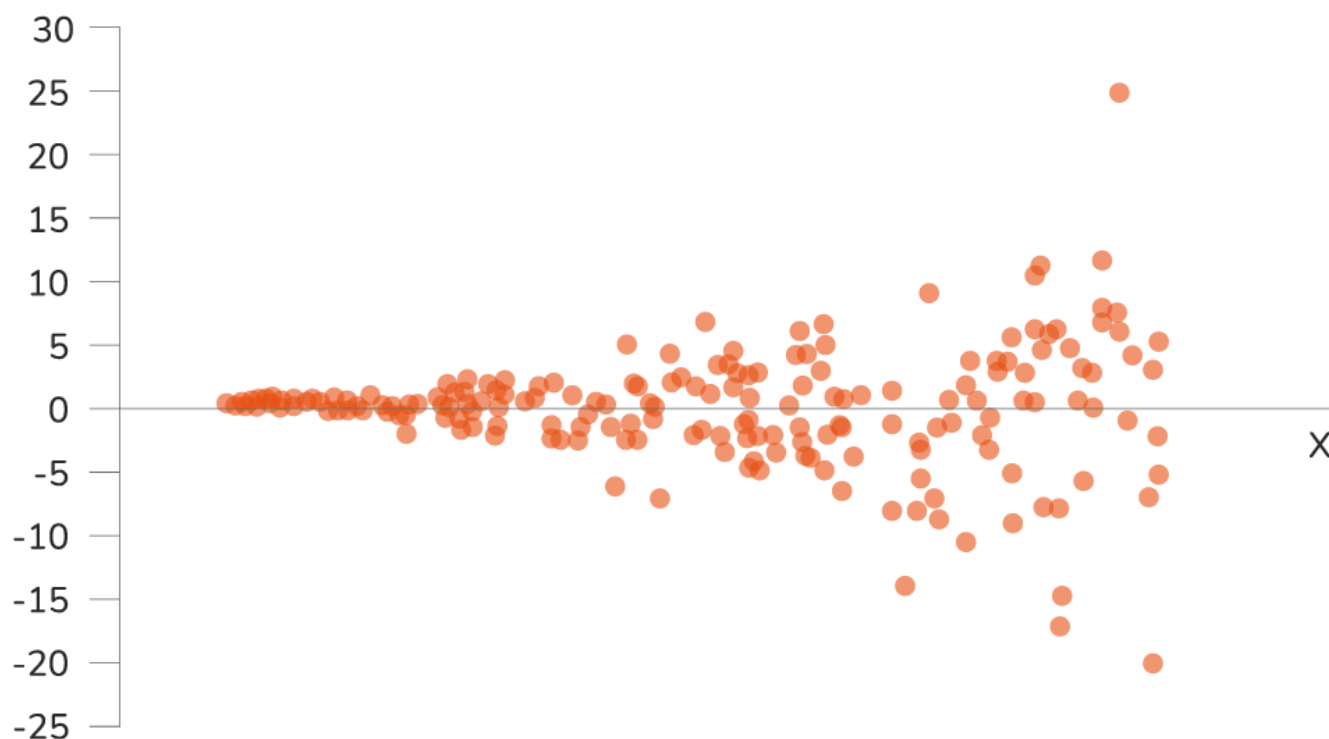


Рисунок 5.2. Поведение остатков регрессии говорит в пользу гетероскедастичности

Анализ графиков не всегда позволяет сделать однозначный вывод по поводу наличия или отсутствия гетероскедастичности, поэтому помимо него могут быть полезны соответствующие формальные статистические тесты. Ниже приводятся два наиболее часто используемых в настоящее время теста.

Тест Бреуша — Пагана

Тестируемая гипотеза в данном тесте состоит в том, что гетероскедастичности в модели нет:

$$H_0 : \sigma_1^2 = \dots = \sigma_n^2$$

Альтернативная гипотеза — дисперсия случайной ошибки ϵ_i некоторым образом зависит от группы переменных:

$$H_1 : \sigma_i^2 = \gamma_0 + \gamma_1 z_i^{(1)} + \dots + \gamma_p z_i^{(p)}$$

Здесь $z_i^{(1)}, z_i^{(2)}, \dots, z_i^{(p)}$ — набор переменных, которые предположительно влияют на дисперсию случайной ошибки. Обычно в качестве таких переменных берутся регрессоры из исходной модели, а также их квадраты.

Процедура осуществления теста устроена так: сначала при помощи обычного МНК оценивается исходная модель (для которой мы хотим проверить отсутствие гетероскедастичности) и вычисляются соответствующие остатки e_i . Далее вычисляется вспомогательное значение $\tilde{\sigma}^2 = \frac{1}{n} \sum e_i^2$. После этого необходимо оценить вспомогательное уравнение, в котором справа стоят переменные, потенциально влияющие на дисперсию случайной ошибки:

$$\frac{e_i^2}{\tilde{\sigma}^2} = \gamma_0 + \gamma_1 z_i^{(1)} + \dots + \gamma_p z_i^{(p)} + u_i$$

Далее вычисляется расчетное значение тестовой статистики по простой формуле: половина от объясненной суммы квадратов во вспомогательном уравнении.

Если верна нулевая гипотеза, то указанная статистика асимптотически имеет распределение Хи-квадрат с p степенями свободы. Поэтому, если расчетное значение больше критического значения, взятого из таблиц распределения χ^2 с p степенями свободы для выбранного исследователем уровня значимости, то следует отвергнуть нулевую гипотезу и заключить, что в данных есть гетероскедастичность (необходимые таблицы доступны, например, в приложении к главе 6). В противном случае можно сделать вывод в пользу гомоскедастичности.

Тест Уайта

Тестируемая гипотеза в данном тесте снова состоит в том, что гетероскедастичности в модели нет:

$$H_0 : \sigma_1^2 = \dots = \sigma_n^2$$

Альтернативная гипотеза — дисперсия случайной ошибки ϵ_i произвольным (возможно, нелинейным) образом зависит от переменных модели.

Процедура теста устроена так: сначала при помощи обычного МНК оценивается исходная модель (для которой мы хотим проверить отсутствие гетероскедастичности) и вычисляются соответствующие остатки e_i . После этого необходимо оценить вспомогательное уравнение, в котором слева стоит e_i^2 , а справа — константа, регрессоры исходного уравнения, их квадраты и попарные произведения¹.

Далее вычисляется расчетное значение тестовой статистики по следующей формуле

$$(R^2_{\text{вспомогательному уравнению}}) * n.$$

Если верна нулевая гипотеза, то указанная статистика асимптотически имеет распределение Хи-квадрат с p степенями свободы (p — число регрессоров во вспомогательном уравнении). Поэтому, если расчетное значение больше критического значения, взятого из таблиц распределения χ^2 с p степенями свободы для выбранного исследователем уровня значимости, то следует отвергнуть нулевую гипотезу и заключить, что в данных есть гетероскедастичность. В противном случае можно сделать вывод в пользу гомоскедастичности².

Закончить обсуждение вопроса выявления гетероскедастичности следует предостережением по поводу **ложной гетероскедастичности**. Ложной гетероскедастичностью называется ситуация, при которой формальные тесты указывают на наличие гетероскедастичности, однако в действительности дело вовсе не в ней, а в неверной спецификации уравнения. Хорошим примером может служить рисунок 4.5б, на котором представлена нелинейная зависимость между парой переменных. Если при этом ошибочно оценить линейную регрессию (соответствующая прямая линия изображена на рисунке), то статистические тесты будут говорить в пользу гетероскедастичности, так как поведение остатков технически будет зависеть от значения регрессора (см. нижний график на этом же рисунке). Однако в действительности гетероскедастичности в модели нет, а есть только нелинейная связь между переменными.

Важно различать истинную и ложную гетероскедастичность, так как они приводят к совершенно разным последствиям. Истинная гетероскедастичность не вызывает смещения оценок коэффициентов модели, в то время как ошибочная спецификация уравнения регрессии вызывает его, то есть является гораздо более серьезной проблемой.

Пример 5.3. Оценка эффективности использования удобрений (окончание)

Для модели, оцененной в примере 5.1, осуществите тест Уайта, используя пятипроцентный уровень значимости. Интерпретируйте полученные результаты.

Решение:

Результаты оценки вспомогательного уравнения для осуществления теста Уайта представлены ниже.

Тест Уайта на гетероскедастичность

MNK, использованы наблюдения 1-200

Зависимая переменная: квадраты остатков регрессии, оцененной в примере 5.1

Коэффициент Ст. ошибка t-статистика P-значение

```
-----
const -2270,16 5654,83 -0,4015 0,6886
FUNG1 0,482935 108,468 0,004452 0,9965
FUNG2 -150,789 95,3876 -1,581 0,1158
GIRB 1,82902 98,0229 0,01866 0,9851
INSEC 129,035 72,6436 1,776 0,0775 *
LABOUR 2,05511 4,84800 0,4239 0,6722
YD0B1 30,5354 43,6171 0,7001 0,4849
YD0B2 3,43369 50,1279 0,06850 0,9455
квадрат_FUNG1 0,0956897 0,0549105 1,743 0,0833 *
FUNG1*FUNG2 0,0294522 0,0492979 0,5974 0,5510
FUNG1*GIRB -0,0332633 0,0480171 -0,6927 0,4895
FUNG1*INSEC 0,0229300 0,0554156 0,4138 0,6796
FUNG1*LABOUR -0,00259087 0,00348456 -0,7435 0,4582
FUNG1*YD0B1 0,0104778 0,0246791 0,4246 0,6717
FUNG1*YD0B2 -0,0536699 0,0501448 -1,070 0,2861
квадрат_FUNG2 0,0919819 0,0646201 1,423 0,1565
FUNG2*GIRB -0,0931636 0,0529473 -1,760 0,0803 *
FUNG2*INSEC -0,0878293 0,0548646 -1,601 0,1113
FUNG2*LABOUR -0,00520969 0,00419928 -1,241 0,2165
FUNG2*YD0B1 0,0829467 0,0471221 1,760 0,0802 *
FUNG2*YD0B2 -0,0118900 0,0437395 -0,2718 0,7861
квадрат_GIRB -0,0598434 0,0581283 -1,030 0,3048
GIRB*INSEC -0,0361947 0,0561232 -0,6449 0,5199
GIRB*LABOUR 0,00279620 0,00413522 0,6762 0,4999
GIRB*YD0B1 0,0287539 0,0384965 0,7469 0,4562
```

```
GIRB*YD0B2 0,0537695 0,0489420 1,099 0,2735
квadrat_INSEC -0,0406052 0,0570708 -0,7115 0,4778
INSEC*LABOUR -0,00562133 0,00477862 -1,176 0,2412
INSEC*YD0B1 0,0367439 0,0296061 1,241 0,2163
INSEC*YD0B2 -0,0599689 0,0474392 -1,264 0,2080
квadrat_LABOUR -0,000342326 0,000950427 -0,3602 0,7192
LABOUR*YD0B1 -0,00130669 0,00215676 -0,6059 0,5454
LABOUR*YD0B2 0,00492378 0,00240166 2,050 0,0419 **
квadrat_YD0B1 -0,0176598 0,00913794 -1,933 0,0550 *
YD0B1*YD0B2 -0,0352604 0,0255207 -1,382 0,1690
квadrat_YD0B2 0,0194422 0,0232499 0,8362 0,4042
Неисправленный R-квadrat = 0,280856
Тестовая статистика: n*R-квadrat = 56,171138,
p-значение = P(Хи-квadrat(35) > 56,171138) = 0,013059
```

В учебных целях мы привели оцененное уравнение полностью, хотя обычно в этом нет нужды, так как для осуществления теста достаточно знать только количество переменных в этом уравнении, его R-квadrat и число наблюдений. Обратите внимание, что число регрессоров тут действительно велико из-за добавления квадратов и попарных произведений переменных из исходного уравнения.

В нашем случае Р-значение для осуществляемого теста (представленное в самом низу таблицы с результатами) составляет 0,013. Это меньше, чем 0,05. Поэтому при уровне значимости 5% следует отвергнуть нулевую гипотезу данного теста. Напомним, что нулевая гипотеза теста Уайта состоит в том, что в модели нет гетероскедастичности. Следовательно, отвергая её, мы должны заключить: в нашем случае в данных наблюдается гетероскедастичность.

-
1. Обратите внимание, что добавлять нужно только такие квадраты и попарные произведения, включение которых в модель не приводит к чистой мультиколлинеарности. Например, квадраты фиктивных переменных добавлять не стоит, так как они будут принимать в точности такие же значения, что и исходные переменные ($0^2 = 0$ и $1^2 = 1$). ↵
 2. В XX веке для выявления гетероскедастичности использовался широкий спектр альтернативных тестов: Голдфелда — Квандта, Спирмена, Глейзера и Парка. Они остались за рамками этой книги, поскольку в современных исследованиях применяются редко. ↵

5.5. Обобщенная линейная модель и обобщенный МНК

В предыдущих параграфах данной главы мы рассматривали случай отказа от одной предпосылки классической линейной модели множественной регрессии. Теперь мы расширим наш анализ и откажемся сразу от двух предпосылок: от предпосылок №№4-5 (о постоянстве дисперсии случайной ошибки и о некоррелированности разных случайных ошибок между собой). В техническом смысле этот параграф несколько сложнее предыдущих (в частности, тут более широко используется линейная алгебра). Поэтому, если вы заинтересованы в том, чтобы разобраться только в прикладных аспектах множественной регрессии, а в соответствующих вычислениях готовы полностью довериться эконометрическому пакету, можете его пропустить.

Отказ от указанных двух предпосылок означает, что ковариационная матрица вектора случайных ошибок (таблица, в которой записаны все ковариации между ε_i и ε_j , см. параграф 3.3) больше не является диагональной матрицей с одинаковыми числами на главной диагонали, как это было в первоначальной классической модели. Теперь ковариационная матрица вектора случайных ошибок Ω — это произвольная ковариационная матрица (разумеется, так как это не совсем любая матрица, а именно ковариационная матрица, то по своим свойствам она является симметричной и положительно определенной).

Модель, в которой сохранены только первые три предпосылки классической линейной модели множественной регрессии, называется **обобщенной линейной моделью множественной регрессии**.

Проанализируем, к каким последствиям приводит отказ от предпосылок №№4-5.

Во-первых, полученная обычным методом наименьших квадратов оценка $\hat{\beta} = (X'X)^{-1}X'y$ остается несмещенной (это свойство мы доказывали, опираясь как раз лишь на первые три предпосылки).

Во-вторых, МНК-оценки хоть и остаются несмещенными, но больше не являются эффективными.

В-третьих, если мы оцениваем ковариационную матрицу вектора оценок коэффициентов (которая нужна для тестирования всевозможных гипотез), то оценка $\hat{V}(\hat{\beta}) = (X'X)^{-1}S^2$ смещена и больше не является корректной.

Чтобы убедиться в этом, посчитаем ковариационную матрицу от $\hat{\beta}$ в условиях обобщенной модели (при этом мы используем свойства ковариационной матрицы, перечисленные в параграфе 3.3):

$$\begin{aligned} V(\hat{\beta}) &= V[(X'X)^{-1}X'y] = V[(X'X)^{-1}X'(X\beta + \varepsilon)] = \\ &= V[(X'X)^{-1}X'\varepsilon] = (X'X)^{-1}X'V[\varepsilon](X'X)^{-1}X' = (X'X)^{-1}X'\Omega(X'X)^{-1}X' = (X'X)^{-1}X'\Omega X(X'X)^{-1} \end{aligned}$$

Так выглядит ковариационная матрица вектора МНК-оценок в обобщенной модели. Ясно, что она не может быть корректно оценена стандартной оценкой $(X'X)^{-1}S^2$. Следовательно, прежней формулой пользоваться нельзя: если мы будем использовать стандартные ошибки, рассчитанные по обычной формуле (предполагая выполнение предпосылок классической линейной модели), то получим некорректные стандартные ошибки, что может привести нас к неверным выводам по поводу значимости или незначимости тех или иных регрессоров.

Таким образом, последствия перехода к обобщенной модели аналогичны тем, что мы наблюдали для случая гетероскедастичности. Это неудивительно, так как гетероскедастичность — частный случай обобщенной линейной модели.

Поэтому для получения эффективных оценок обычный МНК не подойдет, и придется воспользоваться альтернативным методом — **обобщенным МНК** (ОМНК, generalized least squares, GLS). Формулу для расчета оценок коэффициентов при помощи ОМНК позволяет получить специальная теорема.

Теорема Айткена

Если

1. модель линейна по параметрам и правильно специфицирована
 $y = X\beta + \varepsilon$,
2. матрица X — детерминированная матрица, имеющая максимальный ранг k ,
3. $E(\varepsilon) = \vec{0}$,
4. $V(\varepsilon) = \Omega$ — произвольная положительно определенная и симметричная матрица,

то оценка вектора коэффициентов модели $\hat{\beta} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$ является:

- несмещенной
- и эффективной, то есть имеет наименьшую ковариационную матрицу в классе всех несмещенных и линейных по y оценок.

Предпосылки теоремы Айткена — это предпосылки обобщенной линейной модели множественной регрессии. Из них первые три — стандартные, как в классической модели, а четвертая ничего особого не требует (y вектора случайных ошибок может быть любая ковариационная матрица без каких-либо дополнительных специальных ограничений). Сама теорема Айткена является аналогом теоремы Гаусса — Маркова для случая обобщенной модели.

Докажем эту теорему.

Из линейной алгебры известно: если матрица Ω симметрична и положительно определена, то существует такая матрица P , что

$$P' \bullet P = \Omega^{-1} \Leftrightarrow P' = \Omega^{-1} \bullet P^{-1}$$

А раз такое представление возможно, то воспользуемся им для замены переменных. От вектора значений зависимой переменной y , перейдем к вектору $P \bullet y$, обозначив его как вектор $\tilde{y} = P \bullet y$. Аналогичным образом введем матрицу $\tilde{X} = P \bullet X$ и вектор ошибок $\tilde{\varepsilon} = P \bullet \varepsilon$.

Вернемся к исходной модели, параметры которой нас и интересуют:

$$y = X\beta + \varepsilon$$

Умножим левую и правую части равенства на матрицу P :

$$Py = PX\beta + P\varepsilon$$

С учетом новых обозначений это равенство можно записать так:

$$y = X\beta + \varepsilon$$

Для новой модели (со звездочками) выполняются предпосылки теоремы Гаусса-Маркова. Чтобы в этом убедиться, достаточно показать, что математическое ожидание вектора случайных ошибок является нулевым вектором (третья предпосылка классической модели) и ковариационная матрица вектора случайных ошибок является диагональной с одинаковыми элементами на главной диагонали (четвертая и пятая предпосылки).

Для этого вычислим математическое ожидание нового вектора ошибок:

$$E(\varepsilon) = E(P\varepsilon) = P \bullet E(\varepsilon) = P \bullet \vec{0} = \vec{0}$$

Теперь вычислим ковариационную матрицу вектора ε :

$$V(\varepsilon) = V(P \bullet \varepsilon) = P \bullet V(\varepsilon) \bullet P' = P \bullet \Omega \bullet P' = P \bullet \Omega \bullet \Omega^{-1} \bullet P^{-1} = I_n$$

Здесь I_n обозначает единичную матрицу размера n на n .

Следовательно, для модели со звездочками выполняются все предпосылки теоремы Гаусса — Маркова. Поэтому получить несмещенную и эффективную оценку вектора коэффициентов можно, применив к этой измененной модели обычный МНК:

$$\hat{\beta} = (X'X)^{-1} X'y$$

Теперь осталось вернуться к исходным обозначениям, чтобы получить формулу несмещенной и эффективной оценки интересующего нас вектора в терминах обобщенной модели:

$$\hat{\beta} = (X'X)^{-1} X'y = ((PX)'PX)^{-1} (PX)'Py = (X'P'PX)^{-1} X'P'Py = (X'\Omega^{-1}P^{-1}PX)^{-1} X'\Omega^{-1}P^{-1}Py = (X'\Omega^{-1}X)^{-1} X'y$$

Что и требовалось доказать.

Взвешенный МНК, который мы обсуждали ранее, — это частный вариант обобщенного МНК (для случая, когда только предпосылка №4 нарушена, а предпосылка №5 сохраняется).

Как и при использовании взвешенного МНК в ситуации применения ОМНК коэффициент R-квадрат не обязан лежать между нулем и единицей и не может быть интерпретирован стандартным образом.

Слабая сторона ОМНК состоит в том, что для его реализации нужно знать не только матрицу регрессоров X с вектором значений зависимой переменной y , но и ковариационную матрицу вектора случайных ошибок Ω . На практике, однако, эта матрица почти никогда не известна. Поэтому в прикладных исследованиях практически всегда вместо ОМНК используется, так называемый, **доступный ОМНК** (его ещё называют **практически реализуемый ОМНК**, feasible GLS). Идея доступного ОМНК состоит в том, что следует сначала оценить матрицу Ω (традиционно обозначим её оценку $\hat{\Omega}$), а уже затем получить оценку вектора коэффициентов модели, заменив в формуле ОМНК Ω на $\hat{\Omega}$:

$$\hat{\beta} = (X'\hat{\Omega}^{-1}X)^{-1} X'\hat{\Omega}^{-1}y.$$

Применение этого подхода осложняется тем, что $\hat{\Omega}$ не может быть оценена непосредственно без дополнительных предпосылок, так как в ней слишком много неизвестных элементов. Действительно, в матрице размер n на n всего n^2 элементов, и оценить их все, имея всего n наблюдений, представляется слишком амбициозной задачей. Даже если воспользоваться тем, что матрица Ω является симметричной, в результате чего достаточно оценить только элементы на главной диагонали и над ней, мы все равно столкнемся с необходимостью оценивать $(n+1)n/2$ элементов, что всегда больше числа доступных нам наблюдений.

Поэтому процедура доступного ОМНК устроена так:

1. Делаются некоторые предпосылки по поводу того, как устроена ковариационная матрица вектора случайных ошибок Ω . На основе этих предпосылок оценивается матрица $\hat{\Omega}$.
2. После этого по формуле $(X'\hat{\Omega}^{-1}X)^{-1} X'\hat{\Omega}^{-1}y$ вычисляется вектор оценок коэффициентов модели.

Из сказанного следует, что доступный ОМНК может быть реализован только в ситуации, когда есть разумные основания сформулировать те или иные предпосылки по поводу матрицы $\hat{\Omega}$. Рассмотрим некоторые примеры таких ситуаций.

Пример 5.4. Автокорреляция и ОМНК-оценка

Рассмотрим линейную модель $y = X\beta + \varepsilon$, для которой дисперсия случайных ошибок постоянна, однако наблюдается так называемая автокорреляция первого порядка:

$$\varepsilon_i = \rho * \varepsilon_{i-1} + u_i$$

Здесь u_i — независимые и одинаково распределенные случайные величины с дисперсией σ_u^2 , а $\rho \in (-1, 1)$ — коэффициент автокорреляции.

(а) Найдите ковариационную матрицу вектора случайных ошибок для представленной модели.

(б) Запишите в явном виде формулу ОМНК-оценки вектора коэффициентов модели, предполагая, что коэффициент ρ известен.

Примечание: в отличие от гетероскедастичности, автокорреляция случайных ошибок обычно наблюдается не в пространственных данных, а во временных рядах. Для временных рядов вполне естественна подобная связь будущих случайных ошибок с предыдущими их значениями.

Решение:

(а) Используя условие о постоянстве дисперсии случайной ошибки, то есть условие $\text{var}(\varepsilon_i) = \text{var}(\varepsilon_{i-1})$, найдем эту дисперсию:

$$\text{var}(\varepsilon_i) = \text{var}(\rho * \varepsilon_{i-1} + u_i)$$

$$\text{var}(\varepsilon_i) = \rho^2 \text{var}(\varepsilon_{i-1}) + \text{var}(u_i)$$

$$\text{var}(\varepsilon_i) = \rho^2 \text{var}(\varepsilon_i) + \sigma_u^2$$

$$\text{var}(\varepsilon_i) = \frac{\sigma_u^2}{1-\rho^2}.$$

Тем самым мы нашли элементы, которые будут стоять на главной диагонали ковариационной матрицы вектора случайных ошибок. Теперь найдем элементы, которые будут находиться непосредственно на соседних с главной диагональю клетках:

$$\text{cov}(\varepsilon_i, \varepsilon_{i-1}) = \text{cov}(\rho * \varepsilon_{i-1} + u_i, \varepsilon_{i-1}) = \rho * \text{cov}(\varepsilon_{i-1}, \varepsilon_{i-1}) + \text{cov}(u_i, \varepsilon_{i-1}) = \rho * \text{var}(\varepsilon_{i-1}) + 0 = \frac{\rho * \sigma_u^2}{1-\rho^2}$$

По аналогии легко убедиться, что

$$\text{cov}(\varepsilon_i, \varepsilon_{i-k}) = \frac{\rho^k * \sigma_u^2}{1-\rho^2} = \frac{\sigma_u^2 * \rho^k}{1-\rho^2}.$$

Следовательно, ковариационная матрица вектора случайных ошибок имеет вид:

$$\Omega = \frac{\sigma_u^2}{1-\rho^2} \begin{pmatrix} 1 & \rho & \dots & \rho^{n-2} & \rho^{n-1} \\ \rho & 1 & \dots & \rho^{n-3} & \rho^{n-2} \\ \dots & \dots & \dots & \dots & \dots \\ \rho^{n-2} & \rho^{n-3} & \dots & 1 & \rho \\ \rho^{n-1} & \rho^{n-2} & \dots & \rho & 1 \end{pmatrix}$$

(б) Вектор ОМНК-оценок коэффициентов имеет вид:

$$\hat{\beta} = (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} y = \left(X' \begin{pmatrix} 1 & \rho & \dots & \rho^{n-2} & \rho^{n-1} \\ \rho & 1 & \dots & \rho^{n-3} & \rho^{n-2} \\ \dots & \dots & \dots & \dots & \dots \\ \rho^{n-2} & \rho^{n-3} & \dots & 1 & \rho \\ \rho^{n-1} & \rho^{n-2} & \dots & \rho & 1 \end{pmatrix}^{-1} X \right)^{-1} * X' \begin{pmatrix} 1 & \rho & \dots & \rho^{n-2} & \rho^{n-1} \\ \rho & 1 & \dots & \rho^{n-3} & \rho^{n-2} \\ \dots & \dots & \dots & \dots & \dots \\ \rho^{n-2} & \rho^{n-3} & \dots & 1 & \rho \\ \rho^{n-1} & \rho^{n-2} & \dots & \rho & 1 \end{pmatrix}^{-1} y$$

Обратите внимание, что дробь $\frac{\sigma_u^2}{1-\rho^2}$ при расчете представленной оценки сокращается. Поэтому для вычисления оценки знать величину σ_u^2 не нужно.

Примечание: если коэффициент автокорреляции ρ неизвестен, то его можно легко оценить. Например, для этого можно применить обычный МНК к исходной регрессии, получить вектор остатков и оценить регрессию $\hat{\varepsilon}_i = \hat{\rho} * \varepsilon_{i-1}$. Полученной оценки $\hat{\rho}$ достаточно, чтобы вычислить ОМНК-оценку вектора параметров модели. Тем самым в представленном примере для применения доступного ОМНК достаточно оценить всего один параметр ковариационной матрицы вектора оценок коэффициентов.

Пример 5.5. Гетероскедастичность и ОМНК-оценка

Рассмотрим линейную модель $y = X\beta + \varepsilon$, для которой выполнены все предпосылки классической линейной модели множественной регрессии за одним исключением: дисперсия случайной ошибки прямо пропорциональна квадрату некоторой известной переменной

$$\text{var}(\varepsilon_i) = \sigma_i^2 = \sigma_0^2 z_i^2 > 0.$$

(а) Найдите ковариационную матрицу вектора случайных ошибок для представленной модели.

(б) Запишите в явном виде формулу ОМНК-оценки вектора коэффициентов модели.

Решение:

(а) Так как в этом случае нарушена только четвертая предпосылка классической линейной модели множественной регрессии, то вне главной диагонали ковариационной матрицы вектора случайных ошибок будут стоять нули.

$$\Omega = \begin{pmatrix} \sigma_0^2 z_1^2 & 0 & \dots & 0 \\ 0 & \sigma_0^2 z_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_0^2 z_n^2 \end{pmatrix}$$

(б) Обратите внимание, что при подстановке в общую формулу для ОМНК-оценки величина σ_0^2 сокращается, следовательно, для оценки вектора коэффициентов знать её не нужно:

$$\hat{\beta} = (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} y = \left(X' \begin{pmatrix} z_1^2 & 0 & \dots & 0 \\ 0 & z_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & z_n^2 \end{pmatrix}^{-1} X \right)^{-1} X' \begin{pmatrix} z_1^2 & 0 & \dots & 0 \\ 0 & z_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & z_n^2 \end{pmatrix}^{-1} y$$

* * *

Ещё одна важная ситуация, когда с успехом может быть применен доступный ОМНК — это модель со случайными эффектами, которую мы рассмотрим в главе, посвященной панельным данным.

Глава 6. Модель со стохастическими регрессорами и асимптотический подход в эконометрике

До этой главы мы предполагали, что объясняющие переменные в нашем уравнении являются детерминированными, а также опирались на свойства регрессоров для конечных выборок. Указанный подход удобен в рамках первого знакомства с эконометрическими моделями, однако имеет ряд ограничений. В современных прикладных исследованиях чаще используется **асимптотический подход**, который заключается в том, чтобы концентрироваться на асимптотических свойствах исследуемых объектов, а не на их свойствах для конечных выборок.

Напомним, что **асимптотические свойства** (оценок параметров, тестовых статистик и доверительных интервалов) — это свойства при увеличении выборки до бесконечности. Формально это свойства, которые мы наблюдаем в пределе, при $n \rightarrow \infty$.

В реальном мире размер выборки, разумеется, всегда конечен. Однако все-таки есть серьезные причины для популярности применения асимптотического подхода. Далее перечислены три из них (в порядке убывания важности):

1. **Реалистичность предпосылок.** Как мы увидим в этой главе, асимптотический подход позволит ослабить некоторые предпосылки наших эконометрических моделей, что сделает их максимально близкими к реальности.
2. **Техническая простота.** Асимптотические свойства эконометрических объектов часто оказываются проще, чем свойства для конечных выборок.
3. **Увеличение доступности данных.** В современном мире данные становятся всё более доступными, и эконометристам чаще удается работать с действительно большими их массивами, которые позволяют быть уверенными, что применение асимптотических свойств вполне уместно¹.

Всё это делает данную главу одной из ключевых с точки зрения понимания современного взгляда на эконометрические исследования.

Для освоения нового подхода нам потребуется вспомнить ряд важных определений и результатов из математической статистики. Этому посвящен первый параграф главы. Во втором параграфе мы сформулируем предпосылки новой модели со стохастическими регрессорами и обсудим, что делает их более реалистичными, чем предпосылки КЛММР.

Третий и четвертый параграфы содержат несколько доказательств свойств МНК-оценок в рамках применения нового подхода. Это «технические» параграфы, и если вы не заинтересованы в формальных доказательствах, используемых в главе фактов, их можно пропустить.

Пятый параграф, напротив, важен прежде всего с прикладной точки зрения: в нём обсуждаются особенности тестирования гипотез и построения доверительных интервалов в рамках асимптотического подхода.

[6.1. Некоторые важные результаты математической статистики](#)

[6.2. Линейная регрессионная модель со стохастическими регрессорами](#)

[6.3. Состоятельность МНК-оценок](#)

[6.4. Асимптотическая нормальность МНК-оценок](#)

[6.5. Тестирование гипотез и построение доверительных интервалов](#)

[Задания для самостоятельного решения](#)

[Приложение 6.A. Состоятельность в условиях гетероскедастичности стандартная ошибка оценки коэффициента: доказательство состоятельности](#)

[Приложение 6.B. Дельта-метод](#)

[Приложение 6.B. Таблицы стандартного нормального распределения и распределения Хи-квадрат](#)

1. Для обозначения особенно огромных массивов часто используется набравший популярность термин большие данные (big data). [↗](#)

6.1. Некоторые важные результаты математической статистики

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

Этот раздел, разумеется, не претендует на то, чтобы заменить собой учебник по теории вероятностей и математической статистике, и мы не будем вспоминать здесь все используемые в эконометрике факты из этих наук. Однако последующие разделы учебника будут восприниматься легче, если перед непосредственным знакомством с ними освежить в памяти некоторые ключевые результаты, касающиеся асимптотической теории.

Сходимость по вероятности. Рассмотрим последовательность случайных величин $X_1, X_2, \dots, X_n, \dots$

Если для любого $\varepsilon > 0$ вероятность события $|X_n - a| > \varepsilon$ стремится к нулю при $n \rightarrow \infty$, то говорят, что число a — это предел по вероятности для последовательности $X_1, X_2, \dots, X_n, \dots$

Для предела по вероятности обычно используют обозначение $X_n \xrightarrow{P} a$ или $\text{plim } X_n = a$. Также в этом случае говорят, что последовательность сходится по вероятности к числу a .

Состоятельность. Оценка параметра называется состоятельной, если её предел по вероятности равен истинному значению оцениваемого параметра: $\hat{\beta} \xrightarrow{P} \beta$.

Говоря проще, если оценка параметра состоятельна, то всё, что вам нужно, чтобы узнать его истинное значение, — собрать достаточно большую выборку. Поэтому эконометристы очень любят состоятельные оценки в сочетании с большими массивами данных.

Достаточное условие состоятельности. Если оценка параметра является несмещенной (или асимптотически несмещенной), и ее дисперсия стремится к нулю при $n \rightarrow \infty$, то эта оценка состоятельна.

Пример 6.1. Регрессия на константу

Рассмотрим модель регрессии на константу:

$$y_i = \theta + \varepsilon_i, i = 1, 2, \dots, n.$$

Пусть для неё выполнены все предпосылки классической линейной модели множественной регрессии. Докажите, что МНК-оценка параметра θ является состоятельной.

Решение:

МНК-оценка параметра θ может быть вычислена по формуле: $\hat{\theta} = \bar{y}$ (см. задание 9 из главы 2).

Она является несмещенной:

$$E(\hat{\theta}) = E\left(\frac{\sum_{i=1}^n y_i}{n}\right) = E\left(\frac{\sum_{i=1}^n (\theta + \varepsilon_i)}{n}\right) = \frac{n\theta + \sum_{i=1}^n E(\varepsilon_i)}{n} = \theta.$$

Её дисперсия равна:

$$\begin{aligned} \text{var}(\hat{\theta}) &= \text{var}\left(\frac{\sum_{i=1}^n (\theta + \varepsilon_i)}{n}\right) = \\ &= \text{var}\left(\theta + \frac{\sum_{i=1}^n \varepsilon_i}{n}\right) = \frac{\text{var}(\sum_{i=1}^n \varepsilon_i)}{n^2} = \frac{n * \sigma^2}{n^2} = \frac{\sigma^2}{n}. \end{aligned}$$

Таким образом, МНК-оценка коэффициента является несмещенной, а при $n \rightarrow \infty$ ее дисперсия стремится к нулю. Поэтому выполнено достаточное условие состоятельности.

Закон больших чисел в форме Чебышёва. Если $Y_1, Y_2, \dots, Y_n, \dots$ — независимые и одинаково распределенные случайные величины, причем $E(Y_i) = \mu$, $\text{var}(Y_i) < \infty$, то $\bar{Y} \xrightarrow{P} \mu$.

Иными словами, для последовательности независимых и одинаково распределенных величин с **конечной** дисперсией среднее значение будет состоятельной оценкой математического ожидания.

Неравенство Коши — Буняковского. Это неравенство является достаточно общим результатом, однако нам в рамках этой главы будет достаточно его частного случая для математических ожиданий:

Пусть ξ и η — случайные величины, для которых определены конечные вторые моменты распределения. Тогда

$$E|\xi * \eta| \leq \sqrt{E(\xi^2) * E(\eta^2)}$$

Замечание. Не удивляйтесь тому, что в англоязычной эконометрической литературе вы такого названия неравенства не встретите. Там этот результат принято называть неравенством Коши — Шварца (the Cauchy-Schwarz inequality).

Сходимость по распределению. Последовательность случайных величин $X_1, X_2, \dots, X_n, \dots$ сходится по распределению к случайной величине ξ , если

$$\lim_{n \rightarrow \infty} P\{X_n < x\} = P\{\xi < x\}$$

для всех точек, где функция $F(x) = P\{\xi < x\}$ непрерывна.

Обозначение сходимости по распределению: $X_n \xrightarrow{d} \xi$.

Обратите внимание на важное отличие сходимости по вероятности от сходимости по распределению. В первом случае речь идёт о том, что последовательность сходится к некоторой (неслучайной) константе, а во втором — к случайной величине.

Сходимость по распределению главным образом понадобится нам для использования центральной предельной теоремы.

Центральная предельная теорема (ЦПТ). Если $Y_1, \dots, Y_n, \dots$ — независимые и одинаково распределённые случайные величины, причём $E(Y_i) = \mu$, $var(Y_i) = \sigma^2 < \infty$, то

$$\frac{\sqrt{n}(\bar{Y} - \mu)}{\sigma} \xrightarrow{d} N(0, 1) \quad (6.1).$$

Выражение (6.1) иногда записывают в одном из эквивалентных вариантов.

Во-первых, напомним (см. пример 6.1), что $var(\bar{Y}) = \frac{\sigma^2}{n}$. Тогда, обозначив $var(\bar{Y}) = \sigma_{\bar{Y}}^2$, можно переписать (6.1) вот так:

$$\frac{(\bar{Y} - \mu)}{\sigma_{\bar{Y}}} \xrightarrow{d} N(0, 1).$$

В этом случае говорят, что распределение $\bar{Y}$ является **асимптотически нормальным** с математическим ожиданием μ и дисперсией $\frac{\sigma^2}{n}$.

Во-вторых, если случайную величину $\frac{\sqrt{n}(\bar{Y} - \mu)}{\sigma}$ домножить на коэффициент σ , то её дисперсия увеличится в σ^2 раз (по свойству дисперсии). Следовательно, выражение (6.1) эквивалентно следующему утверждению:

$$\sqrt{n}(\bar{Y} - \mu) \xrightarrow{d} N(0, \sigma^2).$$

Центральная предельная теорема даёт уверенность в том, что при указанных предпосылках и достаточно большой выборке среднее значение будет иметь приблизительно нормальное распределение. Причем для этого не требуется, чтобы отдельные случайные величины $Y_1, \dots, Y_n, \dots$ сами имели нормальное распределение. Они могут быть распределены как угодно, лишь бы имели конечную дисперсию.

Для эконометристов этот результат важен, так как во многих ситуациях гарантирует асимптотическую нормальность оценок коэффициентов, что позволяет легко тестировать гипотезы и вообще упрощает работу, так как свойства нормального распределения изучены очень хорошо.

Теорема Слуцкого (теорема Манна — Вальда).

Если

1. $X_n \xrightarrow{P} a$ (то есть последовательность случайных величин $X_1, X_2, \dots, X_n, \dots$ сходится по вероятности к константе a),
2. функция $g(x)$ непрерывна в точке a и некоторой её окрестности,

то $g(X_n) \xrightarrow{P} g(a)$.

Пример 6.2. Регрессия на константу (продолжение)

Вернемся к нашему примеру 6.1. Вычислите предел по вероятности для $\hat{\theta}^2$.

Решение:

Так как $g(x) = x^2$ — это непрерывная функция, и $\hat{\theta} = \bar{y}$ сходится по вероятности к θ , то по только что сформулированной теореме получаем, что $\hat{\theta}^2$ сходится по вероятности к θ^2 .

Замечание 1. Эта теорема верна и в случае, когда X_n — это случайный вектор, и a — это вектор констант.

Замечание 2. Эта теорема называется теоремой Слуцкого только в русскоязычной традиции. Если обратиться к англоязычным учебникам по статистике и эконометрике, то там она называется теоремой Манна — Вальда (**Mann-Wald theorem**) или даже теоремой о непрерывном отображении (**continuous mapping theorem**). А теоремой Слуцкого там называется вот такой результат:

Теорема Слуцкого. Если

1. $X_n \xrightarrow{P} a$ (то есть последовательность случайных величин $X_1, X_2, \dots, X_n, \dots$ сходится по вероятности к константе a),
2. $Y_n \xrightarrow{d} \xi$ (то есть последовательность случайных величин $Y_1, Y_2, \dots, Y_n, \dots$ сходится по распределению к случайной величине ξ),

то

$$X_n + Y_n \xrightarrow{d} a + \xi,$$

$$X_n * Y_n \xrightarrow{d} a * \xi,$$

-- .

$$\frac{Y_n}{X_n} \xrightarrow{d} \frac{\xi}{a},$$

Для последнего соотношения также требуется, чтобы случайная величина X_n не равнялась нулю с единичной вероятностью.

Обратите внимание: чтобы указанные соотношения выполнялись, каждый из пределов в отдельности должен существовать.

Обе сформулированные теоремы пригодятся нам в дальнейшем. Для определенности в контексте доказательств в данной главе договоримся, что, когда мы ссылаемся на теорему Slutsky, мы подразумеваем последний из двух вариантов.

6.2. Линейная регрессионная модель со стохастическими регрессорами

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

Сформулируем новый набор предпосылок, который будем называть предпосылками линейной модели со стохастическими регрессорами. Начнем с модели парной регрессии.

Предпосылки линейной модели со стохастическим регрессором (случай парной регрессии):

1. Модель линейна по параметрам и правильно специфицирована:

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i, \quad i = 1, 2, \dots, n.$$

2. Наблюдения $\{(x_i, y_i), i = 1, \dots, n\}$ независимы и одинаково распределены.

3. x_i и y_i имеют ненулевые конечные четвертые моменты распределения $E(x_i^4) < \infty, E(y_i^4) < \infty$.

4. Случайные ошибки имеют нулевое условное математическое ожидание при заданном x_i : $E(\varepsilon_i | x_i) = 0$.

Сравним предпосылки этой модели с предпосылками классической линейной модели парной регрессии (КЛМНР) из главы 2.

Первая предпосылка стандартна и остается без изменений.

Вторая предпосылка в КЛМНР требовала, чтобы регрессоры были неслучайными величинами. Теперь мы отказываемся от неё, допуская, что объясняющие переменные могут быть случайными. При этом мы требуем, чтобы наблюдения $\{(x_i, y_i), i = 1, \dots, n\}$ были независимыми и одинаково распределенными (independent and identically distributed, i.i.d.).

Это требование вовсе **не означает**, что y_i не зависит от x_i (ясно, что в этом случае анализировать модель их взаимосвязи было бы бессмысленно). Зато оно говорит о том, что векторы $(x_1, y_1), (x_2, y_2), (x_3, y_3) \dots$ независимы друг от друга в вероятностном смысле. Иными словами, отдельные наблюдения в нашей модели не влияют друг на друга.

Для пространственных данных эта предпосылка практически всегда выполняется¹. В то же время следует помнить, что при работе с временными рядами эта предпосылка часто нарушается, так как для временных рядов естественно предполагать, что будущие значения переменных зависят от прошлых². Поскольку пока мы в основном концентрируемся на пространственных данных, для нас она остается весьма реалистичной.

Лирическое отступление о неслучайных и случайных регрессорах

Отвлечемся ненадолго от технических деталей и обратимся к вопросу: как следует думать об объясняющих переменных с содержательной точки зрения? Следует ли считать их скорее детерминированными величинами или скорее случайными?

Ответ, разумеется, зависит от того, с какими данными вы работаете, и какова процедура их сбора.

Представим, например, что вы анализируете зависимость логарифма реального ВВП от номера года. То есть оцениваете параметры линии тренда для временного ряда:

$$\ln y_t = \beta_1 + \beta_2 \cdot t + \varepsilon_t$$

Здесь y_t — ВВП в год t . В данном примере регрессор (номер года t) вполне естественно считать неслучайным (детерминированным). Действительно, мы точно знаем, что в принятой нами системе летоисчисления за 2020-ым годом последует 2021-ый, а затем наступит 2022-ой. Никакой случайности тут нет.

Теперь представим, что вас интересуют параметры следующей модели для инфляции:

$$\pi_t = \beta_1 + \beta_2 \pi_{t-1} + \beta_3 x_t + \varepsilon_t$$

π_t — это уровень инфляции в год t , а x_t — это, например, отклонение фактического ВВП от потенциального ВВП в год t ³. Обратите внимание: здесь предполагается, что инфляция в текущем периоде зависит от инфляции в прошлом периоде. Однако инфляция прошлого периода π_{t-1} , в свою очередь, зависит от ε_{t-1} , а значит уж точно является случайной величиной. Следовательно, в данном примере по крайней мере один из регрессоров (переменная π_{t-1}) заведомо является случайным (стохастическим).

В двух приведенных примерах детерминированная или стохастическая природа объясняющих переменных может быть определена однозначно из соображений здравого смысла. В то же время, во многих ситуациях решение о том, как воспринимать регрессоры — как неслучайные величины или как случайные — это исключительно вопрос технического удобства. В частности, при использовании асимптотического подхода второй вариант более удобен, поэтому в современных эконометрических приложениях по умолчанию используют его.

Третья предпосылка выглядит достаточно устрашающе. Однако в действительности никак не ограничивает исследователя. По существу, она означает, что очень большие выбросы в данных маловероятны. Это техническая предпосылка, которая, как мы увидим в дальнейшем, позволяет гарантировать асимптотическую нормальность оценок коэффициентов. Это даст нам возможность тестировать гипотезы и строить доверительные интервалы.

Проверить эту предпосылку сложно, однако она достаточно слабая, и потому на практике обычно считают, что она выполнена. Во всяком случае, легко согласиться с тем, что она выполняется гораздо чаще, чем предпосылка КЛМНР №6 о нормальности случайных ошибок. А ведь именно её она, в сущности, заменяет.

Четвертая предпосылка играет ключевую роль в получении корректных результатов эконометрического моделирования. В последующих параграфах и главах мы увидим, что именно вопрос о выполнении или нарушении этой предпосылки оказывается в центре дискуссии об уместности применения тех или иных методов и спецификаций моделей в различных ситуациях.

Содержательно эта предпосылка говорит о том, что «прочие факторы», которые «спрятаны» в случайной ошибке ε_i , никак не связаны с регрессором. Поэтому знание x_i никак не влияет на ожидания по поводу случайной величины ε_i .

Чтобы на конкретных числах «пощупать» эту предпосылку, а заодно вспомнить, что такое условное математическое ожидание и как его считать, рассмотрим следующий простой пример.

Пример 6.3. Об условном математическом ожидании

Пусть известен совместный закон распределения случайных величин x_i и ε_i .

	$\varepsilon_i = -1$	$\varepsilon_i = 0$	$\varepsilon_i = 1$
$x_i = 0$	0,2	0,1	0,2
$x_i = 1$	0,1	0,3	0,1

(а) Проверьте, выполняется ли в данном случае предпосылка №4 об условном математическом ожидании случайной ошибки?

(б) Вычислите безусловное математическое ожидание случайной ошибки.

(в) Вычислите $\text{cov}(\varepsilon_i, x_i)$.

Решение:

(а) Напомним, что по определению условным математическим ожиданием случайной величины ε_i при условии x_i называется математическое ожидание условного распределения случайной величины ε_i при условии x_i .

Запишем закон условного распределения ε_i при условии, что $x_i = 0$. Для этого отметим, что вероятность события $x_i = 0$ в нашем примере составляет $0,2 + 0,1 + 0,2 = 0,5$.

	$\varepsilon_i = -1$	$\varepsilon_i = 0$	$\varepsilon_i = 1$
$P(\varepsilon_i x_i = 0)$	$\frac{0,2}{0,5}$	$\frac{0,1}{0,5}$	$\frac{0,2}{0,5}$

Зная этот закон распределения, легко посчитать математическое ожидание:

$$E(\varepsilon_i | x_i = 0) = -1 * \frac{0,2}{0,5} + 0 * \frac{0,1}{0,5} + 1 * \frac{0,2}{0,5} = 0$$

Аналогично получаем условное математическое ожидание ε_i при условии, что $x_i = 1$.

	$\varepsilon_i = -1$	$\varepsilon_i = 0$	$\varepsilon_i = 1$
$P(\varepsilon_i x_i = 1)$	$\frac{0,1}{0,5}$	$\frac{0,3}{0,5}$	$\frac{0,1}{0,5}$

$$E(\varepsilon_i | x_i = 1) = -1 * \frac{0,1}{0,5} + 0 * \frac{0,3}{0,5} + 1 * \frac{0,1}{0,5} = 0$$

Таким образом, для любого возможного значения x_i условие $E(\varepsilon_i | x_i) = 0$ соблюдается. То есть предпосылка выполнена.

$$\begin{aligned} \text{(б)} E(\varepsilon_i) &= P(\varepsilon_i = -1) * (-1) + P(\varepsilon_i = 0) * 0 + P(\varepsilon_i = 1) * 1 = \\ &= 0,3 * (-1) + 0,4 * 0 + 0,3 * 1 = 0 \end{aligned}$$

Следовательно, безусловное математическое ожидание случайной ошибки тоже равно нулю.

$$\text{(в)} \text{cov}(\varepsilon_i, x_i) = E(\varepsilon_i x_i) - E(\varepsilon_i) * E(x_i) = E(\varepsilon_i x_i) - 0 * E(x_i) = E(\varepsilon_i x_i)$$

$$\begin{aligned} E(\varepsilon_i x_i) &= 0,2 * (-1) * 0 + 0,1 * 0 * 0 + 0,2 * 1 * 0 + \\ &+ 0,1 * (-1) * 1 + 0,3 * 0 * 1 + 0,1 * 1 * 1 = 0 \end{aligned}$$

В нашем примере оказалось, что предпосылке №4 соответствует выполнение условий $E(\varepsilon_i) = 0$ и $\text{cov}(\varepsilon_i, x_i) = 0$. На самом деле это не случайный результат. Его можно обобщить, доказав два важных следствия из предпосылки №4.

Следствие 1. Если случайные ошибки имеют нулевое **условное** математическое ожидание при заданном x_i : $E(\varepsilon_i | x_i) = 0$, то они имеют нулевое **безусловное** математическое ожидание: $E(\varepsilon_i) = 0$

Доказательство этого следствия является хорошим примером применения **закона повторного математического ожидания**.

Напомним формулировку закона повторного математического ожидания:

$$E(\xi) = E(E(\xi | \eta))$$

В нашем случае в соответствии с этим законом:

$$E(\varepsilon_i) = E(E(\varepsilon_i | x_i)) = E(0) = 0.$$

Поэтому, сформулировав предпосылку №4, мы не нуждаемся в том, чтобы отдельно формулировать предположение по поводу безусловного математического ожидания случайной ошибки, которое мы делаем в КЛМНР.

Подчеркнем, что обратное утверждение, вообще говоря, неверно. Вполне возможна ситуация, когда безусловное математическое ожидание случайной ошибки равно нулю, а её условное математическое ожидание при условии x_i — нет. см. пример 6.4 далее.

Следствие 2. Если случайные ошибки имеют нулевое условное математическое ожидание при любом заданном x_i : $E(\varepsilon_i | x_i) = 0$, то регрессор и случайная ошибка не коррелированы друг с другом: $\text{cov}(\varepsilon_i, x_i) = 0$.

Для доказательства сначала отметим, что по свойству теоретической ковариации:

$$\text{cov}(\varepsilon_i, x_i) = E(\varepsilon_i x_i) - E(\varepsilon_i)E(x_i) = E(\varepsilon_i x_i) - 0 * E(x_i) = E(\varepsilon_i x_i).$$

А затем снова воспользуемся законом повторного математического ожидания:

$$E(\varepsilon_i x_i) = E(E(\varepsilon_i x_i | x_i)) = E(x_i E(\varepsilon_i | x_i)) = E(x_i * 0) = E(0) = 0$$

Регрессор, который не коррелирован со случайной ошибкой модели, обычно называют **экзогенным регрессором**. Таким образом, предпосылку №4 иногда называют предпосылкой об экзогенности регрессора.

Если же объясняющая переменная в модели, наоборот, коррелирована со случайной ошибкой $\text{cov}(\varepsilon_i, x_i) \neq 0$, то её называют **эндогенным регрессором**.

Пример 6.4. Об условном математическом ожидании (продолжение)

Пусть теперь совместный закон распределения x_i и ε_i имеет такой вид:

	$\varepsilon_i = -1$	$\varepsilon_i = 0$	$\varepsilon_i = 1$
$x_i = 0$	0,3	0,1	0,1
$x_i = 1$	0,1	0,1	0,3

Покажите, что в этом случае условие $E(\varepsilon_i) = 0$ выполнено, а условие $E(\varepsilon_i|x_i) = 0$ нарушается.

Решение:

$$E(\varepsilon_i) = P(\varepsilon_i = -1) * (-1) + P(\varepsilon_i = 0) * 0 + P(\varepsilon_i = 1) * (1) = \\ = 0,4 * (-1) + 0,2 * 0 + 0,4 * 1 = 0$$

Чтобы показать, что предпосылка $E(\varepsilon_i|x_i) = 0$ не выполняется, достаточно привести любое значение x_i , для которого указанное равенство нарушено. Рассмотрим, например, случай $x_i = 0$.

$$E(\varepsilon_i|x_i = 0) = -1 * \frac{0,3}{0,5} + 0 * \frac{0,1}{0,5} + 1 * \frac{0,1}{0,5} = -0,4$$

Следовательно, предпосылка $E(\varepsilon_i|x_i) = 0$ не выполняется: регрессор в модели является эндогенным.

Выполнение четырех предпосылок линейной модели со стохастическими регрессорами (случай парной регрессии) гарантирует, что применение МНК будет приводить к хорошим результатам. Говоря более строго, эти гарантии можно сформулировать в виде следующей теоремы:

Теорема о состоятельности и асимптотической нормальности МНК-оценок в парной регрессии. Если предпосылки №1–4 выполнены, то МНК-оценки коэффициентов β_1 и β_2 состоятельны и асимптотически нормальны.

Доказательство этой теоремы приводится в параграфах 6.3 и 6.4. В первом из них доказывается состоятельность, а во втором — асимптотическая нормальность. Однако прежде, чем переходить к доказательству, обсудим значение теоремы для прикладных исследований. Забегая вперед, отметим, что оно велико.

Первый из результатов — состоятельность — даёт нам уверенность, что при достаточно слабых предположениях МНК будет обеспечивать верные ответы на интересующие нас вопросы о мире. Для получения этих ответов нужно лишь собрать достаточно много данных, чтобы асимптотические свойства были применимы. В практических исследованиях вполне хватает нескольких сотен точек (хотя, конечно, когда речь идет об асимптотических методах, то чем больше, тем лучше).

Второй результат — асимптотическая нормальность — позволяет нам легко тестировать гипотезы и строить доверительные интервалы, не делая жестких предположений о распределении отдельных случайных ошибок и отдельных переменных. Детали см. в параграфе 6.5. Это ценно потому, что на практике обычно нет никакой уверенности в том, что случайные ошибки модели распределены нормально. А ведь в рамках КЛМНР, как вы помните, мы были вынуждены делать такую предпосылку.

Отметим также, что в рамках нашей новой модели, в отличие от КЛМНР, мы не требуем гомоскедастичности. Действительно, мы сделали предположение по поводу того, что константой должно быть **условное математическое ожидание** случайной ошибки $E(\varepsilon_i|x_i)$, однако по поводу **условной дисперсии** случайной ошибки $\text{var}(\varepsilon_i|x_i)$ мы никаких предпосылок не делали. Следовательно, эта величина может меняться при изменении x_i , то есть в модели может наблюдаться гетероскедастичность (в таком случае её также называют условной гетероскедастичностью).

Аналогичный набор предпосылок и аналогичная теорема могут быть, разумеется, сформулированы и для множественной регрессии:

Предпосылки линейной модели со стохастическими регрессорами (случай множественной регрессии):

1. Модель линейна по параметрам:

$$y_i = \beta_1 + \beta_2 * x_i^{(2)} + \beta_3 * x_i^{(3)} + \dots + \beta_k * x_i^{(k)} + \varepsilon_i, i = 1, 2, \dots, n.$$

2. Наблюдения $\{(x_i^{(2)}, \dots, x_i^{(k)}, y_i), i = 1, \dots, n\}$ независимы и одинаково распределены.

3. $x_i^{(2)}, \dots, x_i^{(k)}, y_i$ имеют ненулевые конечные четвертые моменты.

4. Случайные ошибки имеют нулевое условное математическое ожидание при заданных значениях регрессоров:

$$E(\varepsilon_i|x_i^{(2)}, \dots, x_i^{(k)}) = 0, i = 1, \dots, n$$

5. В модели с вероятностью единица отсутствует чистая мультиколлинеарность.

Теорема о состоятельности и асимптотической нормальности МНК-оценок (случай множественной регрессии). Если предпосылки №1–5 выполнены, то МНК-оценки коэффициентов модели множественной регрессии состоятельны и асимптотически нормальны.

Легко видеть, что набор предпосылок полностью идентичен случаю парной регрессии за одним исключением: нам пришлось добавить требование отсутствия мультиколлинеарности. Как мы знаем, при его нарушении МНК-оценки в модели множественной регрессии в принципе невозможно определить однозначно. Упоминание вероятности в формулировке предпосылки связано с тем, что теперь регрессоры являются стохастическими, то есть при каждой реализации их набор может отличаться.

Таблица 6.1. Сопоставление различных регрессионных моделей

Название модели	Классическая линейная модель множественной регрессии	Обобщенная линейная модель множественной регрессии	Линейная модель со стохастическими регрессорами
Где эта модель описана	В параграфе 3.2 (а также для случая парной регрессии в параграфе 2.3)	В параграфе 5.5	В параграфе 6.2
Предположение о детерминированности (неслучайности) регрессоров	Требуется	Требуется	Не требуется
Предположение о нормальности случайных ошибок	Требуется для тестирования гипотез	Требуется для тестирования гипотез	Не требуется
Предположение об отсутствии гетероскедастичности	Требуется	Не требуется	Не требуется

В таблице 6.1 содержится сопоставление предпосылок трёх основных моделей, в условиях которых мы исследуем свойства МНК-оценок. Из неё легко видеть, что предпосылки нашей новой модели, действительно, являются сравнительно более мягкими, что делает её максимально реалистичной моделью для практической работы с пространственными данными.

1. Исключение составляет специфический класс моделей пространственной автокорреляции, которые обычно рассматриваются отдельно. [↩](#)
2. Пример такой ситуации приведен далее в лирическом отступлении о неслучайных и случайных регрессорах. [↩](#)
3. Макроэкономист узнает в такой спецификации одну из возможных версий современной кривой Филлипса с адаптивными инфляционными ожиданиями. Однако даже человек, незнакомый с макроэкономическими моделями, наверняка согласится с тем, что если инфляция была высока в прошлом месяце, то и в этом она тоже наверняка будет высокой. Иными словами, текущая инфляция зависит от своих прошлых значений, что и отражено в данной модели. [↩](#)

6.3. Состоятельность МНК-оценок

В этом параграфе мы докажем, что если выполнены предпосылки линейной модели со стохастическим регрессором:

1. Модель представима следующим образом:

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i, i = 1, 2, \dots, n,$$

1. Наблюдения $\{(x_i, y_i), i = 1, \dots, n\}$ независимы и одинаково распределены,
2. x_i и y_i имеют ненулевые конечные четвертые моменты распределения $E(x_i^4) < \infty, E(y_i^4) < \infty$,
3. Случайные ошибки имеют нулевое условное математическое ожидание при заданном x_i : $E(\varepsilon_i | x_i) = 0$,

то МНК-оценка коэффициента β_2 является состоятельной.

Иными словами, мы докажем первую часть теоремы, сформулированной в параграфе 6.2.

В процессе доказательства мы несколько раз будем использовать тот факт, что в условиях перечисленных предпосылок 1–4 выборочные моменты сходятся по вероятности к своим теоретическим аналогам. Например:

$$\widehat{var}(x) \xrightarrow{p} var(x_i)$$

$$\widehat{cov}(x, y) \xrightarrow{p} cov(x_i, y_i)$$

Вполне возможно, что вы хорошо знакомы с этим фактом из курса математической статистики. Однако мы все-таки докажем одно из этих утверждений, чтобы продемонстрировать, почему сформулированные предпосылки действительно важны.

Утверждение: если предпосылки 2–3 выполнены, то $\widehat{cov}(x, y) \xrightarrow{p} cov(x_i, y_i)$

Доказательство:

$$cov(x_i, y_i) = E(x_i y_i) - E(x_i) * E(y_i)$$

$$\widehat{cov}(x, y) = \overline{xy} - \bar{x} * \bar{y}$$

Так как в силу предпосылки №2 все x_i — независимые и одинаково распределенные, выполняется закон больших чисел:

$$\bar{x} \xrightarrow{p} E(x_i)$$

Аналогично: $\bar{y} \xrightarrow{p} E(y_i)$

Следовательно, по теореме Slutsky $\bar{x} * \bar{y} \xrightarrow{p} E(x_i) * E(y_i)$

По закону больших чисел: $\overline{xy} \xrightarrow{p} E(x_i y_i)$

Наконец, применив теорему Slutsky для разности, получим:

$$\overline{xy} - \bar{x} * \bar{y} \xrightarrow{p} E(x_i y_i) - E(x_i) * E(y_i) = cov(x_i, y_i)$$

Зачем для доказательства нам необходима предпосылка №3 о том, что $E(x_i^4) < \infty, E(y_i^4) < \infty$?

Чтобы применить закон больших чисел (в форме Чебышёва) к последовательности $\{x_i, i = 1, 2, \dots\}$, нам необходимо, чтобы у x_i существовали конечная дисперсия и математическое ожидание. Из конечности момента распределения четвертого порядка следует и конечность моментов распределения первого и второго порядков, следовательно, математическое ожидание и дисперсия x_i существуют. Аналогично для y_i .

Наконец, чтобы применить закон больших чисел к последовательности произведений $\{x_i y_i, i = 1, 2, \dots\}$, нам необходимо, чтобы у $x_i y_i$ существовали конечная дисперсия и математическое ожидание.

$$E(x_i^2 * y_i^2) \leq \sqrt{E(x_i^4) * E(y_i^4)} < \infty$$

Здесь первый переход следует из неравенства Коши-Буняковского, а второй — из предпосылки №3.

Это значит, что случайная величина $x_i y_i$ имеет конечный второй момент. Тогда она имеет конечный первый момент $E x_i y_i$ и конечную дисперсию

$$var(x_i y_i) = E(x_i^2 * y_i^2) - (E x_i y_i)^2$$

Что и требовалось доказать.

Утверждение $\widehat{var}(x) \xrightarrow{p} var(x_i)$ доказывается аналогичным образом.

Теперь мы можем непосредственно доказать состоятельность МНК-оценки коэффициента при переменной в модели парной регрессии.

$$\widehat{\beta}_2 = \frac{\widehat{cov}(x, y)}{\widehat{var}(x)} \xrightarrow{p} \frac{cov(x_i, y_i)}{var(x_i)} = \frac{cov(x_i, \beta_1 + \beta_2 * x_i + \varepsilon_i)}{var(x_i)} =$$

$$= \frac{\text{cov}(x_i, \beta_2 * x_i + \varepsilon_i)}{\text{var}(x_i)} = \frac{\beta_2 * \text{cov}(x_i, x_i) + \text{cov}(x_i, \varepsilon_i)}{\text{var}(x_i)} = \beta_2 + \frac{\text{cov}(x_i, \varepsilon_i)}{\text{var}(x_i)}$$

Таким образом, мы получили важное соотношение:

$$\widehat{\beta}_2 \xrightarrow{p} \beta_2 + \frac{\text{cov}(x_i, \varepsilon_i)}{\text{var}(x_i)} \quad (6.1)$$

В параграфе 6.2 мы доказали, что из предпосылки $E(\varepsilon_i | x_i) = 0$ следует равенство нулю соответствующей ковариации: $\text{cov}(x_i, \varepsilon_i) = 0$. Поэтому в случае, когда данная предпосылка об экзогенности регрессора выполнена, получаем:

$$\widehat{\beta}_2 \xrightarrow{p} \beta_2 + \frac{\text{cov}(x_i, \varepsilon_i)}{\text{var}(x_i)} = \beta_2 + \frac{0}{\text{var}(x_i)} = \beta_2$$

Тем самым мы получили, что $\widehat{\beta}_2 \xrightarrow{p} \beta_2$, то есть что при увеличении выборки МНК-оценка коэффициента сходится по вероятности к истинному значению этого коэффициента, а значит, является состоятельной. Что и требовалось доказать.

Из этого доказательства становится ясно, почему критически важно выполнение предпосылки №4 об экзогенности регрессора. Представим, что она нарушена, то есть $\text{cov}(x_i, \varepsilon_i) \neq 0$. Пусть, например, регрессор положительно коррелирован со случайной ошибкой: $\text{cov}(x_i, \varepsilon_i) > 0$. В этом случае:

$$\widehat{\beta}_2 \xrightarrow{p} \beta_2 + \frac{\text{cov}(x_i, \varepsilon_i)}{\text{var}(x_i)} > \beta_2.$$

Здесь МНК-оценка окажется несостоятельной и завышенной (и наоборот, если $\text{cov}(x_i, \varepsilon_i) < 0$, то она будет несостоятельной и заниженной).

6.4. Асимптотическая нормальность МНК-оценок

В этом параграфе мы докажем, что МНК-оценка коэффициента β_2 имеет асимптотически нормальное распределение, если выполнены предпосылки линейной модели со стохастическим регрессором:

1. Модель представима следующим образом:

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i, i = 1, 2, \dots, n,$$

1. Наблюдения $\{(x_i, y_i), i = 1, \dots, n\}$ независимы и одинаково распределены,
2. x_i и y_i имеют ненулевые конечные четвертые моменты распределения $E(x_i^4) < \infty, E(y_i^4) < \infty$,
3. Случайные ошибки имеют нулевое условное математическое ожидание при заданном x_i : $E(\varepsilon_i | x_i) = 0$.

Иными словами, мы докажем вторую часть теоремы, сформулированной в параграфе 6.2.

Для этого мы докажем, что

$$\sqrt{n}(\hat{\beta}_2 - \beta_2) \xrightarrow{d} N\left(0, \frac{\text{var}((x_i - \mu_x)\varepsilon_i)}{(\text{var}(x_i))^2}\right).$$

Здесь μ_x обозначает математическое ожидание регрессора, то есть $E(x_i) = \mu_x$.

Как было показано во второй главе (см. равенство (2.2) в параграфе 2.4), оценка $\hat{\beta}_2$ может быть представлена следующим образом:

$$\hat{\beta}_2 = \beta_2 + \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

С учетом этого представления имеем:

$$\begin{aligned} \sqrt{n}(\hat{\beta}_2 - \beta_2) &= \sqrt{n}\left(\beta_2 + \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} - \beta_2\right) = \frac{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} = \\ &= \frac{\sqrt{\frac{1}{n} \sum_{i=1}^n ((x_i - \mu_x) - (\bar{x} - \mu_x))\varepsilon_i}}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} = \\ &= \frac{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu_x)\varepsilon_i}}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} - \frac{(\bar{x} - \mu_x)\sqrt{\frac{1}{n} \sum_{i=1}^n \varepsilon_i}}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

Для удобства дальнейших выкладок введем такие обозначения:

$$\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu_x)\varepsilon_i} = A_n,$$

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = B_n,$$

$$(\bar{x} - \mu_x)\sqrt{\frac{1}{n} \sum_{i=1}^n \varepsilon_i} = C_n.$$

С учетом новых обозначений имеем:

$$\sqrt{n}(\hat{\beta}_2 - \beta_2) = \frac{A_n}{B_n} - \frac{C_n}{B_n}$$

Рассмотрим последовательности случайных величин A_n, B_n, C_n по отдельности.

Как мы обсудили в параграфе 6.3 в рамках наших предпосылок выборочная дисперсия регрессора сходится к своему теоретическому аналогу:

$$B_n = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \widehat{\text{var}}(x) \xrightarrow{p} \text{var}(x_i)$$

Далее, чтобы выяснить, куда сходится последовательность C_n , применим закон больших чисел к первому её множителю и центральную предельную теорему ко второму. Получим:

$$(\bar{x} - \mu_x) \xrightarrow{p} (\mu_x - \mu_x) = 0$$

$$\sqrt{\frac{1}{n}} \sum_{i=1}^n \varepsilon_i \xrightarrow{d} N(0, \sigma_\varepsilon^2)$$

Применим к произведению этих множителей теорему Slutsky и увидим, что оно сходится по распределению к произведению нуля и нормальной случайной величины. То есть к нулю:

$$C_n = (\bar{x} - \mu_x) * \sqrt{\frac{1}{n}} \sum_{i=1}^n \varepsilon_i \xrightarrow{d} 0$$

Осталось исследовать последовательность A_n :

$$A_n = \sqrt{\frac{1}{n}} \sum_{i=1}^n (x_i - \mu_x) \varepsilon_i = \sqrt{\frac{1}{n}} \sum_{i=1}^n v_i$$

Здесь $v_i = (x_i - \mu_x) \varepsilon_i$

Покажем, что A_n имеет асимптотически нормальное распределение. Для этого применим центральную предельную теорему. Чтобы её применение было корректным, следует убедиться в выполнении её условий. Для этого необходимо доказать, что дисперсия v_i конечна

$$\text{var}(v_i) = E(v_i - E v_i)^2 = E(v_i)^2 = E((x_i - \mu_x) \varepsilon_i)^2 = E((x_i - \mu_x)^2 \varepsilon_i^2)$$

По неравенству Коши-Буняковского.

$$E((x_i - \mu_x)^2 \varepsilon_i^2) \leq \sqrt{E(x_i - \mu_x)^4 E \varepsilon_i^4}$$

По предпосылке №3:

$$\sqrt{E(x_i - \mu_x)^4 E \varepsilon_i^4} < \infty.$$

Строго говоря, в предпосылке №3 накладывается ограничение на моменты распределения переменных x_i и y_i , а не переменной ε_i . Однако в силу предпосылки №1 случайная величина ε_i линейно выражается через x_i и y_i . Поэтому, раз данные переменные имеют конечные четвертые моменты распределения, то и для ε_i это тоже верно.

Итак, доказано, что $\text{var}(v_i) < \infty$, поэтому к случайным величинам v_i применима центральная предельная теорема. В силу этой теоремы

$$A_n = \sqrt{\frac{1}{n}} \sum_{i=1}^n (x_i - \mu_x) \varepsilon_i = \sqrt{\frac{1}{n}} \sum_{i=1}^n v_i \xrightarrow{d} \delta$$

Где случайная величина δ имеет распределение $N(0, \text{var}(v_i))$.

Обобщая все сказанное выше и снова применяя теорему Slutsky, получаем:

$$\sqrt{n}(\hat{\beta}_2 - \beta_2) = \frac{A_n}{B_n} - \frac{C_n}{B_n} \xrightarrow{d} \frac{\delta}{\text{var}(x_i)} - 0,$$

где случайная величина δ имеет распределение $N(0, \text{var}(v_i))$. Следовательно, случайная величина $\frac{\delta}{\text{var}(x_i)}$ имеет распределение $N\left(0, \frac{\text{var}(v_i)}{(\text{var}(x_i))^2}\right)$.

Таким образом, мы доказали, что

$$\sqrt{n}(\hat{\beta}_2 - \beta_2) \xrightarrow{d} N\left(0, \frac{\text{var}(v_i)}{(\text{var}(x_i))^2}\right)$$

Иными словами, величина $\sqrt{n}(\hat{\beta}_2 - \beta_2)$ имеет асимптотически нормальное распределение с математическим ожиданием 0 и дисперсией

$$\frac{\text{var}(v_i)}{(\text{var}(x_i))^2}.$$

Следовательно, величина $(\hat{\beta}_2 - \beta_2)$ имеет асимптотически нормальное распределение с математическим ожиданием 0 и дисперсией

$$\frac{\text{var}(v_i)}{n(\text{var}(x_i))^2}.$$

(Так как по свойству дисперсии, разделив случайную величину $\sqrt{n}(\widehat{\beta}_2 - \beta_2)$ на $\sqrt{n}$, мы уменьшим её дисперсию в n раз.)

И наконец, величина $\widehat{\beta}_2$ имеет асимптотически нормальное распределение с математическим ожиданием β_2 и той же дисперсией

$$\frac{\text{var}(v_i)}{n(\text{var}(x_i))^2}.$$

(Так как добавив к случайной величине $(\widehat{\beta}_2 - \beta_2)$ константу β_2 , мы увеличим её матожидание на эту константу, оставив дисперсию неизменной.)

Осталось вспомнить, что $v_i = (x_i - \mu_x)\varepsilon_i$, и записать, что тем самым МНК-оценка коэффициента при регрессоре в модели парной регрессии $\widehat{\beta}_2$ имеет асимптотически нормальное распределение с математическим ожиданием β_2 и дисперсией

$$\text{var}(\widehat{\beta}_2) = \frac{\text{var}((x_i - \mu_x)\varepsilon_i)}{n * (\text{var}(x_i))^2}.$$

Что и требовалось доказать.

Последняя из формул даёт подсказку, как можно получить состоятельную оценку дисперсии случайной ошибки. Для этого нужно теоретические дисперсии в числителе и знаменателе указанной дроби заменить их оценками, а случайные ошибки заменить остатками регрессии. Например, вот так:

$$\widehat{\text{var}}(\widehat{\beta}_2) = \frac{\frac{1}{n-2} \sum_{i=1}^n (x_i - \bar{x})^2 e_i^2}{n * (\widehat{\text{var}}(x))^2}$$

В приложении 6.А к этой главе показано, что такая оценка является состоятельной (даже в условиях гетероскедастичности) оценкой дисперсии $\widehat{\beta}_2$.

6.5. Тестирование гипотез и построение доверительных интервалов

Во второй и третьей главах, обсуждая тестирование гипотез относительно отдельных коэффициентов, мы использовали тот факт, что отношение $\frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$ имеет t-распределение Стьюдента с $(n - k)$ степенями свободы. Однако в асимптотическом случае это число степеней свободы стремится к бесконечности. Из математической статистики известно, что при неограниченном увеличении числа степеней свободы случайная величина, имеющая распределение Стьюдента, сходится к нормальной случайной величине. Это несколько упрощает процедуру тестирования гипотезы о незначимости коэффициента, которая теперь устроена так:

Процедура тестирования незначимости коэффициента в модели множественной регрессии при использовании асимптотического подхода:

1. Формулируем тестируемую гипотезу $H_0 : \beta_j = 0$ («переменная $x^{(j)}$ не влияет на переменную y ») и альтернативную гипотезу $H_1 : \beta_j \neq 0$ («переменная $x^{(j)}$ влияет на переменную y »).
2. Находим расчетное значение тестовой статистики по формуле $\frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$.
3. Вычисляем Р-значение по формуле

$$P\text{-значение} = 2 * \Phi \left(- \left| \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \right| \right).$$

Здесь $\Phi(\bullet)$ обозначает функцию стандартного нормального распределения.

1. Выбираем уровень значимости α .
2. Если Р-значение меньше уровня значимости:

$$2 * \Phi \left(- \left| \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \right| \right) < \alpha,$$

то следует отвергнуть гипотезу $H_0 : \beta_j = 0$ и сделать вывод в пользу альтернативной гипотезы, то есть заключить, что переменная $x^{(j)}$ влияет на переменную y . В этом случае переменную $x^{(j)}$ называют статистически значимой при уровне значимости α .

Замечание 1. Как и прежде, вместо вычисления Р-значения можно сравнивать расчетное значение тестовой статистики с критическим значением из таблицы стандартного нормального распределения. Например, при уровне значимости 1% ($\alpha = 0,01$) получаем следующее условие для отвержения нулевой гипотезы:

$$2 * \Phi \left(- \left| \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \right| \right) < 0,01.$$

Это условие эквивалентно неравенству

$$\left| \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \right| > 2,58.$$

Поэтому при использовании асимптотического подхода критическое значение тестовой статистики при уровне значимости 1% составляет 2,58. Аналогично легко проверить, что при уровне значимости 5% оно равно 1,96.

Отсюда следует, что 99-процентный и 95-процентный асимптотические доверительные интервалы для коэффициента, соответственно, составляют:

$$(\hat{\beta}_j - se(\hat{\beta}_j) * 2,58; \hat{\beta}_j + se(\hat{\beta}_j) * 2,58)$$

и

$$(\hat{\beta}_j - se(\hat{\beta}_j) * 1,96; \hat{\beta}_j + se(\hat{\beta}_j) * 1,96).$$

В Приложении 6.Б обсуждается построение доверительных интервалов для нелинейных относительно коэффициентов выражений. Оказывается, что асимптотический подход позволяет решать и такую задачу. Для этого применяется так называемый дельта-метод.

Замечание 2. Аналогичным образом можно тестировать гипотезу $H_0 : \beta_j = c$, где c — это некоторая константа. В этом случае процедура тестирования остается такой же с одним исключением: расчетное значение тестовой статистики будет иметь вид

$\frac{\hat{\beta}_j - c}{se(\hat{\beta}_j)}$ и, следовательно, Р-значение рассчитывается по формуле:

$$P\text{-значение} = 2 * \Phi \left(- \left| \frac{\hat{\beta}_j - c}{se(\hat{\beta}_j)} \right| \right).$$

Замечание 3. Не забудьте, что, так как в рамках нашей новой модели мы отказались от предположения о гомоскедастичности, стандартные ошибки для осуществления этого теста должны быть получены на основе формул стандартных ошибок, состоятельных в условиях гетероскедастичности. К счастью, все современные эконометрические пакеты умеют легко рассчитывать их автоматически.

В случае тестирования гипотез по поводу выполнения сразу нескольких линейных ограничений также можно использовать все стандартные варианты F-теста, которые подробно рассмотрены в параграфе 3.5.

Единственное отличие состоит в том, что теперь если верна нулевая гипотеза данного теста, то расчетное значение тестовой статистики имеет не распределение Фишера с q и $(n - k)$ степенями свободы $F(q, n - k)$. Теперь это распределение Фишера с q и ∞ степеней свободы $F(q, \infty)$, так как $n \rightarrow \infty$. Поэтому для получения критических значений следует использовать таблицу распределения Фишера $F^\alpha(q, \infty)$ или таблицу распределения Хи-квадрат, так как случайная величина с распределением $\chi^2(q)$ в q раз больше случайной величины с распределением $F(q, \infty)$.

Ограничение такого подхода к тестированию совместных ограничений состоит в том, что стандартный F-тест опять же требует гомоскедастичности случайных ошибок. Поэтому, чтобы получить корректные в условиях гетероскедастичности результаты тестирования гипотезы, используется специальное обобщение F-теста. В качестве такого обобщения можно применять, например, тест Вальда.

Идея теста Вальда похожа на идею использования состоятельных в условиях гетероскедастичности стандартных ошибок: расчетное значение тестовой статистики корректируется с учетом возможной гетероскедастичности. Поэтому вас не должно удивлять, если при использовании робастных стандартных ошибок ваш эконометрический пакет для проверки значимости уравнения или для сравнения «короткой» и «длинной» регрессий начинает использовать тест Вальда вместо F-теста (что приводит к соответствующей корректировке значений тестовых статистик). Интерпретировать его результаты можно аналогичным образом.

Более полное описание процедуры теста Вальда требует знакомства с матричной формой записи для множественной регрессии (параграф 3.3) и обобщенной линейной моделью множественной регрессии (параграф 5.5). Если вы разобрались и с тем, и с другим, то можно читать дальше.

Пусть тестируемая гипотеза, как и в параграфе 3.6, имеет вид:

$$H\beta = r.$$

Здесь β — вектор коэффициентов модели. H — матрица размера q на k , r — вектор-столбец длины q . В свою очередь, q — количество тестируемых ограничений, то есть количество уравнений в системе ограничений, k — число коэффициентов в модели.

Расчетное значение тестовой статистики теста Вальда для тестирования гипотезы $H\beta = r$ имеет вид:

$$(H\hat{\beta} - r)' (H(X'\Omega^{-1}X)^{-1}H')^{-1} (H\hat{\beta} - r)$$

Здесь Ω — ковариационная матрица вектора случайных ошибок.

Если верна тестируемая гипотеза, то эта величина имеет распределение Хи-квадрат с q степенями свободы.

Таким образом, в случае, если расчетное значение тестовой статистики больше критического значения из таблиц распределения Хи-квадрат при заданном уровне значимости, то гипотеза о выполнении ограничения $H\beta = r$ должна быть отвергнута.

На практике ковариационная матрица Ω обычно не известна. Поэтому в формуле тестовой статистики её заменяют оценкой $\hat{\Omega}$. В частности, если выполнены предпосылки линейной модели со стохастическими регрессорами о независимости отдельных наблюдений, то $\hat{\Omega}$ будет диагональной матрицей, где на главной диагонали стоят оценки дисперсий случайных ошибок, полученные по процедуре из главы 5 (см. параграф 5.3, случай 2).

Пример 6.5. Тестирование гипотез в условиях асимптотического подхода

Руководство крупной торговой сети планирует выяснить, помогает ли тренинг по продажам увеличить эффективность работы менеджеров по продажам.

Для решения этой задачи были собраны следующие данные о двух тысячах менеджеров:

sales — объем продаж данного менеджера (в тысячах рублей за период);

training — фиктивная переменная, равная единице, если в самом начале данного периода менеджер прошел тренинг по продажам (работники, которые направлялись на курсы, выбирались из общей совокупности работников компании при помощи специальной лотереи);

female — фиктивная переменная, равная единице для менеджеров-женщин и нулю для мужчин;

experience — опыт работы менеджера в годах;

capital — фиктивная переменная, равная единице, если менеджер работает в столичном отделении компании, и равная нулю в противном случае;

IQ — все менеджеры при приеме на работу в данную компанию проходят IQ-тест, эта переменная характеризует результаты менеджера.

(а) Оценка регрессии переменной sales на переменные training, female, experience, capital и IQ дала следующие результаты:

Модель 1: МНК, использованы наблюдения 1-2000

Зависимая переменная: sales

Робастные оценки стандартных ошибок (с поправкой на гетероскедастичность), вариант HC1

	Коэффициент	Ст. ошибка
const	−24,3731	6,70733
training	18,8678	1,81762
female	0,392756	1,75849
experience	5,30055	0,569546
capital	3,29574	1,60932
IQ	1,72884	0,0647299
Сумма кв. остатков	2574369	
		F(5, 1994)
		180,6451

Используя асимптотический подход:

(а) вычислите Р-значение для коэффициента при переменной female. Проверьте значимость указанной переменной. Интерпретируйте полученный результат.

(б) Постройте 95-процентный доверительный интервал для коэффициента при переменной training. Интерпретируйте полученный результат.

(в) Проверьте незначимость уравнения в целом, используя уровень значимости 1%.

Решение:

(а) Р-значение равно:

$$2 * \Phi \left(- \left| \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \right| \right) = 2 * \Phi \left(\frac{-0,392756}{1,75849} \right) = 2 * \Phi(-0,223348) =$$

$$= 2 * 0,4116 = 0,8232$$

Так как Р-значение больше одной сотой, пяти сотых и десяти сотых, то соответствующий коэффициент не является значимым ни на однопроцентном, ни на пятипроцентном, ни на десятипроцентном уровнях значимости. Следовательно, пол менеджера не влияет на объем его продаж.

Примечание: Значение функции стандартного нормального распределения в нужной точке $\Phi(-0,223348)$ можно вычислить, например, в MS Excel, используя команду =НОРМ.СТ.РАСП(-0,2235;1).

(б) Асимптотический доверительный интервал имеет вид:

$$\left(\hat{\beta}_j - se(\hat{\beta}_j) * 1,96; \hat{\beta}_j + se(\hat{\beta}_j) * 1,96 \right)$$

$$(18,8678 - 1,81762 * 1,96; 18,8678 + 1,81762 * 1,96)$$

$$(15,31; 22,43)$$

Так как этот 95-процентный доверительный интервал не содержит ноль, можно заключить, что посещение тренинга значимо влияет на объем продаж менеджера (при 5-процентном уровне значимости). Причем с вероятностью 95% увеличение объем продаж в результате этого посещения лежит в пределах от 15,31 тысячи рублей до 22,43 тысяч рублей за период.

(в) Расчетное значение тестовой статистики дано по условию и составляет 180,6451. Критическое значение может быть взято из таблиц распределения Фишера: $F^\alpha(q, \infty) = F^{0,01}(5, \infty) = 3,02$.

Или можно получить тот же самый результат, воспользовавшись распределением Хи-квадрат. Критическое значение из таблицы распределения Хи-квадрат для 5 степеней свободы и уровня значимости 1% составляет 15,086. Его следует разделить на $q = 5$, что приведет к получению той же самой величины 3,02.

Поскольку расчетное значение больше критического, гипотеза о незначимости уравнения в целом отвергается. Уравнение в целом значимо (при однопроцентном уровне).

Приложение 6.А. Состоятельная в условиях гетероскедастичности стандартная ошибка оценки коэффициента: доказательство состоятельности

В параграфе 6.4 мы доказали, что асимптотическая дисперсия оценки коэффициента в модели парной регрессии равна:

$$\text{var}(\hat{\beta}_2) = \frac{\text{var}((x_i - \mu_x)\varepsilon_i)}{n * (\text{var}(x_i))^2}.$$

Поскольку в рамках нашей модели мы допускаем наличие гетероскедастичности, оценка этой дисперсии должна быть состоятельной в условиях гетероскедастичности (робастной к гетероскедастичности).

Такая оценка может быть вычислена по формуле:

$$\widehat{\text{var}}(\hat{\beta}_2) = \frac{1}{n} \frac{\frac{1}{n-2} \sum_{i=1}^n (x_i - \bar{x})^2 e_i^2}{\widehat{\text{var}}(x)^2}$$

Если извлечь корень из оценки дисперсии $\widehat{\beta}_2$, то мы получим состоятельную в условиях гетероскедастичности стандартную ошибку оценки коэффициента:

$$se(\hat{\beta}_2) = \sqrt{\widehat{\text{var}}(\hat{\beta}_2)} = \sqrt{\frac{1}{n} \frac{\frac{1}{n-2} \sum_{i=1}^n (x_i - \bar{x})^2 e_i^2}{\widehat{\text{var}}(x)^2}}.$$

Именно эта стандартная ошибка уже встречалась вам в параграфе 5.2 главы про гетероскедастичность. Теперь мы знаем достаточно для того, чтобы доказать её состоятельность.

Для этого мы докажем, что отношение оценки дисперсии $\widehat{\text{var}}(\hat{\beta}_2)$ к истинной теоретической дисперсии $\text{var}(\hat{\beta}_2)$ сходится по вероятности к единице:

$$\frac{\frac{1}{n} \frac{\frac{1}{n-2} \sum_{i=1}^n (x_i - \bar{x})^2 e_i^2}{\widehat{\text{var}}(x)^2}}{\frac{\text{var}((x_i - \mu_x)\varepsilon_i)}{n * (\text{var}(x_i))^2}} \xrightarrow{p} 1$$

Перепишем указанную дробь следующим образом:

$$\frac{\frac{1}{n} \frac{\frac{1}{n-2} \sum_{i=1}^n (x_i - \bar{x})^2 e_i^2}{\widehat{\text{var}}(x)^2}}{\frac{\text{var}((x_i - \mu_x)\varepsilon_i)}{n * (\text{var}(x_i))^2}} = \frac{\frac{n}{n-2} * \left(\frac{\text{var}(x_i)}{\widehat{\text{var}}(x)}\right)^2 * \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 e_i^2}{\text{var}((x_i - \mu_x)\varepsilon_i)}$$

Первый множитель представляет собой неслучайную величину, которая сходится к 1.

Второй множитель сходится по вероятности к 1, так как выборочная дисперсия регрессора сходится к теоретической (см. параграф 6.2).

Осталось доказать, что третий множитель тоже сходится к 1. Рассмотрим его подробнее.

$$\frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 e_i^2}{\text{var}((x_i - \mu_x)\varepsilon_i)} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 e_i^2}{E((x_i - \mu_x)^2 \varepsilon_i^2)}$$

Чтобы доказать, что эта дробь сходится по вероятности к единице, достаточно доказать, что

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 e_i^2 - E((x_i - \mu_x)^2 \varepsilon_i^2) \xrightarrow{p} 0$$

Сделаем это в два шага:

Шаг 1. Докажем, что $\frac{1}{n} \sum_{i=1}^n (x_i - \mu_x)^2 \varepsilon_i^2 \xrightarrow{p} E((x_i - \mu_x)^2 \varepsilon_i^2)$

Шаг 2. Докажем, что $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 e_i^2 - \frac{1}{n} \sum_{i=1}^n (x_i - \mu_x)^2 \varepsilon_i^2 \xrightarrow{p} 0$

Краткое описание каждого из двух указанных шагов приводится ниже.

Шаг 1. Сделаем дополнительную предпосылку:

$$E(x_i^8) < \infty, E(\varepsilon_i^8) < \infty$$

Докажем, что $v_i = (x_i - \mu_x)^2 \varepsilon_i^2$ удовлетворяют предпосылкам закона больших чисел. Для этого нужно доказать, что $\text{var}(v_i) < \infty$.

$$\text{var}(v_i) = E\left((x_i - \mu_x)^2 \varepsilon_i^2\right)^2 = E\left((x_i - \mu_x)^4 \varepsilon_i^4\right) \leq \sqrt{E(x_i - \mu_x)^8 * E\varepsilon_i^8} < \infty$$

Здесь предпоследнее неравенство следует из неравенства Коши-Буняковского, а последнее — из сформулированной нами дополнительной предпосылки.

Следовательно, случайная величина v_i имеет конечный второй начальный момент распределения, который в данном случае равен дисперсии. Значит, она имеет конечный первый момент распределения, то есть матожидание. Тогда все требования закона больших чисел выполнены. В соответствии с ним, получаем, что:

$$\frac{1}{n} \sum_{i=1}^n (x_i - \mu_x)^2 \varepsilon_i^2 \xrightarrow{p} E\left((x_i - \mu_x)^2 \varepsilon_i^2\right)$$

Шаг 2. В рамках второго шага представим анализируемое выражение следующим образом:

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 e_i^2 - \frac{1}{n} \sum_{i=1}^n (x_i - \mu_x)^2 \varepsilon_i^2 = \\ & = \frac{1}{n} \sum_{i=1}^n \left((x_i - \bar{x})^2 e_i^2 - (x_i - \mu_x)^2 \varepsilon_i^2 \right) = \\ & = \frac{1}{n} \sum_{i=1}^n \left((x_i - \bar{x})^2 (\beta_1 + \beta_2 x_i + \varepsilon_i - \widehat{\beta}_1 - \widehat{\beta}_2 x_i)^2 - (x_i - \mu_x)^2 \varepsilon_i^2 \right) \end{aligned}$$

Далее достаточно раскрыть скобки, привести подобные и представить это выражение в виде суммы слагаемых, каждое из которых по теореме Slutsky сходится к нулю. Для доказательства сходимости к нулю каждого из слагаемых достаточно будет снова применить теорему Slutsky и неравенство Коши-Буняковского.

Приложение 6.Б. Дельта-метод

До сих пор мы умели строить доверительные интервалы для отдельных параметров и их линейных комбинаций. Например, для β_1 или суммы $\beta_1 + 5\beta_2$.

Иногда нужно построить доверительный интервал для нелинейного по параметрам выражения. Например, в полиномиальной модели $y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \varepsilon_i$ нам может быть интересен доверительный интервал для вершины параболы, то есть для $-\frac{\beta_2}{2\beta_3}$.

Асимптотическая теория позволяет решить эту задачу при помощи так называемого дельта-метода¹. Опишем его.

Пусть у нас есть некоторая состоятельная и асимптотически нормальная оценка параметра, то есть:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \xi, \text{ где } \xi \sim N(0, \text{var}(\xi)).$$

Или, иными словами, величина $\hat{\beta}$ имеет асимптотически нормальное распределение с математическим ожиданием β и дисперсией $\frac{\text{var}(\xi)}{n}$.

Какое распределение имеет функция от этой оценки $g(\hat{\beta})$?

Чтобы ответить на этот вопрос, вспомним разложение функции в ряд Тейлора в окрестности точки x_0 :

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + o(x), \quad x \rightarrow x_0$$

Применим это разложение к функции от нашей оценки параметра:

$$\begin{aligned} g(\hat{\beta}) &= g(\beta) + g'(\beta)(\hat{\beta} - \beta) + o(\hat{\beta}) \\ g(\hat{\beta}) - g(\beta) &= g'(\beta)(\hat{\beta} - \beta) + o(\hat{\beta}) \\ \sqrt{n}(g(\hat{\beta}) - g(\beta)) &= g'(\beta)\sqrt{n}(\hat{\beta} - \beta) + \sqrt{n} * o(\hat{\beta}) \end{aligned}$$

Так как $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \xi$, где $\xi \sim N(0, \sigma^2)$, то

$$\sqrt{n}(g(\hat{\beta}) - g(\beta)) \xrightarrow{d} g'(\beta)\xi$$

По свойству дисперсии

$$\text{var}(g'(\beta)\xi) = (g'(\beta))^2 * \text{var}(\xi)$$

Следовательно,

$$\sqrt{n}(g(\hat{\beta}) - g(\beta)) \xrightarrow{d} N(0, g'(\beta)^2 * \text{var}(\xi))$$

Поэтому случайная величина $g(\hat{\beta})$ будет иметь асимптотически нормальное распределение с математическим ожиданием $g(\beta)$ и дисперсией $g'(\beta)^2 * \text{var}(\xi)/n$.

На практике величина $\text{var}(\xi)$ неизвестна, но для целей построения асимптотического доверительного интервала можно заменить её оценкой.

Чтобы понять, как работает дельта-метод, рассмотрим модель парной регрессии $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, для которой выполнены все предпосылки линейной модели со стохастическим регрессором из параграфа 6.2. Построим доверительный интервал для функции от оценки коэффициента при переменной $g(\hat{\beta}_2)$.

В параграфе 6.4 мы доказали, что

$$\sqrt{n}(\hat{\beta}_2 - \beta_2) \xrightarrow{d} N\left(0, \frac{\text{var}((x_i - \mu_x)\varepsilon_i)}{(\text{var}(x_i))^2}\right)$$

Следовательно, в обозначениях данного приложения:

$$\text{var}(\xi) = \frac{\text{var}((x_i - \mu_x)\varepsilon_i)}{(\text{var}(x_i))^2}$$

Таким образом, случайная величина $g(\hat{\beta}_2)$ будет иметь асимптотически нормальное распределение с математическим ожиданием $g(\beta_2)$ и дисперсией $g'(\beta_2)^2 * \frac{\text{var}((x_i - \mu_x)\varepsilon_i)}{(\text{var}(x_i))^2 * n}$. Отметим, что дробь в последнем произведении — это просто дисперсия МНК-оценки коэффициента при переменной, так как в конце параграфа 6.4 мы доказали, что:

$$\text{var}(\widehat{\beta}_2) = \frac{\text{var}((x_i - \mu_x)\varepsilon_i)}{(\text{var}(x_i))^2 * n}.$$

Поэтому случайная величина $g(\widehat{\beta}_2)$ будет иметь асимптотически нормальное распределение с математическим ожиданием $g(\beta_2)$ и дисперсией $g'(\beta_2)^2 * \text{var}(\widehat{\beta}_2)$.

Заменим в последнем выражении неизвестные случайные величины их оценками и получим оценку дисперсии $g(\widehat{\beta}_2)$, которая будет равна: $g'(\widehat{\beta}_2)^2 * \text{var}(\widehat{\beta}_2)$. Если извлечь из этой величины корень, то получим соответствующую стандартную ошибку:

$$\sqrt{g'(\widehat{\beta}_2)^2 * \text{var}(\widehat{\beta}_2)} = |g'(\widehat{\beta}_2)| * se(\widehat{\beta}_2)$$

95-процентный асимптотический доверительный интервал для величины $g(\beta_2)$ будет иметь вид:

$$(g(\widehat{\beta}_2) - 1,96 * |g'(\widehat{\beta}_2)| * se(\widehat{\beta}_2); g(\widehat{\beta}_2) + 1,96 * |g'(\widehat{\beta}_2)| * se(\widehat{\beta}_2)).$$

Аналогичным образом могут быть построены доверительные интервалы и для коэффициентов в модели множественной регрессии.

Пример 6.6. Применение дельта-метода для парной регрессии

Оценка параметров модели при помощи МНК позволила получить следующие результаты:

$$\hat{y}_i = \underset{(0,2)}{2,3} + \underset{(0,1)}{4,0}x_i$$

Постройте 95-процентный асимптотический доверительный интервал для величины $(\beta_2)^3$.

Решение:

$$\begin{aligned} & (g(\widehat{\beta}_2) - 1,96 * g'(\widehat{\beta}_2) * se(\widehat{\beta}_2); g(\widehat{\beta}_2) + 1,96 * g'(\widehat{\beta}_2) * se(\widehat{\beta}_2)) \\ & ((\widehat{\beta}_2)^3 - 1,96 * 3 * (\widehat{\beta}_2)^2 * se(\widehat{\beta}_2); (\widehat{\beta}_2)^3 + 1,96 * 3 * (\widehat{\beta}_2)^2 * se(\widehat{\beta}_2)) \\ & (4^3 - 1,96 * 3 * 4^2 * 0,1; 4^3 + 1,96 * 3 * 4^2 * 0,1) \\ & (54,6; 73,4) \end{aligned}$$

Полученный нами результат может быть обобщен и на многомерный случай:

Теорема о дельта-методе.

Если

1. $\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{d} \xi$ (где ξ — случайный вектор),
2. $g(x)$ — вектор непрерывно дифференцируемых функций в окрестности точки β ,

то

$$\sqrt{n}(g(\widehat{\beta}) - g(\beta)) \xrightarrow{d} G(\beta)^T \xi,$$

$G(x)$ — это матрица частных производных функций из вектора $g(x)$. Число строк в этой матрице равно длине вектора $\widehat{\beta}$, а число столбцов — вектора $g(x)$.

В частности, **если** ξ имеет распределение $N(\vec{0}, V)$, **то**

$$\sqrt{n}(g(\widehat{\beta}) - g(\beta)) \xrightarrow{d} N(\vec{0}, G(\beta)^T V G(\beta))$$

Примечание: здесь значок T означает транспонирование.

1. В русскоязычной литературе он также иногда называется методом построения доверительного интервала со стабилизацией дисперсии. [↗](#)

Приложение 6.В. Таблицы стандартного нормального распределения и распределения Хи-квадрат

Критические значения t-статистик для больших выборок из стандартного нормального распределения			
	Уровень значимости		
	10%	5%	1%
Двусторонний тест	1,64	1,96	2,58

Критические значения для Хи-квадрат распределения Для уровней значимости 5% и 1%												
k	1	2	3	4	5	6	7	8	9	10	11	12
5%	3,84	5,99	7,81	9,49	11,07	12,59	14,07	15,51	16,92	18,31	19,68	21,03
1%	6,63	9,21	11,34	13,28	15,09	16,81	18,48	20,09	21,67	23,21	24,72	26,22

Глава 7. Проблемы спецификации уравнения регрессии

В этой главе мы сконцентрируемся на том, как при помощи эконометрики получать корректные ответы на вопросы о причинно-следственных связях. Чтобы это сделать, нужно верно специфицировать вашу модель. Под верной спецификацией будем понимать такую, которая позволяет получить состоятельные оценки коэффициентов при интересующих вас переменных. А также получить состоятельные стандартные ошибки для тестирования гипотез.

Глава будет устроена так: мы будем перечислять типичные ловушки, которые приводят к неверной спецификации. Далее для каждой такой ловушки мы будем указывать возможные способы избежать её и устранить проблему.

В каких-то случаях мы будем опираться на уже знакомые вам концепции и понятия. В некоторых же ситуациях мы будем, наоборот, ссылаться на более продвинутые методы и модели, с которыми нам ещё предстоит разобраться в следующих главах учебника (надеюсь, это станет для вас дополнительной мотивацией все-таки дочитать его до конца).

Напомним, что в предыдущей главе мы сформулировали два важных определения:

- Эндогенный регрессор — регрессор, который коррелирован со случайными ошибками модели.
- Экзогенный регрессор — регрессор, который не коррелирован со случайными ошибками модели.

Кроме того, в той же главе мы выяснили, что для состоятельности оценки коэффициента при переменной необходимо, чтобы эта переменная была экзогенной (точнее, необходимо выполнение предпосылки №4 линейной регрессионной модели со стохастическими регрессорами из главы 6). Если же регрессор эндогенный, результаты вашего моделирования нельзя интерпретировать в терминах причинно-следственных связей. Нарушение предпосылки №4 об экзогенности регрессора — это самая частая проблема при проведении прикладных исследований на пространственных и панельных данных. Поэтому важно понимать, в каких случаях вам следует опасаться её возникновения. Есть следующие типичные ситуации:

1. Эндогенность регрессора из-за пропуска существенной переменной. В качестве важного частного случая тут также следует указать проблему эндогенности из-за самоотбора.
2. Эндогенность регрессора из-за выбора неверной функциональной формы связи.
3. Эндогенность регрессора из-за двусторонней причинно-следственной связи.
4. Эндогенность регрессора из-за ошибок измерения.

В последующих четырех параграфах главы мы подробно обсудим каждый из этих пунктов. В пятом параграфе мы поговорим о других (помимо эндогенности) проблемах, которые могут делать выводы эконометрических исследований необоснованными. В каждом случае мы также укажем основные возможные пути преодоления перечисленных трудностей.

[7.1. Эндогенность из-за пропуска существенной переменной](#)

[7.2. Эндогенность из-за выбора неверной функциональной формы связи](#)

[7.3. Эндогенность из-за двусторонней причинно-следственной связи](#)

[7.4. Эндогенность из-за ошибок измерения](#)

[7.5. Другие \(помимо эндогенности\) потенциальные угрозы обоснованности выводов эконометрического исследования](#)

[7.6. Чек-лист эконометриста](#)

[Задания для самостоятельного решения](#)

7.1. Эндогенность из-за пропуска существенной переменной

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

Пусть выполнены все предпосылки линейной модели со стохастическими регрессорами, и на зависимую переменную влияют два фактора:

$$y_i = \beta_1 + \beta_2 * x_i + \beta_3 * w_i + \varepsilon_i, \quad \beta_3 \neq 0 \quad (7.1)$$

[Файл с примером](#)

Представим, что мы игнорируем второй фактор и оцениваем парную регрессию переменной y по переменной x .

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i.$$

В третьей главе мы показали, что в этом случае МНК-оценка будет, вообще говоря, смещена, что само по себе является достаточно серьезной проблемой. На самом деле, в большинстве ситуаций (кроме одного частного случая) она будет ещё и несостоятельной. То есть возникшую проблему нельзя будет компенсировать использованием сколь угодно большого массива данных. Докажем это:

$$\begin{aligned} \hat{\beta}_2 &= \frac{\widehat{\text{cov}}(x, y)}{\widehat{\text{var}}(x)} \xrightarrow{\text{P}} \frac{\text{cov}(x, y_i)}{\text{var}(x_i)} = \frac{\text{cov}(x_i, \beta_1 + \beta_2 * x_i + \beta_3 * w_i + \varepsilon_i)}{\text{var}(x_i)} = \\ &= \frac{\beta_2 * \text{cov}(x_i, x_i) + \beta_3 * \text{cov}(x_i, w_i) + \text{cov}(x_i, \varepsilon_i)}{\text{var}(x_i)} = \\ &= \beta_2 + \beta_3 \frac{\text{cov}(x_i, w_i)}{\text{var}(x_i)} + \frac{\text{cov}(x_i, \varepsilon_i)}{\text{var}(x_i)} = \beta_2 + \beta_3 \frac{\text{cov}(x_i, w_i)}{\text{var}(x_i)} \end{aligned}$$

Из этого соотношения ясно, что если, например, $\beta_3 > 0$ и $\text{cov}(x_i, w_i) > 0$, то МНК-оценка коэффициента β_2 в парной регрессии будет несостоятельна и завышена:

$$\hat{\beta}_2 \xrightarrow{\text{P}} \beta_2 + \beta_3 \frac{\text{cov}(x_i, w_i)}{\text{var}(x_i)} > \beta_2.$$

Отметим, что в этом случае интересующая нас переменная $\mathbf{x}$ действительно оказывается эндогенной (что, собственно, и приводит к проблеме). Действительно: при пропуске переменной $\mathbf{w}$ этот пропущенный фактор как бы остается внутри случайной ошибки. Иными словами, исходную модель можно переписать вот так:

$$y_i = \beta_1 + \beta_2 * x_i + u_i,$$

Где $u_i = \beta_3 * w_i + \varepsilon_i$. Поэтому $\text{cov}(x_i, u_i) = \text{cov}(x_i, \beta_3 * w_i + \varepsilon_i) = \beta_3 \text{cov}(x_i, w_i) + \text{cov}(x_i, \varepsilon_i) = \beta_3 \text{cov}(x_i, w_i) > 0$. То есть, хотя регрессор $\mathbf{x}$ не коррелирован со случайной ошибкой исходной модели (ε), однако он коррелирован со случайной ошибкой оцениваемой парной регрессии ($\mathbf{u}$).

Единственный случай, в котором регрессор остается экзогенным, а МНК-оценка останется состоятельной — это ситуация некоррелированности интересующего нас регрессора и пропущенной переменной: $\text{cov}(x_i, w_i) = 0$. Если это так, то:

$$\hat{\beta}_2 \xrightarrow{\text{P}} \beta_2 + \beta_3 \frac{0}{\text{var}(x_i)} = \beta_2.$$

Как можно решить проблему эндогенности регрессора из-за пропуска существенной переменной? Ответ на этот вопрос зависит от того, в какой из двух возможных ситуаций мы находимся: в простой или в сложной. Рассмотрим их последовательно.

Ситуация 1 (простая). Пропущенная переменная наблюдаема

Эта ситуация, в которой у вас есть данные о пропущенной переменной. Тогда, как вы наверняка уже догадались, для решения проблемы нужно просто добавить пропущенную переменную в модель. В реальных исследованиях таких пропущенных переменных обычно не одна, а несколько (так как мир устроен сложно, и на зависимую переменную обычно влияют сразу много факторов). Что ж, тогда нужно добавить их все.

Переменные, которые вы добавляете в модель, чтобы устранить смещение оценки нужного вам коэффициента, называют контрольными.

Сформулируем два определения:

- **Переменная интереса** (variable of interest) — фактор, влияние которого на зависимую переменную нас интересует.
- **Контрольные переменные** (control variables) — переменные, которые мы включаем в модель для того, чтобы избежать смещения коэффициента при интересующей нас переменной.

Получение состоятельных оценок коэффициентов при контрольных переменных для исследователя зачастую не критично. И в целом получить состоятельные оценки коэффициентов при каждой переменной в модели множественной регрессии — это часто слишком амбициозная задача. Состоятельно оценить влияние переменной интереса — это уже успех.

Обратите внимание: с технической точки зрения (например, с точки зрения формул для вычисления МНК-оценок) нет разницы между контрольными переменными и переменными интереса. Разделение связано только с содержательными соображениями: интересующим вас исследовательским вопросом.

Пусть, например, ваш вопрос: влияет ли образование на уровень доходов? В этом случае в регрессии:

$$\text{EARNINGS}_i = \beta_1 + \beta_2 S_i + \beta_3 \text{EXP}_i + \beta_4 \text{FEMALE}_i + \varepsilon_i$$

переменная S , обозначающая число лет обучения респондента, будет переменной интереса. А переменные EXP и FEMALE , характеризующие опыт работы и пол респондента, соответственно, будут контрольными.

Однако если ваша статья посвящена исследованию дискриминации на рынке труда, то вы можете оценивать ту же самую модель, считая переменной интереса регрессор **FEMALE**, так как именно он обозначает интересную вам характеристику работника. Контрольными переменными в этом случае будут факторы **S** и **EXP**.

Так как включение дополнительных переменных в модель позволяет избежать несостоятельности коэффициентов, кажется, что разумно включать их в модель как можно больше. Просто на всякий случай. Приведет ли это к каким-то проблемам? Иными словами, к каким последствиям приводит включение в модель несущественного регрессора? Несущественным мы будем называть регрессор, который на самом деле не оказывает никакого влияния на зависимую переменную.

Последствия включения в модель несущественной переменной:

1. Коэффициенты при прочих переменных остаются несмещенными и состоятельными. Действительно, если новый регрессор не влияет на зависимую переменную, то можно всё равно считать, что он входит в модель, просто истинный коэффициент при нём равен нулю.

2. Из-за необходимости оценивать большее количество коэффициентов, а также из-за вероятной мультиколлинеарности увеличивается дисперсия оценок коэффициентов, то есть снижается точность модели.

Таким образом, включить в модель лишнюю переменную не так страшно, как пропустить нужную. Ведь в первом случае нет несостоятельности оценок, а во втором она возникает. Однако слишком много несущественных переменных включать в уравнение нецелесообразно, так как это негативно сказывается на точности ваших результатов.

Поэтому хочется иметь набор правил, чтобы принимать решение, включать ли ту или иную переменную в уравнение. Ниже приведены некоторые соображения по этому поводу.

Критерии для включения переменной в модель:

1. Роль переменной в уравнении опирается на прочные теоретические основания. Ну или хотя бы на здравый смысл.

2. Переменная статистически значима

3. Оценки других коэффициентов сильно меняются при включении новой переменной в модель. Это значит, что до этого они страдали от смещения из-за пропуска существенной переменной. Теперь вы эту существенную переменную добавили, и смещение пропало.

4. Скорректированный R-квадрат существенно увеличивается в результате включения переменной в модель

Ситуация 2 (сложная). Пропущенная переменная не наблюдаема

Эта ситуация, в которой у вас отсутствуют данные о пропущенной переменной и достать их невозможно. В этом случае говорят, что пропущенная переменная является ненаблюдаемой.

Важным частным случаем является проблема самоотбора. Представим, что вы снова обратились к одному из наших любимых примеров: оценке влияния образования на уровень дохода. Вполне возможно, что индивиды принимают решение, получать ли им образование или нет, в зависимости от некоторого не наблюдаемого исследователем фактора. Скажем, уровня таланта. Более талантливые индивиды после школы чаще принимают решение продолжить обучение в университете. Это и называется самоотбором. Тогда в терминах нашего уравнения (6.1) y_i — это доход i -го индивида, x_i — уровень его образования, а w_i — ненаблюдаемый уровень таланта. Причем $\text{cov}(x_i, w_i) > 0$, так как талантливые люди чаще решают получить высшее образование, и $\beta_3 > 0$, так как талант в среднем способствует получению более высоких доходов.

Ясно, что это как раз тот случай, когда

$$\widehat{\beta_2} \xrightarrow{P} \beta_2 + \beta_3 \frac{\text{cov}(x_i, w_i)}{\text{var}(x_i)} > \beta_2,$$

и МНК-оценка будет завышать пользу от образования.

Так как в этой ситуации мы не можем просто включить в уравнение нужную переменную, придется придумать что-то другое. К счастью, можно указать целых четыре **пути решения проблемы эндогенности из-за пропуска ненаблюдаемой существенной переменной**:

Рассмотрим каждый из этих путей.

Замещающие переменные. Замещающей переменной называется переменная, которая тесно коррелирована с ненаблюдаемой существенной переменной и при этом является наблюдаемой.

В нашем примере с образованием такой переменной могли бы быть результаты IQ-теста. Ясно, что никакой тест не способен в полной мере описать природные способности человека, однако также ясно, что результаты хороших тестов будут коррелированы с этими способностями.

Другой пример: представим, что вам важно измерить то, насколько удобно вести бизнес в данной стране. Удобство ведения бизнеса зависит от множества параметров, которые трудно измерить количественно, например, характеристики законодательства; то, насколько это законодательство соблюдается; уровень коррупции и так далее. Поэтому напрямую переменную «удобство ведения бизнеса» включить в регрессионное уравнение не получится. Однако вместо неё вы можете использовать в качестве замещающей переменной один из многочисленных индексов, которые рассчитываются разными службами для оценки качества бизнес-среды. Например, индекс Ease of Doing Business Index (индекс легкости ведения бизнеса), который рассчитывается Мировым банком для сопоставления степени простоты ведения предпринимательской деятельности в разных странах.

Включение замещающей переменной устраняет несостоятельность оценки коэффициента при регрессоре из-за пропуска существенного ненаблюдаемого фактора.

Инструментальные переменные. В некоторых случаях у вас отсутствует переменная, тесно коррелированная с пропущенным ненаблюдаемым фактором, поэтому применить подход с замещающими переменными тоже невозможно. Зато вы можете отыскать данные о переменной, которая, наоборот, **вообще не коррелирована** с пропущенным ненаблюдаемым фактором и при этом коррелирована с вашей переменной интереса.

В нашем примере в странах, где образование преимущественно платное, такой переменной мог бы быть, например, доход родителей индивида. Вряд ли богатство родителей гарантирует талант ребенка, однако оно даёт ему гораздо больше возможностей для продолжения обучения.

Оказывается, что такую переменную также можно применить для решения эндогенности. Как это сделать, мы подробно обсудим в главе 8.

Модели с фиксированными эффектами. Если вы располагаете данными за несколько периодов и пропущенный ненаблюдаемый фактор с течением времени меняется медленно (или вообще не меняется), то для получения состоятельных оценок коэффициентов подойдут модели, использующие панельные данные. В частности, модели с фиксированными эффектами. Их применение подробно обсуждается в главе 9.

Контролируемый эксперимент. Представим, что в нашем примере был выпущен закон, в соответствии с которым в определенном регионе страны индивидам отныне запрещено самостоятельно решать, сколько лет они будут учиться. Теперь это определяется случайным образом, скажем, в ходе специальной лотереи. Законом строго предписывается, что каждому следует учиться ровно столько, сколько ему выпало в лотерее, и государству удастся обеспечить его неукоснительное выполнение.

Конечно, на практике трудно представить себе такую ситуацию, однако, используя свою силу воображения, вы сможете это сделать. Окажется, что в этом случае талант индивида больше никак не коррелирован с продолжительностью обучения (так как эта продолжительность определяется случайным образом, независимым от таланта): $\text{cov}(x_i, w_i) = 0$. Поэтому теперь МНК-оценка отдачи от образования снова станет состоятельной:

$$\widehat{\beta}_2 \xrightarrow{p} \beta_2 + \beta_3 \frac{0}{\text{var}(x_i)} = \beta_2.$$

Мы описали пример контролируемого эксперимента. Можно видеть, что такой подход также решает проблему эндогенности. В главе 1 мы уже упоминали преимущества использования экспериментальных данных. Более детально мы обсудим этот вопрос в главе 11, где будет освещен ряд соответствующих продвинутых методов¹.

В этом параграфе мы проанализировали широкий арсенал методов устранения эндогенности, вызванной пропуском существенной переменной. В следующих параграфах нас ждет аналогичное обсуждение для других источников эндогенности.

1. Примеры контролируемых экспериментов в оценке эффективности образования также вовсе не всегда являются утопией. Так, в 80-х годах XX века в Далласе несколько тысяч школьников в ходе контролируемого эксперимента случайным образом распределялись по классам разного размера. Детали этой истории вы можете найти в главе 11 данного учебника или в работе Kreuger (1999) *Experimental Estimates of Education Production Functions* // *The Quarterly Journal of Economics*. [↗](#)

7.2. Эндогенность из-за выбора неверной функциональной формы связи

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

Неверная функциональная форма уравнения приводит к эндогенности регрессора. Действительно, представим, что в нашем примере про образование отдача от него является убывающей и описывается квадратичной функцией:

$$y_i = \beta_1 + \beta_2 * x_i + \beta_3 * x_i^2 + \varepsilon_i, \quad \beta_2 > 0, \quad \beta_3 < 0$$

Рассмотрение линейной модели вместо нелинейной эквивалентно пропуску существенной переменной (в данном случае это переменная x^2) и приводит к таким же последствиям¹.

Для решения указанной проблемы следует отыскать корректную функциональную форму связи. Соображения, которые позволяют это сделать, изложены в параграфе 4.3 главы 4. Поэтому здесь мы лишь кратко напомним, что для выявления верной функциональной формы могут быть полезны следующие шаги:

1. осуществите графический анализ исходных данных и графический анализ остатков оцененного уравнения регрессии,
2. опирайтесь на экономическую теорию или другие содержательные соображения по поводу природы анализируемых переменных (в конце концов, эконометрику можно применять для ответов не только на экономические вопросы),
3. используйте формальные статистические критерии.

1. Если истинная модель является не квадратичной, а какой-либо иной, например, линейно-логарифмической, то это сохраняет вывод об эндогенности регрессора в неверно специфицированном уравнении. Формально в этом нетрудно убедиться, вспомнив, что логарифм можно аппроксимировать квадратичной функцией или полиномом более высокого порядка. [↪](#)

7.3. Эндогенность из-за двусторонней причинно-следственной связи

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

Следующая важная ситуация, которая приводит к эндогенности — одновременная причинно-следственная связь (её также называют двусторонней или двунаправленной).

Снова воспользуемся примером. Представим, что нас интересует, насколько эффективно работает полиция в некотором городе. Мы полагаем, что увеличение числа полицейских в определенном районе должно приводить к уменьшению числа преступлений в этом районе:

$$y_i = \beta_1 + \beta_2 * x_i + \varepsilon_i, \quad \beta_2 < 0,$$

где y_i — количество преступлений в i -ом районе, x_i — число полицейских в i -ом районе. Если в какой-то район мы отправили много полицейских, то количество преступлений в нём должно сокращаться. Говоря совсем кратко, получается, что чем больше x , тем меньше y . Поэтому мы и ожидаем, что $\beta_2 < 0$. Предположим, что городской департамент полиции планирует количественно оценить указанный эффект и нанял для этого нас. Для достижения указанной цели мы нуждаемся в состоятельной оценке коэффициента β_2 .

Проблема, однако, в том, что кроме влияния числа полицейских на число преступлений наверняка существует и обратная причинно-следственная связь. Скорее всего, чем более криминогенным является данный район, тем больше туда будет направлено служителей правопорядка. В виде уравнения это влияние можно описать, например, вот так:

$$x_i = \alpha_1 + \alpha_2 * y_i + u_i, \quad \alpha_2 > 0$$

Положительный коэффициент α_2 отражает гипотезу о том, что чем выше в районе преступность, тем больше туда будет отправлено полицейских.

В нашем примере не только x влияет на y , но и y влияет на x . Поэтому мы имеем дело не с отдельным уравнением, а с системой одновременных уравнений:

$$\begin{cases} y_i = \beta_1 + \beta_2 * x_i + \varepsilon_i, & \beta_2 < 0 \\ x_i = \alpha_1 + \alpha_2 * y_i + u_i, & \alpha_2 > 0 \end{cases} \quad (7.2)$$

Покажем, что, даже если случайные ошибки двух уравнений не связаны друг с другом: $\text{cov}(u_i, \varepsilon_i) = 0$, регрессор в нашей модели все равно будет эндогенным. Для этого вычислим его ковариацию со случайной ошибкой в первом уравнении системы. Чтобы это сделать, удобно провести следующее преобразование: выразить эндогенные переменные x_i и y_i через экзогенные ε_i и u_i .

Подставим первое уравнение системы во второе:

$$x_i = \alpha_1 + \alpha_2 * (\beta_1 + \beta_2 * x_i + \varepsilon_i) + u_i$$

Выразив x_i , получаем:

$$x_i = \frac{\alpha_1 + \alpha_2 \beta_1 + \alpha_2 \varepsilon_i + u_i}{1 - \alpha_2 \beta_2},$$

Аналогичную операцию можно проделать со второй переменной. Получим новую систему:

$$\begin{cases} x_i = \frac{\alpha_1 + \alpha_2 \beta_1 + \alpha_2 \varepsilon_i + u_i}{1 - \alpha_2 \beta_2} \\ y_i = \frac{\beta_1 + \alpha_1 \beta_2 + \varepsilon_i + \beta_2 u_i}{1 - \alpha_2 \beta_2} \end{cases} \quad (7.3)$$

Система (7.2) называется системой в структурной форме. Система (7.3) называется системой в приведенной форме, так как в ней эндогенные переменные выражены через экзогенные.

Такое представление даёт возможность легко посчитать ковариацию между x_i и ε_i . Для этого достаточно воспользоваться первым уравнением системы (7.3):

$$\begin{aligned} \text{cov}(x_i, \varepsilon_i) &= \text{cov}\left(\frac{\alpha_1 + \alpha_2 \beta_1 + \alpha_2 \varepsilon_i + u_i}{1 - \alpha_2 \beta_2}, \varepsilon_i\right) = \frac{\alpha_2 * \text{cov}(\varepsilon_i, \varepsilon_i)}{1 - \alpha_2 \beta_2} = \\ &= \frac{\alpha_2 * \text{var}(\varepsilon_i)}{1 - \alpha_2 \beta_2} > 0, \end{aligned}$$

Чтобы определить знак последней дроби, мы воспользовались первоначальными ограничениями на параметры: $\alpha_2 > 0$ и $\beta_2 < 0$.

Оказывается, что из-за того, что ситуация описывается не одним уравнением, а системой, регрессор x эндогенен: он коррелирован со случайной ошибкой.

Как мы доказали ранее в главе 6 (см. равенство (6.1) в параграфе 6.3), МНК-оценки коэффициентов сходятся по вероятности к следующей величине:

$$\hat{\beta}_2 \xrightarrow{\text{p}} \beta_2 + \frac{\text{cov}(x_i, \varepsilon_i)}{\text{var}(x_i)}.$$

В нашем случае это означает следующее:

$$\hat{\beta}_2 \xrightarrow{\text{p}} \beta_2 + \frac{\text{cov}(x_i, \varepsilon_i)}{\text{var}(x_i)} = \left(\beta_2 + \frac{1}{\text{var}(x_i)} * \frac{\alpha_2 * \text{var}(\varepsilon_i)}{1 - \alpha_2 \beta_2} \right) > \beta_2$$

Мы показали, что МНК-оценка будет несостоятельна и завышена. Так как $\beta_2 < 0$, то в нашем примере при использовании обычного МНК мы часто будем получать оценку, сдвинутую к нулю. Иными словами, мы будем недооценивать эффективность полиции. Если эффект завышения силен, то мы получим оценку коэффициента $\hat{\beta}_2$ настолько близкую к нулю, что даже сочтем её

незначимой, то есть сделаем ошибочный вывод о бесполезности полиции. В крайнем случае возможна даже ситуация $\hat{\beta}_2 > 0$, что может привести нас к решению о том, что полиция способствует увеличению, а вовсе не снижению количества преступлений.

Этот пример показывает, что в ситуации двусторонней причинно-следственной связи обычный МНК применять не следует, так как его использование приведёт к несостоятельности интересующей нас оценки даже при использовании больших выборок.

Проблема одновременной причинно-следственной связи и сопутствующие ей трудности при оценивании встречаются в исследованиях часто. Гораздо чаще, чем хотелось бы прикладным эконометристам. Представим, например, традиционную проблему идентификации при оценке последствий макроэкономической политики: вопрос о том, как изменение ключевой ставки процента¹ влияет на уровень инфляции? При ответе на этот вопрос придется принимать во внимание, что, с одной стороны, увеличение ключевой ставки сдерживает рост инфляции (то есть ключевая ставка влияет на инфляцию). Однако, с другой стороны, увеличение инфляции может побудить центральный банк изменить свою политику (то есть инфляция влияет на политику центрального банка и, следовательно, на ключевую ставку). Даже простой анализ конкурентного равновесия на рынке некоторого товара приведёт нас к системе по меньшей мере из двух одновременных уравнений: кривой спроса и кривой предложения, что снова потребует анализа одновременной причинно-следственной связи.

Обратите внимание, что в нашем примере про полицию в модели нет никаких пропущенных переменных, поэтому включение их в уравнение не может решить проблему. Придется использовать другие **пути получения состоятельной оценки коэффициента в условиях двусторонней причинно-следственной связи**: (1) используйте инструментальные переменные, (2) используйте методы, основанные на оценке систем одновременных уравнений (вместо оценки отдельных уравнений), (3) осуществите контролируемый эксперимент.

Применительно к пространственным и панельным данным в современных исследованиях более распространен первый подход. Мы обсудим его в главе 8. Второй подход в настоящее время в основном используются для временных рядов² и потому выходит за рамки данного издания нашего вводного учебника.

Наконец, использование экспериментов подробно обсуждается в главе 11. По аналогии с предыдущими примерами вы вполне можете догадаться, как может быть устроен эксперимент в данном случае. Департаменту полиции следует определять количество полицейских в ряде районов случайным образом, то есть независимо от наблюдаемого там уровня преступности. В этом случае обратная причинно-следственная связь в указанных районах пропадет, что позволит состоятельно оценить коэффициент β_2 при помощи обычного МНК.

1. Напомним, что ключевая ставка — это ставка, по которой центральный банк предоставляет кредиты коммерческим банкам. Она играет решающую роль при установлении процентных ставок по банковским кредитам. [↗](#)

2. См. модели структурных векторных авторегрессий (structural vector autoregression, SVAR) или векторной коррекции ошибок (vector error correction model, VECM). [↗](#)

7.4. Эндогенность из-за ошибок измерения

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

Пусть переменные y_i и x_i^* связаны точным соотношением

$$y_i = \beta_1 + \beta_2 * x_i^*$$

Однако вместо точных значений регрессора мы наблюдаем измеренные с ошибкой значения: $x_i = x_i^* + \varepsilon_i$, $\text{cov}(x_i^*, \varepsilon_i) = 0$.

Мы оцениваем методом наименьших квадратов уравнение

$$y_i = \beta_1 + \beta_2 * x_i + u_i.$$

Покажем, что и в этом случае МНК-оценка $\widehat{\beta}_2$ будет несостоятельной.

Так как $y_i = \beta_1 + \beta_2 * (x_i - \varepsilon_i) = \beta_1 + \beta_2 * x_i - \beta_2 * \varepsilon_i$, то

$$u_i = -\beta_2 * \varepsilon_i.$$

$$\begin{aligned} \widehat{\beta}_2 &\xrightarrow{P} \beta_2 + \frac{\text{cov}(x_i, u_i)}{\text{var}(x_i)} = \beta_2 + \frac{\text{cov}(x_i^* + \varepsilon_i, -\beta_2 * \varepsilon_i)}{\text{var}(x_i)} = \\ &= \beta_2 - \beta_2 \frac{\text{cov}(x_i^*, \varepsilon_i) + \text{cov}(\varepsilon_i, \varepsilon_i)}{\text{var}(x_i)} = \beta_2 - \beta_2 \frac{\text{var}(\varepsilon_i)}{\text{var}(x_i)} = \\ &= \beta_2 - \beta_2 \frac{\text{var}(\varepsilon_i)}{\text{var}(\varepsilon_i) + \text{var}(x_i^*)} = \frac{\text{var}(x_i^*)}{\text{var}(\varepsilon_i) + \text{var}(x_i^*)} * \beta_2 \end{aligned}$$

Величина $\left| \frac{\text{var}(x_i^*)}{\text{var}(\varepsilon_i) + \text{var}(x_i^*)} \right| < 1$, поэтому независимо от знака β_2 эта оценка несостоятельна и смещена к нулю.

Можно привести много примеров ситуаций, когда в эконометрическом исследовании приходится мириться с ошибками измерения. Скажем, если вашим регрессором является уровень безработицы или валовой внутренний продукт, вы неизбежно столкнетесь с этой проблемой, так как статистические службы не могут измерить указанные показатели идеально точно.

Исследования, опирающиеся на индивидуальные данные, также иногда связаны с ошибками измерений. Типичная ситуация тут — использование данных, основанных на опросах. Если регрессором в вашей модели является возраст индивида, информация о котором собрана в ходе опроса (например, в процессе переписи населения), то, скорее всего, в измерениях будут содержаться неточности: демографам хорошо известно, что многие индивиды склонны при ответах на вопросы о возрасте округлять его до чисел, кратных пяти или десяти годам. Похожий эффект возникает и в случае ответов на вопросы о доходе.

Конечно, в условиях ошибок измерений всегда можно посоветовать исследователю найти данные поточнее. Это хороший совет. Однако, к сожалению, на практике последовать ему бывает трудно, поэтому приходится использовать альтернативный путь.

Как мы выясним в главе 8, проблема ошибок измерения также может быть решена при помощи инструментальных переменных.

Иногда, однако, эту проблему просто игнорируют. Мотивация тут такая: если вам интересна не количественная оценка силы влияния переменной x на переменную y , а просто сам факт наличия или отсутствия этого влияния, то, получив статистически значимый коэффициент при регрессоре, вы можете не предпринимать дальнейших корректировок. Действительно, мы точно знаем, что ошибки измерения сдвигают оценку коэффициента к нулю. Поэтому, если коэффициент оказался значимым даже в условиях ошибок измерения, то после их устранения он тем более должен быть значим.

7.5. Другие (помимо эндогенности) потенциальные угрозы обоснованности выводов эконометрического исследования

Хотя эндогенность является наиболее существенным препятствием для получения надежных выводов при помощи эконометрики, есть и другие аспекты, о которых не следует забывать в процессе моделирования. В этом разделе мы прокомментируем основные из них:

- Мультиколлинеарность
- Нарушение предпосылок о гомоскедастичности или о независимости случайных ошибок
- Неоднородность выборки
- Угрозы внешней обоснованности выводов

Мультиколлинеарность

Мультиколлинеарность не является первоочередной проблемой, так как её наличие в модели не приводит к смещению оценок коэффициентов. Поэтому если вы не наблюдаете серьезных негативных последствий мультиколлинеарности, то и бороться с ней не нужно. Напомним (см. детали в параграфе 3.1), что под серьезными проблемами в этом контексте мы понимаем сильные проявления следующих традиционных симптомов мультиколлинеарности:

— **Неустойчивость результатов.** Небольшое изменение исходных данных приводит к существенному изменению оценок коэффициентов.

— **Незначимость большинства переменных.** Каждая переменная в отдельности является незначимой, а уравнение в целом является значимым и характеризуется высоким R^2 .

— **Неправдоподобность результатов.** Оценки коэффициентов имеют неправильные с точки зрения экономической теории знаки или неоправданно большие значения.

Однако даже если вы столкнулись с этими признаками, то прежде, чем бороться с мультиколлинеарностью, подумайте, не обусловлены ли указанные странности эндогенностью регрессоров? Например, неправдоподобные знаки коэффициентов могут быть вызваны смещением из-за пропуска существенной переменной, а незначимость переменной интереса может объясняться ошибками измерения. В такой ситуации нужно устранять не мультиколлинеарность, а эти, гораздо более критичные, проблемы.

Гетероскедастичность случайных ошибок или коррелированность случайных ошибок, относящихся к разным наблюдениям

В условиях гетероскедастичности МНК-оценки коэффициентов остаются несмещенными и состоятельными (см. детали в главе 5). Поэтому, как и мультиколлинеарность, эта проблема является менее критичной, чем проблема эндогенности.

Основная рекомендация в связи с возможной гетероскедастичностью — не забывайте использовать состоятельные в условиях гетероскедастичности (робастные) стандартные ошибки. Иначе результаты тестирования гипотез и построенные вами доверительные интервалы будут некорректными (даже при том, что сами коэффициенты модели будут оценены состоятельно).

В некоторых моделях случайные ошибки, относящиеся к разным наблюдениям, могут быть коррелированы друг с другом. Одним из примеров такой ситуации является автокорреляция, рассмотренная в конце параграфа 5.5. Другой типичный пример — модели на панельных данных. Он будет проанализирован в главе 9. Там будет показано, что в этом случае для решения проблемы также достаточно использовать специальный тип робастных стандартных ошибок.

Неоднородность выборки

Неоднородностью выборки будем называть ситуацию, когда часть наблюдений существенным образом отличается от основного массива данных. Подобные резко отличающиеся наблюдения называют выбросами. Как правило, следует исключать выбросы из выборки.

Представим, например, что вы исследуете влияние неравенства на экономический рост, используя межстрановые данные. Все страны в вашей выборке являются развитыми европейскими экономиками за единственным исключением: одна страна оказалась развивающейся экономикой из Южной Америки. Например, Венесуэлой. Не имея ничего против Венесуэлы, отметим, что, по всей видимости, её институты, уровень экономического развития и многие другие характеристики довольно сильно отличаются от аналогичных характеристик типичной развитой европейской страны. Поэтому, наверняка, и модель экономического роста для такой страны будет другой. Чтобы в полной мере учесть это отличие, пришлось бы добавить неоправданно много контрольных переменных. Поэтому лучше просто исключить Венесуэлу из выборки.

Другая похожая ситуация — исследования на региональных данных по России. В таких работах из выборки часто исключают, например, Москву, так как она резко отличается от других регионов сразу по ряду характеристик: плотность и численность населения, размер валового регионального продукта, средняя заработная плата и так далее.

Примеры неоднородности данных легко привести и на микроуровне. Представим исследователя, анализирующего влияние физических характеристик баскетболиста на его доходы в профессиональном спорте. Если массив данных, на который опирается исследователь, состоит из информации о двух сотнях спортсменов, играющих в российских баскетбольных турнирах, и о его любимом Майкле Джордане¹, то информацию о последнем, увы, придется из выборки исключить.

Формально, наличие существенных выбросов нарушает предпосылку №2 линейной модели со стохастическими регрессорами, что может приводить к проблемам с состоятельностью и асимптотической нормальностью оценок. Однако обычно целесообразность исключения нетипичных наблюдений понятна и просто из соображений здравого смысла. Вряд ли статистика великого Майкла Джордана, игравшего в НБА в прошлом веке, может рассказать какую-то полезную историю о заработках типичного российского баскетболиста в наши дни. Скорее уж она непоправимо исказит результаты.

Обычно потенциальная неоднородность данных может быть обнаружена на этапе их первичного анализа. В частности, анализа гистограмм распределения данных и диаграмм рассеяния. Кроме того, для выявления отдельных неоднородных наблюдений или целых подвыборок полезно хорошо понимать содержательную сторону исследуемого вопроса.

Угрозы внешней обоснованности выводов

Приведенный выше пример про баскетболистов может быть полезен, чтобы обсудить ещё один важный вопрос: корректное определение границ, в рамках которых можно обобщать выводы эконометрического исследования.

В соответствии с определением (Stock, Watson, 2010) будем называть выводы эконометрического моделирования по поводу причинно-следственных связей **внутренне обоснованными**, если они применимы к проанализированной в исследовании генеральной совокупности.

Соблюдение рекомендаций данной главы позволяет гарантировать внутреннюю обоснованность за счет обеспечения состоятельности оценок коэффициентов и корректного определения их стандартных ошибок. При этом, однако, важно четко определять рамки анализируемой генеральной совокупности. Скажем, если (как в примере выше) исследование осуществлено на основе информации о двух сотнях российских баскетболистов, то анализируемой генеральной совокупностью можно считать всех (именно) российских (именно) баскетболистов. Это значит, что полученные выводы не следует распространять на всех баскетболистов мира или на представителей других видов спорта.

Более того, если в процессе более пристального рассмотрения данных окажется, что все баскетболисты в выборке были мужчинами в возрасте от 20 до 25 лет, то границы анализируемой генеральной совокупности должны быть сужены до этой категории спортсменов и не могут быть распространены на игроков старше сорока или женщин-баскетболисток. Пусть даже они из России.

Аналогично, на основе исследования по данным развитых европейских экономик можно делать выводы относительно развитых стран из Европы, но не по поводу, например, африканских стран. Во втором случае они будут необоснованными из-за игнорирования важных институциональных различий².

Наконец, выводы не могут считаться обоснованными при слишком малом размере выборки. Надежными можно считать только результаты, опирающиеся на достаточно большие массивы данных. В силу свойств оценок эконометрических моделей, описанных в главах 3 и 6, не стоит всерьез полагаться на множественные регрессии, оцененные по 10–20 точкам. Даже если формальные характеристики этих уравнений выглядят убедительно.

В некоторых случаях эконометрист может претендовать на то, что выводы его исследования являются не только внутренне, но и внешне обоснованными. Выводы эконометрического исследования называются **внешне обоснованными**, если их можно обобщить с исследованной генеральной совокупности и заданных условий на другие генеральные совокупности и условия. Например, если вы используете данные обследования студентов Новосибирского государственного университета, чтобы формулировать те или иные суждения по поводу склонности к списыванию у студентов Санкт-Петербургского государственного университета.

Для обеспечения внешней обоснованности необходимо дополнительно к требованиям обоснованности внутренней гарантировать выполнение еще двух условий:

- **Отсутствие различий в генеральных совокупностях.** В нашем примере это означает, что студенты НГУ и СПбГУ должны быть похожи друг на друга по своим основным характеристикам.
- **Отсутствие различий в условиях.** Даже если студенты НГУ и СПбГУ полностью идентичны, они могут находиться в разных институциональных условиях. Это может делать выводы, полученные на основе анализа одной группы студентов, неприменимыми по отношению к другой группе. Например, если речь идет о склонности к списыванию, то выводы могут быть искажены различием уровней строгости наказаний за нарушения академической этики, принятых в разных университетах.

Пример 7.1. Стоимость колготок в Московских оптовых торговых фирмах.

Массив данных о стоимости продукции двух фирм-производителей колготок осенью 1997 года вы можете найти в файле Tights.xlsx. Советуем проделать все вычисления, описанные в решении этого примера, чтобы лучше разобраться в деталях.

Переменные:

PRICE — цена колготок в рублях 1997 г.,

DEN — плотность колготок,

POLYAMID — % содержания полиамида,

LYKRA — % содержания лайкры,

COTTON — % содержания хлопка,

WOOL — % содержания шерсти,

FIRM = $\begin{cases} 0, & \text{если производитель колготок — фирма Levante,} \\ 1, & \text{если производитель колготок — фирма Golden Lady.} \end{cases}$

Вас интересует ответ на следующий исследовательский вопрос: *если рассматривать продукцию с одинаковыми характеристиками, будет ли различаться цена для двух фирм? Если да, то какая фирма устанавливает более высокие цены?*

(а) На основании анализа средних значений по группам скажите, различаются ли цены на колготки у разных производителей? Каковы недостатки такого подхода?

Решение пункта (а):

Ограничив выборку условием FIRM=1, получаем подвыборку колготок, произведенных фирмой GoldenLady. Среднее значение цены для этой подвыборки составляет 15206 рублей.

Ограничив выборку условием FIRM=0, получаем подвыборку колготок, произведенных фирмой Levante. Среднее значение цены для этой подвыборки составляет 16382 рубля.

Получается, что колготки Levante дороже в среднем примерно на 1000 рублей. Недостаток такого подхода в том, что он мало что говорит по поводу готовности потребителей доплачивать именно за товар данной фирмы-производителя. Возможно, дело не в бренде, а, например, в составе колготок. Иными словами, простое сравнение средних не позволяет учесть прочие важные факторы.

(б) Постройте уравнение зависимости цены колготок от их плотности, состава и производителя. Не забудьте использовать состоятельные в условиях гетероскедастичности стандартные ошибки.

Решение пункта (б):

Модель 1: МНК, использованы наблюдения 1-74

Зависимая переменная: Price

	Коэффициент	Ст. ошибка	t-статистика	P-значение	
const	-4644,35	59396,1	-0,07819	0,9379	
DEN	176,448	39,6128	4,454	<0,0001	***
polyamid	103,653	597,609	0,1734	0,8628	
lykra	391,328	544,700	0,7184	0,4750	
cotton	156,537	584,205	0,2679	0,7896	
wool	476,064	1150,73	0,4137	0,6804	
Сумма кв. остатков	2,16e+09		Ст. ошибка модели	5641,841	
R-квадрат	0,490681		Испр. R-квадрат	0,453231	
F(5, 68)	13,10230		P-значение (F)	6,17e-09	

(в) Можно ли считать, что уравнения, описывающие цены колготок для двух рассматриваемых фирм, отличаются друг от друга? Проведите тест Чоу.

Решение пункта (в)

Для осуществления теста Чоу нужно оценить новую модель, добавив в неё соответствующую переменную сдвига и все необходимые переменные наклона, как это показано в таблице ниже:

Расширенная регрессия для теста Чоу
МНК, использованы наблюдения 1-74
Зависимая переменная: Price
Пропущены из-за совершенной коллинеарности: fi_wool

	Коэффициент	Ст. ошибка	t-статистика	P-значение	
const	-152808	118428	-1,290	0,2017	
DEN	221,721	51,9178	4,271	6,70e-05	***
polyamid	1554,64	1178,86	1,319	0,1920	
lykra	1922,61	1212,67	1,585	0,1179	
cotton	1595,87	1173,40	1,360	0,1787	
wool	3333,75	2346,56	1,421	0,1603	
firm	-296013	168911	-1,752	0,0846	*
firm*DEN	-96,1882	81,2445	-1,184	0,2409	
firm*polyamid	2992,42	1679,25	1,782	0,0796	*
firm*lykra	3114,42	1798,28	1,732	0,0882	*
firm*cotton	3011,39	1698,10	1,773	0,0810	*
Среднее зав. перемен	15841,89	Ст. откл. зав. перемен	7629,898		
Сумма кв. остатков	1,98e+09	Ст. ошибка модели	5612,979		
R-квадрат	0,532946	Испр. R-квадрат	0,458810		
F(10, 63)	7,188806	P-значение (F)	1,93e-07		
Лог. правдоподобие	-737,8770	Крит. Акаике	1497,754		
Крит. Шварца	1523,099	Крит. Хеннана-Куинна	1507,864		

Тест Чоу для структурных изменений в точке firm

F(5, 63) = 1,14022 p-значение 0,349

Примечание: используется F-статистика, состоятельная в условиях гетероскедастичности

Из таблицы видно, что P-значение для теста Чоу составляет 0,349, что больше, чем 0,05. Таким образом даже при использовании пятипроцентного уровня значимости мы не отвергаем гипотезу об отсутствии структурного сдвига между моделями для цены колготок двух разных производителей.

Есть некоторый соблазн остановить на этом процесс нашего исследования. Казалось бы, мы получили ответ на сформулированный в начале задания исследовательский вопрос: бренд не имеет значения. Однако, прежде, чем это сделать, следует более обстоятельно проанализировать качество полученной модели.

(г) Как можно объяснить большое количество незначимых переменных в уравнении из пункта (б)? Может быть, существует мультиколлинеарность? Чему равно значение коэффициентов VIF для модели из предыдущего пункта?

Как правило, колготки делают только из этих 4-х материалов: хлопка, шерсти, лайкры и полиамида (поэтому обычно сумма долей этих 4 элементов должна составлять 100%). Для проверки этого предположения создадим новую переменную, представляющую собой остаток:

$$\text{REST} = 100 - \text{COTTON} - \text{LYKRA} - \text{POLYAMID} - \text{WOOL}$$

Для каких наблюдений эта переменная не равна нулю? Исключите их из выборки.

Оцените регрессию заново, исключив из неё переменную POLYAMID, чтобы устранить мультиколлинеарность.

Решение пункта (г):

Из результатов оценивания модели в пункте (б) видим, что все переменные, кроме переменной DEN, незначимы. Полученный результат можно объяснить мультиколлинеарностью. Действительно, в большинстве случаев колготки состоят из четырех материалов, учтенных в модели: COTTON, LYKRA, POLYAMID и WOOL. Следовательно, обычно сумма долей четырех этих составляющих в составе колготок будет близка к 100%, то есть между указанными переменными существует почти строгая линейная связь.

Для проверки нашего предположения вычислим значения коэффициентов VIF:

DEN	1,746
polyamid	251,545
lykra	89,774
cotton	93,498
wool	57,581

Легко видеть, что для перечисленных выше четырех переменных коэффициенты VIF существенно больше 10, что свидетельствует о сильной мультиколлинеарности.

Для проверки нашего предположения о том, что практически всегда колготки на 100% состоят из хлопка, лайкры, полиамида и шерсти, создадим новую переменную, представляющую из себя остаток:

$$\text{REST} = 100 - \text{COTTON} - \text{LYKRA} - \text{POLYAMID} - \text{WOOL}$$

Проанализировав значения этой переменной в выборке, выясняем, что лишь для двух наблюдений переменная $\text{REST} \neq 0$. Исключим из выборки эти нетипичные наблюдения.

Далее, рассмотрев значения переменной POLYAMID, видим, что для всех наблюдений его значение велико и близко к 100%. Исключив его из регрессии, мы по существу будем воспринимать его как материал «по умолчанию». Это позволит в значительной мере решить проблему мультиколлинеарности. Оценим на новой выборке регрессию с переменными COTTON, DEN, LYKRA и WOOL.

Модель 2: МНК, использованы наблюдения 1-72

Зависимая переменная: Price

Робастные оценки стандартных ошибок (с поправкой на гетероскедастичность), вариант HC1

	Коэффициент	Ст. ошибка	t-статистика	P-значение	
const	4850,27	1800,05	2,695	0,0089	***
DEN	175,399	44,4396	3,947	0,0002	***
lykra	367,032	92,4580	3,970	0,0002	***
cotton	58,2704	61,7878	0,9431	0,3490	
wool	-249,251	554,391	-0,4496	0,6545	
Среднее зав. перемен	15561,11		Ст. откл. зав. перемен	7307,194	
Сумма кв. остатков	2,10e+09		Ст. ошибка модели	5599,897	
R-квадрат	0,445789		Испр. R-квадрат	0,412702	
F(4, 67)	25,40284		P-значение (F)	7,92e-13	
Лог. правдоподобие	-720,9688		Крит. Акаике	1451,938	
Крит. Шварца	1463,321		Крит. Хеннана-Куинна	1456,469	

Вычислив коэффициенты VIF для новой модели, видим, что мультиколлинеарность действительно устранена (так как все коэффициенты заметно меньше 10):

DEN	1,496
lykra	3,109
cotton	1,430
wool	3,147

(д) Можно ли что-то еще сделать для получения более однородных данных? Посмотрим на данные подробнее.

Во-первых, обратимся к переменной WOOL: по шерсти только два наблюдения не равны 0. Скорее всего, это другой вид колготок, который описывается другой моделью, поэтому исключим их.

Во-вторых, проанализируем дополнительно данные по переменной COTTON. Только 3 наблюдения принимают очень высокие значения. Скорее всего, это совершенно другой вид колготок. Исключим из выборки все наблюдения, для которых содержание хлопка составляет 40% или выше.

Снова оцените модель из предыдущего пункта, используя новую ограниченную выборку. То есть оцените регрессию переменной PRICE на константу и переменные DEN, LYKRA, COTTON (переменную WOOL больше не имеет смысла включать в уравнение, так как теперь она для всех наблюдений равна нулю).

Как устранение мультиколлинеарности и переход к более однородным данным сказались на значимости оценок коэффициентов в данном примере?

Решение пункта (д):

Исключение переменных по указанным в данном пункте критериям приводит к выборке из 67 наблюдений. Оценим параметры регрессионной модели по этим данным.

Модель 3: МНК, использованы наблюдения 1-67

Зависимая переменная: Price

Робастные оценки стандартных ошибок (с поправкой на гетероскедастичность), вариант HC1

	Коэффициент	Ст. ошибка	t-статистика	P-значение	
const	5128,89	1596,22	3,213	0,0021	***
DEN	160,371	43,6176	3,677	0,0005	***
lykra	256,324	73,8774	3,470	0,0009	***
cotton	2085,87	529,272	3,941	0,0002	***
Сумма кв. остатков	1,66e+09		Ст. ошибка модели	5134,673	
R-квадрат	0,472643		Испр. R-квадрат	0,447531	
F(3, 63)	17,17920		P-значение (F)	2,91e-08	

Отметим, что устранение мультиколлинеарности и использование однородной выборки приводит к получению статистически значимых коэффициентов при всех переменных в модели.

(е) Вернемся к интересующему нас вопросу о важности фирмы-производителя. Для модели из предыдущего пункта проведите тест Чоу, чтобы получить ответ на этот вопрос. Используйте пятипроцентный уровень значимости.

Если выяснилось, что структурный сдвиг между моделями ценообразования разных фирм существует, то оцените модель из пункта (д) заново, добавив в неё переменную FIRM. То есть оцените регрессию переменной PRICE на константу и переменные DEN, LYKRA, COTTON, FIRM. Интерпретируйте полученный результат.

Решение пункта (е):

Результаты теста Чоу для модели из пункта (д) представлены ниже. Обратите внимание, что соответствующее P-значение равно 0,014, то есть меньше, чем 0,05. Следовательно, при уровне значимости 5% мы можем заключить, что структурный сдвиг существует и добавление в модель переменной, характеризующей фирму производителя, скорее всего, является оправданным.

Расширенная регрессия для теста Чоу

МНК, использованы наблюдения 1-67

Зависимая переменная: Price

Робастные оценки стандартных ошибок (с поправкой на гетероскедастичность), вариант HC1

	Коэффициент	Ст. ошибка	t-статистика	P-значение	
const	3084,33	1738,94	1,774	0,0813	*
DEN	203,996	71,8443	2,839	0,0062	***
lykra	228,206	114,550	1,992	0,0510	*
cotton	2244,67	693,342	3,237	0,0020	***
firm	1950,90	2892,84	0,6744	0,5027	
firm*DEN	-80,0647	81,1609	-0,9865	0,3279	
firm*lykra	243,035	153,016	1,588	0,1176	
firm*cotton	-27,6460	1129,56	-0,02448	0,9806	
Среднее зав. перемен	14856,72	Ст. откл. зав. перемен	6908,103		
Сумма кв. остатков	1,50e+09	Ст. ошибка модели	5046,583		
R-квадрат	0,522927	Испр. R-квадрат	0,466325		
F(7, 59)	12,99036	P-значение (F)	5,58e-10		
Лог. правдоподобие	-662,0824	Крит. Акаике	1340,165		
Крит. Шварца	1357,802	Крит. Хеннана-Куинна	1347,144		

Тест Чоу для структурных изменений в точке firm

Хи-квадрат(4) = 13,7517 p-значение 0,0081

F-статистика: F(4, 59) = 3,43792 p-значение 0,014

Ниже представлены результаты оценивания с включением новой переменной. Легко видеть, что коэффициент при переменной FIRM статистически значим при пятипроцентном уровне и составляет около двух с половиной тысяч. Следовательно, при прочих равных условиях (то есть при одинаковой плотности и сходном составе) колготки фирмы Golden Lady стоят на две с половиной тысячи рублей больше, чем колготки фирмы Levante³.

Модель 4: МНК, использованы наблюдения 1-67

Зависимая переменная: Price

Робастные оценки стандартных ошибок (с поправкой на гетероскедастичность), вариант HC1

	Коэффициент	Ст. ошибка	t-статистика	P-значение	
const	2728,33	1396,87	1,953	0,0553	*
DEN	166,223	43,0672	3,860	0,0003	***
lykra	329,380	75,4975	4,363	<0,0001	***
cotton	2203,03	543,331	4,055	0,0001	***
firm	2566,69	1266,24	2,027	0,0470	**
Сумма кв. остатков		1,57e+09	Ст. ошибка модели	5036,107	
R-квадрат		0,500748	Испр. R-квадрат	0,468538	
F(4, 62)		23,97258	P-значение (F)	5,16e-12	

(ж) Измените функциональную форму модели на логарифмически-линейную. То есть оцените регрессию **логарифма** переменной PRICE на константу и переменные DEN, LYKRA, COTTON, FIRM. Интерпретируйте полученный результат.

Решение пункта (ж)

Результаты оценивания логарифмически-линейной модели представлены ниже:

Модель 5: МНК, использованы наблюдения 1-67

Зависимая переменная: l_Price

Робастные оценки стандартных ошибок (с поправкой на гетероскедастичность), вариант HC1

	Коэффициент	Ст. ошибка	t-статистика	P-значение	
const	8,44591	0,123112	68,60	<0,0001	***
DEN	0,0115825	0,00241486	4,796	<0,0001	***
lykra	0,0352771	0,00608287	5,799	<0,0001	***
cotton	0,148974	0,0284931	5,228	<0,0001	***
firm	0,286642	0,0766869	3,738	0,0004	***
Сумма кв. остатков		6,106496	Ст. ошибка модели	0,313834	
R-квадрат		0,608580	Испр. R-квадрат	0,583327	
F(4, 62)		29,38280	P-значение (F)	1,03e-13	

Легко видеть, что и в данном случае переменная, характеризующая фирму-производителя, статистически значима. Интерпретировать коэффициент при этой переменной можно так: при прочих равных условиях (то есть при одинаковой плотности и сходном составе) колготки фирмы Golden Lady стоят на $(e^{0,29} - 1) * 100\% = 33\%$ больше, чем колготки фирмы Levante.

(з) Осуществите тесты Рамсея для моделей из пунктов (е) и (ж). Используйте пятипроцентный уровень значимости. В соответствии с результатами тестов сделайте вывод, какая функциональная форма связи в данном случае является более оправданной: линейная или логарифмически-линейная?

Решение пункта (з)

Для модели из пункта (е) Р-значения теста Рамсея составляет 0,46 (то есть больше, чем 0,05), а для модели из пункта (ж) оно равно 0,015 (то есть меньше, чем 0,05).

Следовательно, в первом случае гипотеза о корректности спецификации не отвергается, а во втором — отвергается. Можно заключить, что более оправдано использование линейной спецификации модели. Впрочем, это не очень важно, так как выводы по поводу знака и значимости коэффициента при интересующей нас переменной совпадают.

(и) На основе анализа всех полученных результатов дайте ответ на исследовательский вопрос, сформулированный в самом начале этого задания.

Решение пункта (и):

Мы получили устойчивый к изменению спецификации модели вывод о том, что при одинаковой плотности и сходном составе колготки фирмы Golden Lady стоят дороже колготок её конкурента.

Примечание: обратите внимание, что в процессе решения задания для получения корректного ответа на интересующий нас вопрос, пришлось проделать долгий путь. На этом пути мы использовали комбинацию нескольких идей по улучшению качества модели, описанных в данной главе: получение однородной выборки за счет устранения выбросов; преодоление негативных последствий сильной мультиколлинеарности; сопоставление различных функциональных форм уравнения регрессии.

-
1. Майкл Джордан — шестикратный чемпион национальной баскетбольной ассоциации, который по версии ряда источников является величайшим баскетболистом всех времён и народов. В контексте нашего примера стоит также отметить, что Forbes называет Майкла Джордана самым высокооплачиваемым спортсменом в мире за всю историю профессионального спорта. [↪](#)
 2. Учитывая очевидность этой рекомендации, кажется удивительным, насколько часто она игнорируется не только начинающими эконометристами, но и вполне респектабельными людьми, дающими рекомендации по выбору мер экономической политики. Примеры провалов реформ, которые проводились в развивающихся странах на основе моделей, разработанных для развитых экономик, приводятся в остроумной книге Уильяма Истерли «В поисках роста». Русскоязычное издание: Истерли В. В поисках роста: Приключения и злоключения экономистов в тропиках / Пер. с англ. — М.: Институт комплексных стратегических исследований, 2006. [↪](#)
 3. Точнее, стоили в 1997 году. [↪](#)

7.6. Чек-лист эконометриста

Подводя итог главе, сформулируем следующий краткий список вопросов, на который вам следует отвечать каждый раз, когда вы хотите решить, заслуживает ли ваша эконометрическая модель доверия:

1. Нет ли в модели пропущенных существенных переменных?
2. Верна ли выбранная функциональная форма?
3. Нет ли в модели эндогенности в результате двусторонней причинно-следственной связи?
4. Не искажены ли выводы из-за сильных ошибок измерения регрессора?
5. Не забыли ли вы использовать состоятельные в данных условиях стандартные ошибки?
6. Если переменная интереса оказалась незначимой, не вызвана ли эта незначимость сильной мультиколлинеарностью?
7. Верно ли определены границы генеральной совокупности, на которую вы обобщаете выводы своей модели?

Постоянное размышление над этими вопросами не гарантирует получения «идеальной модели» во всех случаях (увы, не существует набора простых рецептов, который бы его гарантировал), однако во многих случаях защитит вас от заведомо некорректных результатов и выводов.

Глава 8. Инструментальные переменные

Инструментальные переменные (инструменты) — это мощное средство для решения проблемы эндогенности. Как мы уже упоминали в прошлой главе, они могут быть полезны в следующих случаях:

- смещение оценки коэффициента из-за пропуска ненаблюдаемой существенной переменной (то есть в ситуации, когда это смещение не может быть устранено за счет включения в модель контрольных переменных);
- ошибки измерения регрессора;
- эндогенность из-за двусторонней причинно-следственной связи.

Как видите, спектр ситуаций, в которых инструментальные переменные полезны, весьма широк. Поэтому целесообразно разобраться с новым для нас специальным методом оценивания, опирающимся на их использование, — двухшаговым методом наименьших квадратов (two-stage least squares). Для его обозначения мы будем использовать аббревиатуру 2МНК (а по-английски — 2SLS или TSLS).

В первом параграфе данной главы мы сформулируем все определения и поясним все ключевые идеи 2МНК для самого простого случая: парной регрессии. Во втором параграфе мы обобщим наш анализ на случай множественной регрессии. В третьем (техническом) параграфе с использованием матричной алгебры будут выведены формулы 2МНК-оценок коэффициентов. Четвертый параграф содержит обсуждение статистических тестов, которые помогают определить, насколько хороши ваши инструменты и насколько уместно их использование в том или ином случае. Наконец, в пятом параграфе (особенно важном с прикладной точки зрения) мы сконцентрируемся на вопросе о том, откуда можно взять подходящие для вашего исследования инструментальные переменные и приведем несколько примеров реальных исследований, использующих этот подход.

[8.1. Двухшаговый МНК: парная регрессия](#)

[8.2. Двухшаговый МНК: множественная регрессия](#)

[8.3. Векторно-матричная форма записи](#)

[8.4. Тесты для моделей, оцененных двухшаговым МНК](#)

[8.5. Где взять подходящие инструменты?](#)

[Задания для самостоятельного решения](#)

8.1. Двухшаговый МНК: парная регрессия

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

Рассмотрим модель парной регрессии $y_i = \beta_1 + \beta_2 * x_i + \varepsilon_i$, в которой x — эндогенный регрессор: $\text{cov}(x_i, \varepsilon_i) \neq 0$. Иными словами, нарушена предпосылка №4 линейной модели с одним регрессором из главы 6. Будем предполагать, что все остальные предпосылки этой модели выполняются.

Как мы выяснили в двух предыдущих главах, если предпосылка об экзогенности регрессора нарушена, то МНК-оценка коэффициента β_2 несостоятельна. Следовательно, обычный МНК использовать нецелесообразно. Поэтому нам нужно предложить другой метод.

Представим, что в нашем распоряжении кроме информации о переменных x и y есть ещё данные о третьей переменной z , которая удовлетворяет двум свойствам:

- релевантность (relevance): $\text{cov}(z_i, x_i) \neq 0$,
- экзогенность (exogeneity): $\text{cov}(z_i, \varepsilon_i) = 0$.

Первое свойство говорит о том, что эта переменная коррелирована с нашим эндогенным регрессором. Второе свойство требует, чтобы эта переменная не была коррелирована со случайной ошибкой модели (то есть, в отличие от регрессора x , являлась экзогенной).

Переменная z называется инструментальной переменной или инструментом (instrumental variable, instrument). Если инструмент удовлетворяет обоим указанным свойствам (экзогенности и релевантности), то его называют валидным.

Графически отношения между переменными x , y и z представлены на рисунке 8.1.

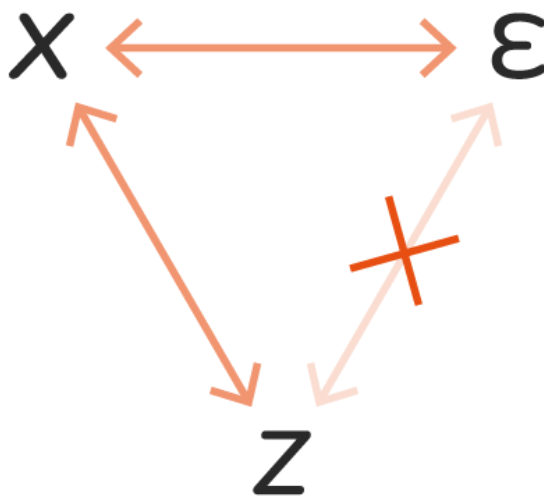


Рисунок 8.1. Взаимосвязи между переменными x , y и z . Не зачёркнутые стрелки означают наличие корреляции. Зачеркнутая стрелка означает её отсутствие.

Наличие валидного инструмента позволяет получить состоятельную оценку коэффициента β_2 , используя двухшаговый МНК. Из названия легко догадаться, что реализация этого метода состоит из двух шагов. На каждом из них применяется обычный МНК:

Первый шаг

Оцениваем регрессию переменной x по переменной z :

$$\hat{x}_i = \hat{\theta}_1 + \hat{\theta}_2 z_i.$$

Получаем прогнозные значения $\hat{x}_i$.

Второй шаг

Оцениваем регрессию переменной y по этим предсказанным значениям $\hat{x}$:

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 \hat{x}_i.$$

Чтобы отличать оценку коэффициента, полученную таким способом, от обычной МНК-оценки, мы иногда будем использовать в обозначении дополнительный индекс: $\widehat{\beta}_2^{TSLs}$.

Прежде чем формально доказывать, что оценка $\widehat{\beta}_2^{TSLs}$ является состоятельной, поясним кратко, почему так получается. Всё дело в том, что переменная $\widehat{x}$, в отличие от исходной переменной x , является экзогенной. Мы можем быть в этом уверены потому, что она линейно выражается через инструмент, а инструмент, в свою очередь, экзогенен по своему свойству.

Покажем, что 2МНК-оценка коэффициента при регрессоре задается формулой:

$$\widehat{\beta}_2^{TSLs} = \frac{\frac{1}{n} \sum (y_i - \bar{y})(z_i - \bar{z})}{\frac{1}{n} \sum (\widehat{x}_i - \bar{\widehat{x}})(z_i - \bar{z})} = \frac{\widehat{cov}(y, z)}{\widehat{cov}(\widehat{x}, z)}$$

Для этого отметим, что в силу стандартных формул для парной регрессии МНК-оценка коэффициента при переменной в регрессии первого шага $\widehat{x}_i = \hat{\theta}_1 + \hat{\theta}_2 z_i$ имеет вид: $\hat{\theta}_2 = \frac{\widehat{cov}(x, z)}{\widehat{var}(z)}$, а МНК-оценка в регрессии второго шага $\widehat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 \widehat{x}_i$ равна

$$\widehat{\beta}_2^{TSLs} = \frac{\widehat{cov}(\widehat{x}, y)}{\widehat{var}(\widehat{x})}. \text{ Следовательно}$$

$$\widehat{\beta}_2^{TSLs} = \frac{\widehat{cov}(\widehat{x}, y)}{\widehat{var}(\widehat{x})} = \frac{\widehat{cov}(\hat{\theta}_1 + \hat{\theta}_2 z, y)}{\widehat{var}(\hat{\theta}_1 + \hat{\theta}_2 z)} = \frac{\hat{\theta}_2 * \widehat{cov}(z, y)}{(\hat{\theta}_2)^2 * \widehat{var}(z)} =$$

$$\frac{\widehat{cov}(z, y)}{\hat{\theta}_2 * \widehat{var}(z)} = \frac{\widehat{cov}(z, y)}{\frac{\widehat{cov}(x, z)}{\widehat{var}(z)} * \widehat{var}(z)} = \frac{\widehat{cov}(y, z)}{\widehat{cov}(x, z)}$$

Теперь можно доказать состоятельность этой оценки.

$$\widehat{\beta}_2^{TSLs} = \frac{\frac{1}{n} \sum (y_i - \bar{y})(z_i - \bar{z})}{\frac{1}{n} \sum (\widehat{x}_i - \bar{\widehat{x}})(z_i - \bar{z})} \xrightarrow{P} \frac{Cov(y_i, z_i)}{Cov(x_i, z_i)} =$$

$$\frac{Cov(\beta_1 + \beta_2 x_i + \varepsilon_i, z_i)}{Cov(x_i, z_i)} = \frac{\beta_2 Cov(x_i, z_i) + Cov(\varepsilon_i, z_i)}{Cov(x_i, z_i)} =$$

$$\{Cov(z_i, \varepsilon_i) = 0\} =$$

$$\frac{\beta_2 Cov(x_i, z_i) + 0}{Cov(x_i, z_i)} = \beta_2$$

Здесь мы также использовали свойство релевантности инструмента (то есть свойство $Cov(z_i, x_i) \neq 0$), чтобы гарантировать, что знаменатель дроби $\frac{Cov(y_i, z_i)}{Cov(x_i, z_i)}$ не равен нулю.

В случае использования двухшагового МНК дисперсии оценок коэффициентов будут отличаться от случая обычного МНК. Следовательно, и стандартные ошибки оценок коэффициентов нужно рассчитывать несколько иначе. Например, для случая гомоскедастичности корректная формула стандартной ошибки коэффициента при регрессоре имеет вид:

$$se(\widehat{\beta}_2) = \sqrt{\frac{s^2}{\widehat{cov}^2(x, z)} * 1}, S^2 = \frac{\sum_{i=1}^n e_i^2}{n-2}.$$

Подчеркнем, что здесь остатки рассчитываются по формуле $e_i = y_i - \widehat{\beta}_1 - \widehat{\beta}_2 x_i$, а не по формуле $e_i = y_i - \widehat{\beta}_1 - \widehat{\beta}_2 \widehat{x}_i$ (так как иначе S^2 окажется несостоятельной оценкой дисперсии случайной ошибки).

Если вы сравните эту формулу стандартной ошибки с привычной формулой из второй главы, то увидите, что отличие состоит в появлении множителя $\frac{1}{\widehat{cov}^2(x, z)}$. Этот множитель демонстрирует важность релевантности инструмента. Ведь если инструмент почти не релевантен, то корреляция между ним и регрессором окажется близкой к нулю. Тогда этот множитель будет велик, и, следовательно, стандартная ошибка оценки коэффициента тоже будет велика, что приведет к низкой точности оценивания. Например, к очень широким доверительным интервалам.

Чтобы эконометрический пакет рассчитывал стандартные ошибки коэффициентов корректно, следует использовать встроенную процедуру двухшагового МНК (в этом случае можно добавить и корректировку стандартных ошибок на случай гетероскедастичности), а не самостоятельно регрессировать y по $\widehat{x}$ при помощи обычного МНК.

За исключением альтернативных стандартных ошибок, процедуры тестирования гипотез и построения доверительных интервалов для коэффициентов модели полностью аналогичны процедурам, описанным в главе 6.

Пример 8.1. Оценка эластичности спроса по цене

На рынке сигарет в некоторой стране функция спроса в i -м регионе имеет вид:

$$\ln Q_i = \beta_1 + \beta_2 * \ln P_i + \varepsilon_i$$

Функция предложения описывается соотношением:

$$\ln Q_i = \gamma_1 + \gamma_2 * \ln P_i + \gamma_3 * \ln T_i + u_i$$

Q_i — количество сигарет в i -м регионе, P_i — цена сигарет в i -м регионе, T_i — налог с продаж в i -м регионе. ε_i — независимые и одинаково распределенные случайные величины, характеризующие шоки спроса (не коррелированы с налогами), u_i — независимые и одинаково распределенные случайные величины, характеризующие шоки предложения.

(а) Объясните, почему МНК-оценка эластичности спроса по цене в рассматриваемой модели будет несостоятельной. Для этого вычислите предел по вероятности для МНК-оценки $\widehat{\beta}_2^{OLS}$. Определите, если это возможно, будет ли МНК давать завышенную или заниженную оценку эластичности спроса?

(6) Предложите процедуру состоятельного оценивания эластичности спроса по цене. Формально обоснуйте свой ответ, вычислив соответствующий предел по вероятности.

Решение:

(а) Выразим эндогенную переменную $\ln P_i$ через экзогенную $\ln T_i$:

$$\beta_1 + \beta_2 \ln P_i + \varepsilon_i = \gamma_1 + \gamma_2 * \ln P_i + \gamma_3 * \ln T_i + u_i$$

$$\ln P_i = \frac{\beta_1 - \gamma_1 - \gamma_3 * \ln T_i + \varepsilon_i - u_i}{\gamma_2 - \beta_2}.$$

$$\text{cov}(\ln P_i, \varepsilon_i) = \text{cov}\left(\frac{\beta_1 - \gamma_1 - \gamma_3 * \ln T_i + \varepsilon_i - u_i}{\gamma_2 - \beta_2}, \varepsilon_i\right) =$$

$$\frac{1}{\gamma_2 - \beta_2} * \text{cov}(\beta_1 - \gamma_1 - \gamma_3 * \ln T_i + \varepsilon_i - u_i, \varepsilon_i) =$$

$$\frac{1}{\gamma_2 - \beta_2} * \text{cov}(-\gamma_3 * \ln T_i + \varepsilon_i - u_i, \varepsilon_i) =$$

$$\frac{1}{\gamma_2 - \beta_2} (-\gamma_3 * \text{cov}(\ln T_i, \varepsilon_i) + \sigma_\varepsilon^2 - \text{cov}(u_i, \varepsilon_i)) =$$

$$\left\{ \begin{array}{l} \text{cov}(\ln T_i, \varepsilon_i) = 0, \text{налоги не коррелированы с шоками спроса} \\ \text{cov}(u_i, \varepsilon_i) = 0 \end{array} \right\} =$$

$$\frac{\sigma_\varepsilon^2}{\gamma_2 - \beta_2}.$$

Таким образом, цена является эндогенной переменной в уравнении спроса, так как она коррелирована со случайными ошибками в этом уравнении. Такой ситуации легко предложить интуитивное объяснение, если представить стандартный график с пересечением кривых спроса и предложения: положительный шок спроса ($\varepsilon > 0$) сдвигает кривую спроса вправо, в результате чего растет равновесная цена.

Поэтому оценка обычного МНК будет несостоятельной:

$$\hat{\beta}_2^{OLS} = \frac{\widehat{\text{cov}}(\ln P, \ln Q)}{\widehat{\text{var}}(\ln P)} = \beta_2 + \frac{\widehat{\text{cov}}(\ln P, \varepsilon)}{\widehat{\text{var}}(\ln P)} \xrightarrow{P}$$

$$\xrightarrow{P} \beta_2 + \frac{\text{cov}(\ln P, \varepsilon_i)}{\text{var}(\ln P_i)} = \beta_2 + \frac{\frac{\sigma_\varepsilon^2}{\gamma_2 - \beta_2}}{\sigma_{\ln P}^2} \neq \beta_2.$$

Поскольку по закону спроса $\beta_2 < 0$, а в силу закона предложения $\gamma_2 > 0$, то $(\gamma_2 - \beta_2) > 0$. Следовательно, $\frac{\sigma_\varepsilon^2}{\gamma_2 - \beta_2} > 0$, и оценка завышена (так как $\beta_2 < 0$, то МНК-оценка, скорее всего, будет ближе к нулю, чем истинное значение коэффициента).

(6) Состоятельную оценку эластичности спроса по цене можно получить, если воспользоваться процедурой двухшагового МНК. Инструмент валиден, если он, во-первых, экзогенен и, во-вторых, релевантен.

В рассматриваемом случае переменная $\ln T_i$ экзогенна: по условию налоги не коррелированы со случайными величинами ε_i , характеризующими шоки спроса: $\text{cov}(\ln T_i, \varepsilon_i) = 0$.

К тому же налоги коррелированы с ценой: если они повышаются, то в результате сдвига кривой предложения увеличивается и равновесная цена. Поэтому $\text{cov}(\ln T_i, \ln P_i) \neq 0$.

Следовательно, инструмент $\ln T_i$ валиден.

На первом шаге оцениваем регрессию $\ln P_i$ по $\ln T_i$ и получаем прогнозные значения $\widehat{\ln P_i}$. На втором шаге оцениваем регрессию $\ln Q_i$ по $\widehat{\ln P_i}$, где уже отсутствует проблема эндогенности.

8.2. Двухшаговый МНК: множественная регрессия

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

Чтобы разобраться, как этот метод устроен для множественной регрессии, договоримся сначала об обозначениях.

Оцениваемое уравнение:

$$y_i = \beta_0 + \beta_1 * x_i^{(1)} + \dots + \beta_p x_i^{(p)} + \beta_{p+1} w_i^{(1)} + \dots + \beta_{p+r} w_i^{(r)} + \varepsilon_i.$$

$x_i^{(1)} \dots x_i^{(p)}$ — эндогенные регрессоры, для состоятельной оценки коэффициентов при которых требуются инструменты;

$w_i^{(1)} \dots w_i^{(r)}$ — экзогенные регрессоры, то есть регрессоры, которые не коррелированы со случайной ошибкой модели;

$z_i^{(1)} \dots z_i^{(m)}$ — инструментальные переменные.

В зависимости от соотношения значений параметров p и m возможны следующие ситуации:

- если $m = p$, то есть число эндогенных переменных совпадает с числом инструментов, то уравнение называется однозначно идентифицируемым (exactly identified)
- если $m > p$, то есть инструментов больше, чем эндогенных переменных, то уравнение называется сверхидентифицируемым (overidentified)
- если $m < p$, то есть инструментов меньше, чем эндогенных регрессоров, то уравнение неидентифицируемо (underidentified).

Как мы покажем ниже, в последнем случае применение двухшагового МНК невозможно. Если же $m \geq p$, то этот метод применим, и его реализация устроена так:

Первый шаг. Оцениваем p регрессий первого шага, в каждой из которых слева стоит один из эндогенных регрессоров ($x_i^{(1)} \dots x_i^{(p)}$), а справа — константа, все инструментальные переменные ($z_i^{(1)} \dots z_i^{(m)}$) и все экзогенные переменные ($w_i^{(1)} \dots w_i^{(r)}$). Оценив эти вспомогательные уравнения, получаем для каждой эндогенной переменной предсказанные значения $\hat{x}_i^{(1)}, \dots, \hat{x}_i^{(p)}$.

Второй шаг. Оцениваем параметры той модели, которая нас интересовала изначально. Только теперь в правой части вместо эндогенных регрессоров $x_i^{(1)} \dots x_i^{(p)}$ ставим их предсказанные значения из регрессий первого шага $\hat{x}_i^{(1)}, \dots, \hat{x}_i^{(p)}$. То есть оцениваем следующее уравнение второго шага:

$$y_i = \beta_0 + \beta_1 * \hat{x}_i^{(1)} + \dots + \beta_p \hat{x}_i^{(p)} + \beta_{p+1} w_i^{(1)} + \dots + \beta_{p+r} w_i^{(r)} + \varepsilon_i$$

Легко видеть, что процедура двухшагового МНК для множественной регрессии в целом аналогична случаю парной регрессии, который мы рассмотрели в предыдущем параграфе. Из описания этой процедуры понятно, почему невозможно идентифицировать уравнение, если инструментов меньше, чем эндогенных регрессоров. Действительно, представим, например, что $p = 2$, а $m = 1$, то есть нам нужно оценить коэффициенты при двух эндогенных переменных, и в нашем распоряжении есть всего один инструмент.

В этом случае в ходе оценивания регрессий первого шага мы получим предсказанные значения этих эндогенных регрессоров, выраженные через один и тот же инструмент: $\hat{x}_i^{(1)} = \hat{\alpha}_0 + \hat{\alpha}_1 z_i^{(1)}$ и $\hat{x}_i^{(2)} = \hat{\gamma}_0 + \hat{\gamma}_1 z_i^{(1)}$. Проблема в том, что раз $\hat{x}_i^{(1)}$ и $\hat{x}_i^{(2)}$ линейно выражаются через одну и ту же переменную, то они линейно выражаются и друг через друга. Иными словами, они являются линейно зависимыми. Поэтому, подставив их в регрессию второго шага, мы столкнемся с ситуацией чистой мультиколлинеарности, что, как мы знаем, делает вычисление оценок коэффициентов невозможным.

В прикладных исследованиях чаще всего встречается ситуация, в которой исследователь концентрируется на получении состоятельной оценки коэффициента при единственном эндогенном регрессоре. То есть в ситуации, когда $p = 1$. В этом случае 2МНК устроен следующим образом:

Регрессия первого шага. Оцениваем единственную регрессию x_i на все инструменты $z_i^{(1)} \dots z_i^{(m)}$ и на все экзогенные переменные $w_i^{(1)} \dots w_i^{(r)}$; получаем предсказанные значения $\hat{x}_i$:

$$\hat{x}_i = \hat{\alpha}_0 + \hat{\alpha}_1 z_i^{(1)} + \dots + \hat{\alpha}_m z_i^{(m)} + \hat{\alpha}_{m+1} w_i^{(1)} + \dots + \hat{\alpha}_{m+r} w_i^{(r)}$$

Регрессия второго шага. Регрессируем y_i на $\hat{x}_i$ и $w_i^{(1)} \dots w_i^{(r)}$:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 * \hat{x}_i + \hat{\beta}_2 w_i^{(1)} + \dots + \hat{\beta}_{1+r} w_i^{(r)}$$

Иногда возможна ситуация, в которой исследователь не уверен в экзогенности факторов $w_i^{(1)} \dots w_i^{(r)}$, однако хочет оставить их в уравнении второго шага в качестве контрольных переменных, чтобы избежать смещения из-за пропуска существенных переменных. В этом случае указанные переменные можно не включать в регрессию первого шага, но включать в регрессию второго шага.

Требования к инструментам в случае множественной регрессии аналогичны случаю парной регрессии:

- **Экзогенность.** Инструменты не должны быть коррелированы со случайными ошибками модели:

$$\text{Cov}(z_i^{(1)}, \varepsilon_i) = 0, \dots, \text{Cov}(z_i^{(m)}, \varepsilon_i) = 0$$

- **Релевантность.** Инструменты должны быть коррелированы с эндогенными регрессорами. Технически это означает, что в регрессии второго шага не должно возникать чистой мультиколлинеарности, не только для конечной выборки, но и при $n \rightarrow \infty$.

Если инструменты удовлетворяют обоим требованиям, то они называются валидными. В этом случае применение двухшагового МНК будет приводить к получению состоятельных оценок коэффициентов.

Важно подчеркнуть, что хотя асимптотические свойства 2МНК-оценок хороши, однако для конечных выборок всё не так радужно. В частности, 2МНК-оценки могут быть смещены. Кроме того, они, вообще говоря, не являются эффективными. С прикладной точки зрения это означает, что двухшаговый МНК лучше применять только на больших выборках, для которых корректно использование асимптотического подхода.

8.3. Векторно-матричная форма записи

Выведем в общем виде формулы 2МНК-оценки вектора коэффициентов для случая множественной регрессии.

В случае использования матричной формы записи можно переписать исходное уравнение модели

$$y_i = \beta_0 + \beta_1 * x_i^{(1)} + \dots + \beta_p x_i^{(p)} + \beta_{p+1} w_i^{(1)} + \dots + \beta_{p+r} w_i^{(r)} + \varepsilon_i$$

следующим образом:

$$y = X\beta + \varepsilon$$

Здесь X — матрица регрессоров, которая имеет $1 + p + r$ столбцов: один столбец из единиц для константы, p столбцов для эндогенных регрессоров $x_i^{(1)} \dots x_i^{(p)}$ и r столбцов для экзогенных переменных $w_i^{(1)} \dots w_i^{(r)}$. Число строк в этой матрице как обычно равно числу наблюдений n .

Дополнительно, нам потребуется матрица Z — n на $(1 + m + r)$ матрица, включающая константу, инструменты и экзогенные регрессоры — все переменные для регрессии первого шага.

На **первом шаге** двухшагового МНК мы оцениваем регрессию X по Z . Это можно записать так:

$$\hat{X} = Z\hat{\alpha}.$$

По формуле для МНК-оценки из параграфа 3.3 $\hat{\alpha} = (Z'Z)^{-1}Z'X$. Следовательно:

$$\hat{X} = Z(Z'Z)^{-1}Z'X.$$

На **втором шаге** двухшагового МНК мы оцениваем регрессию y по $\hat{X}$:

$$\hat{y} = \hat{X}\hat{\beta}^{TSLs}.$$

По формуле для МНК-оценки из параграфа 3.3 $\hat{\beta}^{TSLs} = (\hat{X}'\hat{X})^{-1}\hat{X}'y$. Следовательно:

$$\begin{aligned}\hat{\beta}^{TSLs} &= (\hat{X}'\hat{X})^{-1}\hat{X}'y = \\ &= \left((Z(Z'Z)^{-1}Z'X)' (Z(Z'Z)^{-1}Z'X) \right)^{-1} (Z(Z'Z)^{-1}Z'X)' y = \\ &= (X'Z(Z'Z)^{-1}Z'Z(Z'Z)^{-1}Z'X)^{-1} (Z(Z'Z)^{-1}Z'X)' y = \\ &= (X'Z(Z'Z)^{-1}Z'X)^{-1} (Z(Z'Z)^{-1}Z'X)' y = \\ &= (X'Z(Z'Z)^{-1}Z'X)^{-1} X'Z(Z'Z)^{-1}Z'y\end{aligned}$$

Таким образом, окончательно имеем следующую формулу для 2МНК-оценки:

$$\hat{\beta}^{TSLs} = (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y$$

Отметим, что для корректности приведенных выкладок необходимо, чтобы матрица $\hat{X}'\hat{X}$ была невырожденной (чтобы существовала обратная к ней матрица $(\hat{X}'\hat{X})^{-1}$). Именно для этого требуется выполнение свойства релевантности инструментов, которое мы сформулировали в конце предыдущего параграфа.

Если число инструментов в точности совпадает с числом эндогенных регрессоров ($m = p$), то матрицы X и Z имеют одинаковую размерность. В этом случае выражение для оценки $\hat{\beta}^{TSLs}$ можно упростить:

$$\begin{aligned}\hat{\beta}^{TSLs} &= (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y = \\ &= (Z'X)^{-1}(Z'Z)(X'Z)^{-1}X'Z(Z'Z)^{-1}Z'y = \\ &= (Z'X)^{-1}(Z'Z)(Z'Z)^{-1}Z'y = \\ &= (Z'X)^{-1}Z'y\end{aligned}$$

Этот частный случай двухшагового МНК иногда называют методом инструментальных переменных (instrumental variables). Таким образом, оценка по методу инструментальных переменных имеет вид:

$$\hat{\beta}^{IV} = (Z'X)^{-1}Z'y.$$

8.4. Тесты для моделей, оцененных двухшаговым МНК

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

2МНК-оценки коэффициентов являются состоятельными, только если используемые инструменты релевантны и экзогенны. Поэтому полезно уметь тестировать выполнение этих условий.

Начнем с релевантности. Рассмотрим случай, когда в модели есть единственная эндогенная переменная:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 w_i^{(1)} + \dots + \beta_{1+r} w_i^{(r)} + \varepsilon_i.$$

Тогда для проверки релевантности сначала следует оценить параметры регрессии первого шага:

$$\hat{x}_i = \hat{\alpha}_0 + \hat{\alpha}_1 z_i^{(1)} + \dots + \hat{\alpha}_m z_i^{(m)} + \hat{\alpha}_{m+1} w_i^{(1)} + \dots + \hat{\alpha}_{m+r} w_i^{(r)}$$

Затем, опираясь на тест для сравнения «короткой» и «длинной» регрессий, вычислить расчетное значение тестовой F-статистики для проверки гипотезы

$$H_0 : \alpha_1 = \dots = \alpha_m = 0.$$

Если эта гипотеза отвергается, то это значит, что инструменты вносят существенный вклад в объяснение изменений эндогенной переменной. Это говорит в пользу их релевантности. Обычно на практике используется следующее правило: если расчетное значение тестовой F-статистики для проверки данной гипотезы больше 10, то инструменты признаются релевантными. (И в целом, чем F-статистика больше, тем лучше.)

Эта рекомендация опирается на тот факт, что при выполнении указанного условия максимально возможное смещение 2МНК-оценки будет не слишком велико — в пределах 10% от истинного значения параметра. Поскольку доказательство данного факта весьма сложно технически, оно выходит за рамки нашего учебника. Однако заинтересованному (и не боящемуся трудностей) читателю мы советуем обратиться за подробностями к статье (Stock, Yogo, 2005).

Если же F-статистика меньше 10, то инструмент называется слабым. Это означает, что он почти не коррелирован с регрессором, то есть объясняет малую долю дисперсии эндогенной переменной.

Если используемые вами инструменты слабые, то:

- точность 2МНК оценок очень низка.
- результаты тестов на значимость могут быть некорректны, так как распределение оценки коэффициента не является нормальным даже асимптотически.

Поэтому, если результаты теста указывают, что вы имеете дело со слабыми инструментами, следует:

- или найти альтернативный набор инструментов, которые будут сильными (к сожалению, иногда сделать это довольно трудно);
- или, если у вас уже есть большое количество инструментов, попробовать отказаться от некоторых из них. Исключение наименее значимых инструментов может способствовать увеличению соответствующей F-статистики в регрессии первого шага (разумеется, при этом по-прежнему должно выполняться условие идентифицируемости $m \geq p$).

Пример 8.2. Оценка эластичности спроса по цене (продолжение)

Исследователь анализирует спрос на сигареты в 48 американских штатах. Он изучает зависимость величины спроса (Q) от цены (P), используя в качестве инструмента для цены ставку налога, взимаемого с производителей в соответствующем штате (T). Ниже представлены результаты применения двухшагового МНК:

Регрессия первого шага:

$$\widehat{\ln P}_i = 4,63 + 0,03 \cdot \ln T_i, R^2 = 0,47$$

(0,03) (0,005)

Регрессия второго шага (в скобках указаны робастные к гетероскедастичности стандартные ошибки для 2МНК):

$$\widehat{\ln Q}_i = 9,72 - 1,08 \cdot \widehat{\ln P}_i$$

(1,53) (0,32)

(а) Является ли инструмент рассматриваемой модели слабым?

(б) Постройте 95-процентный доверительный интервал для эластичности спроса по цене. Можно ли на основе полученного интервала утверждать, что цена значимо влияет на потребление сигарет? Можно ли на основе полученного интервала утверждать, что спрос на сигареты является эластичным?

Решение:

(а) Чтобы проверить, является ли инструмент слабым, в данном случае необходимо вычислить расчетное значение соответствующей F-статистики для проверки гипотезы $H_0 : \alpha_2 = 0$ в уравнении:

$$\ln P_i = \alpha_1 + \alpha_2 \cdot \ln T_i + u_i.$$

$$F_{расч} = \frac{\frac{R^2}{1-R^2} \cdot n - 2}{2-1} = \frac{\frac{0,47}{0,53} \cdot 48 - 2}{2-1} = 40,8.$$

Поскольку $F_{расч} > 10$, можно заключить, что инструмент релевантен (то есть не является слабым).

(б) Теперь построим 95-процентный доверительный интервал:

$$\begin{aligned} & (\widehat{\beta}_2 - 1,96 * se(\widehat{\beta}_2), \widehat{\beta}_2 + 1,96 * se(\widehat{\beta}_2)) \\ & (-1,08 - 1,96 * 0,32, -1,08 + 1,96 * 0,32) \\ & (-1,71, -0,45) \end{aligned}$$

Поскольку доверительный интервал не содержит ноль, можно утверждать, что цена значимо влияет на величину спроса.

Однако на основе полученного доверительного интервала нельзя утверждать, что спрос на сигареты является эластичным, так как в доверительный интервал входят как значения больше минус единицы, так и меньше неё.

Перейдем теперь к тестированию экзогенности инструмента. Для этого можно применить тест Саргана (Sargan test или the overidentifying restrictions test). Также иногда его называют J-тестом.

Тест доступен только в том случае, если число инструментов превышает число эндогенных регрессоров: $m > p$.

Нулевая гипотеза теста состоит в том, что все инструменты экзогенны. Альтернативная гипотеза — в том, что хотя бы один из инструментов эндогенен.

Для осуществления теста следует оценить параметры вспомогательной модели регрессии следующего вида:

$$e_i = \gamma_0 + \gamma_1 z_i^{(1)} + \dots + \gamma_m z_i^{(m)} + \gamma_{m+1} w_i^{(1)} + \dots + \gamma_{m+r} w_i^{(r)} + v_i$$

Здесь e_i — остатки, полученные в ходе оценивания регрессии второго шага двухшагового МНК.

Далее следует вычислить расчетное значение тестовой статистики по следующей формуле: $J_{\text{statistic}} = m * F$, где F — расчетное значение F-статистики для проверки гипотезы $\gamma_1 = \dots = \gamma_m = 0$ в этом вспомогательном уравнении.

Если верна нулевая гипотеза, то $J_{\text{statistic}}$ имеет распределение $\chi^2(m - p)$. Поэтому для принятия решения следует сравнивать расчетное значение статистики с критическим значением из таблиц распределения Хи-квадрат с $(m - p)$ степенями свободы. Как обычно, нулевая гипотеза не отвергается при заданном уровне значимости, если расчетное значение меньше критического. Это означает, что все инструменты экзогенны.

Если же нулевая гипотеза отклоняется, то нужно сделать вывод о том, что, по крайней мере, некоторые из инструментов эндогенны, а значит, 2МНК-оценки несостоятельны. Подчеркнем, что в этой ситуации вовсе не следует отказываться абсолютно от всех используемых инструментов, ведь тест говорит о том, что эндогенны лишь какие-то из них (но не обязательно все сразу). Поэтому в такой ситуации разумной идеей будет отказ только от части инструментов, повторная оценка уравнения и повторное осуществление теста. К сожалению, тест не указывает, с каким конкретно из инструментов возникла проблема, так что тут придется принимать решение самостоятельно.

Ещё раз подчеркнем, что проведение теста Саргана возможно лишь при $m > p$. Если же $m = p$, то тестировать экзогенность инструментов в принципе невозможно. Это означает, что при обосновании валидности используемых инструментальных переменных придется опираться на содержательные соображения. Следовательно, если вы используете 2МНК, то содержательное понимание моделируемого процесса особенно важно. Некоторые полезные примеры для такого случая приведены в следующем параграфе.

Третий тест, который мы обсудим в этом разделе — тест Хаусмана. Он помогает принять решение о том, нужно ли в принципе использование 2МНК в вашей модели, или можно ограничиться обычным МНК. Идея теста состоит в следующем: МНК-оценки будут состоятельны только в том случае, если регрессоры экзогенны, а 2МНК-оценки будут состоятельны независимо от того, эндогенен регрессор или экзогенен. Следовательно, если МНК- и 2МНК-оценки параметров отличаются не слишком сильно, то это аргумент в пользу применимости обычного МНК. Если же МНК- и 2МНК-оценки совсем не похожи друг на друга, то значит, МНК дает несостоятельные результаты, и пользоваться им не следует. Поэтому расчетное значение тестовой статистики опирается на сравнение векторов оценок, полученных при помощи обычного МНК и двухшагового.

Представление формулы тестовой статистики требует использования матричной записи (параграф 3.3 и параграф 8.3):

$$\begin{aligned} & (\widehat{\beta}_{2\text{МНК}} - \widehat{\beta}_{\text{МНК}})' (\widehat{V}(\widehat{\beta}_{2\text{МНК}}) - \widehat{V}(\widehat{\beta}_{\text{МНК}}))^{-1} (\widehat{\beta}_{2\text{МНК}} - \widehat{\beta}_{\text{МНК}}) \\ & \widehat{V}(\widehat{\beta}_{\text{МНК}}) \text{ — оценка ковариационной матрицы вектора МНК-оценок} \\ & \widehat{V}(\widehat{\beta}_{2\text{МНК}}) \text{ — оценка ковариационной матрицы вектора 2МНК-оценок} \end{aligned}$$

Нулевая гипотеза теста Хаусмана состоит в том, что МНК-оценки коэффициентов модели состоятельны. Если она верна, то указанная тестовая статистика имеет распределение $\chi^2(k)$, где k — суммарное число переменных в регрессии второго шага.

Если нулевая гипотеза не отвергается, следует использовать для оценки коэффициентов обычный МНК, так как он будет давать состоятельные результаты (и к тому же более точные, чем 2МНК). В случае отвержения нулевой гипотезы придется заключить, что МНК-оценки несостоятельны, и остановить свой выбор на 2МНК.

У теста Хаусмана есть ограничение: его применение корректно только, когда используемые инструменты валидны. Если вывод о том, что можно ограничиться обычным МНК, получен на основе использования теста Хаусмана с сомнительным набором инструментальных переменных, то и сам вывод будет оставаться спорным. Иными словами, для корректного проведения теста Хаусмана все равно придется искать набор годных инструментов. И это дает нам дополнительную мотивацию, чтобы перейти к следующему параграфу.

8.5. Где взять подходящие инструменты?

2МНК является панацеей для многих проблем с эндогенностью при условии, что у вас есть валидные инструменты. При этом, однако, остается вопрос о том, где их взять.

К сожалению, единого готового алгоритма, позволяющего автоматически получать инструментальные переменные для заданной модели, не существует. Поэтому использование двухшагового МНК требует не только владения эконометрической техникой, но и хорошего содержательного понимания исследуемой проблемы.

Именно опираясь на это понимание, следует выбирать потенциальных кандидатов в инструменты — переменные, не коррелированные со случайными ошибками модели и при этом коррелированные с регрессором. От того, насколько убедительно вы объясните читателю, почему предлагаемый вами инструмент должен быть валиден, зависит, будет ли ваша работа вызывать доверие или нет.

Чтобы продемонстрировать, какой способ думать об инструментах может помочь найти их, мы приведем в этом параграфе три истории. Каждая из них опирается на реальные исследования, собравшие большое количество цитирований. Я советую вам не ограничиваться чтением учебника, но и посмотреть сами эти статьи.

История 1. Влияет ли образование на доход?

В главе 7 (параграф 7.1) мы уже обсудили, что в регрессии дохода индивида на его уровень образования (количество лет обучения) регрессор будет эндогенным из-за пропуска ненаблюдаемой существенной переменной — уровня таланта и мотивации индивида. Действительно, более талантливые люди в среднем чаще продолжают обучение в университете, чем менее талантливые. В результате положительная корреляция между числом лет обучения индивида и его доходом может быть обусловлена не пользой от окончания вуза, а тем, что более талантливые люди одновременно и больше учились, и больше зарабатывают теперь.

Для устранения этой эндогенности нам нужно подобрать инструмент, который будет коррелирован с числом лет обучения (релевантен), но при этом не коррелирован с уровнем таланта (экзогенен).

В работе Card (1995)¹ в качестве такого инструмента используется расстояние от места жительства индивида до ближайшего колледжа. Речь идет о том месте, где жил индивид в момент принятия решения о продолжении обучения после школы. На первый взгляд эта переменная выглядит довольно экзотичной, однако в действительности она отвечает всем требованиям:

- переменная релевантна: хотя расстояние до ближайшего учебного заведения вряд ли является главным фактором принятия решения о продолжении обучения, однако если от дома до колледжа совсем близко, это несколько увеличивает шансы на то, что индивид примет решение продолжить обучение. Поэтому в данных наблюдается положительная корреляция между близостью к колледжу и числом лет обучения.
- переменная экзогенна: тот факт, что ваш дом расположен рядом с колледжем, сам по себе не делает вас более талантливым. А значит, близость к колледжу не должна быть коррелирована с уровнем таланта индивида.

Расчеты, проделанные Кардом, выявляют устойчивое значимое позитивное влияние образования на уровень дохода. Вы сможете воспроизвести их, разобрав одно из заданий для самостоятельного решения в конце главы.

Альтернативными инструментами, используемыми в моделях отдачи от образования, являются продолжительность обучения отца и продолжительность обучения матери индивида. Логика здесь такая: наличие у родителей ученой степени само по себе не делает их детей более талантливыми (экзогенность), однако увеличивает вероятность того, что ребёнок под их влиянием сам решит получить хорошее образование (релевантность).

История 2. Влияют ли дети на предложение труда?

В работе (Angrist, Evans, 1998)^[1] авторы исследуют, в какой степени дополнительный ребенок снижает готовность женщины работать. Иными словами, как дети влияют на предложение труда.

На первый взгляд ответ на этот вопрос кажется вполне очевидным: в большинстве случаев мать уделяет ребенку много времени и сил. Особенно, в первые годы его жизни. Это естественным образом снижает ее готовность выходить на рынок труда.

Однако при попытке количественно оценить данный эффект мы сталкиваемся с эндогенностью из-за обратной причинно-следственной связи: ведь не только наличие у женщины маленьких детей влияет на её готовность работать, но и востребованность женщины на рынке труда может влиять на ее склонность заводить детей. Для женщин, которые успешны в своей профессиональной деятельности (зарабатывают много денег и/или просто очень любят свою работу), решение отказаться от карьеры ради ребенка (пусть даже временно) связано с высокими альтернативными издержками.

Для устранения подобной эндогенности, требуется найти инструментальную переменную, которая будет коррелирована с решением женщины завести еще одного ребенка, но при этом не будет подвержена влиянию успешности карьеры этой женщины. Выполнение первого условия будет означать релевантность инструмента, а второго условия — его экзогенность.

В общем случае решить такую проблему достаточно трудно, однако авторы работы придумали, как найти подходящий инструмент для семей, в которых уже есть двое детей.

Они предположили, что в семьях, где первые дети одного и того же пола, вероятность того, что родители решатся завести еще одного ребенка, выше (по сравнению с семьями, где дети разнополые). Это связано с тем, что часть семей мечтают иметь и сына, и дочь. Проверка такого предположения на данных о нескольких тысячах американских семей подтвердила догадку авторов. Следовательно, бинарная переменная, которая равна единице для семей с однополыми детьми и равна нулю для семей с разнополыми детьми, является релевантным инструментом.

Одновременно, эта переменная является экзогенным инструментом, так как пол детей не зависит от карьерных успехов их родителей.

Расчеты с использованием этого инструмента позволили оценить, в какой степени дополнительный ребенок сказывается на предложении труда семьи. Подобные количественные оценки могут быть полезны для выбора мер демографической политики: например, поддержки многодетных семей.

[1] Angrist, J.D., and Evans, W.N. (1998). Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size *American Economic Review*, **88**, 450–477.

История 3. Влияет ли защита прав собственности на экономический рост?

На момент написания этих строк статья (Acemoglu, Robinson, 2001)³ имеет около двух тысяч цитирований и занимает 8 место в списке самых цитируемых статей из топ-5 научных экономических журналов в мире (речь о работах, опубликованных с 1990 года по настоящее время)⁴.

Авторы исследования проверяют гипотезу о том, что качество институтов имеет значение для обеспечения высоких темпов экономического роста. Точнее, о том, что страны, в которых права собственности защищены хорошо, имеют преимущество в накоплении капитала, что в конечном счете обеспечивает более высокий долгосрочный ВВП на душу населения по сравнению со странами, которые не уделяют внимания защите прав собственности.

[Acemoglu.dta](#)

Вы можете реплицировать (повторить) расчёты авторов, используя файл *Acemoglu*, который доступен среди прочих наборов данных для этого учебника. Там содержится информация о следующих переменных:

- `countryn` — название страны;
- `shortnam` — краткий код страны;
- `prot` — мера защиты прав собственности в данной стране (мера устроена так, что более высокие значения соответствуют более хорошей защите прав собственности). Это переменная интереса в анализируемой работе;
- `lgdp` — логарифм ВВП на душу населения в 1995 году. Это зависимая переменная;
- `logmort` — логарифм уровня смертности колонистов (см. детали далее);
- `latitude` — широта, на которой расположена столица данной страны (измерена как расстояние от экватора и нормирована таким образом, чтобы изменяться в пределах от нуля до единицы);
- `eu90` — доля населения европейского происхождения в данной стране по состоянию на 1975 год.

Если оценить параметры модели парной регрессии подушевого ВВП на защиту прав собственности, мы обнаружим положительную корреляцию между этими переменными. см. таблицу 8.1 и рисунок 8.1.

Таблица 8.1. Зависимость логарифма подушевого ВВП (`lgdp`) от переменной качества институтов (`prot`). Обычный МНК

Зависимая переменная: `lgdp`

Робастные оценки стандартных ошибок (с поправкой на гетероскедастичность), вариант HC1

	Коэффициент	Ст. ошибка	t-статистика	P-значение	
const	4,66038	0,320131	14,56	<0,0001	***
prot	0,522107	0,0499225	10,46	<0,0001	***
Среднее зав. перемен	8,062237		Ст. откл. зав. перемен	1,043359	
Сумма кв. остатков	31,53971		Ст. ошибка модели	0,713236	
R-квадрат	0,540115		Испр. R-квадрат	0,532697	
F(1, 62)	109,3770		P-значение (F)	2,57e-15	
Лог. правдоподобие	-68,16772		Крит. Акаике	140,3354	
Крит. Шварца	144,6532		Крит. Хеннана-Куинна	142,0364	

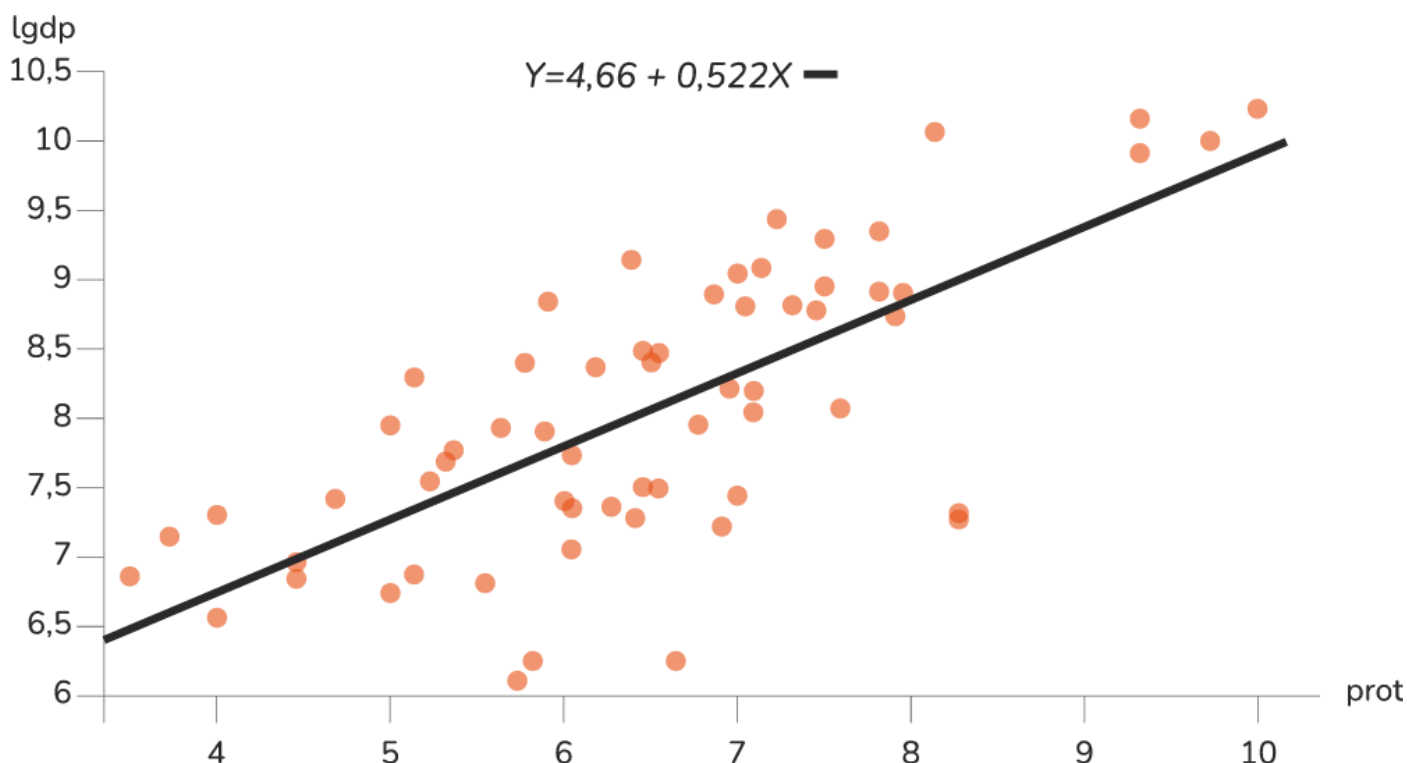


Рисунок 8.1. Зависимость логарифма подушевого ВВП (lgdp) от переменной качества институтов (prot)

Полученные результаты не стоит интерпретировать как причинно-следственную связь по трём причинам:

- Регрессор может быть эндогенен из-за пропущенных существенных переменных. На подушевой ВВП могут влиять другие факторы помимо качества институтов;
- Регрессор может быть эндогенен из-за двусторонней причинно-следственной связи. Возможно, дело не в том, что страны с лучшими институтами становятся богаче, а, наоборот, богатые страны могут позволить себе более хорошие институты.
- Наконец, регрессор наверняка эндогенен из-за ошибок его измерения. Действительно, вряд ли используемый индекс (равно как и любой другой) может идеально охарактеризовать такое сложное понятие, как защита прав собственности.

Если первая проблема может быть решена при помощи включения в модель контрольных переменных, то для преодоления двух других трудностей требуется инструмент.

При выборе инструмента авторы останавливаются на смертности поселенцев во время колонизации данной страны. Разумеется, такой подход требует ограничить выборку только странами, которые являются бывшими колониями, что авторы и делают. Этот инструмент отрицательно коррелирован с качеством институтов: в колониях, где смертность завоевателей была высока из-за скверных условий жизни (например, из-за климата и специфических заболеваний), они не были заинтересованы в создании институтов, направленных на защиту прав собственности, так как их интересовало только извлечение ресурсов из подконтрольных территорий, а жить там сами они не собирались. Это дает надежду на релевантность инструмента. Далее мы проверим это предположение на данных.

Также ясно, что этот инструмент экзогенен, потому что никакие шоки, связанные с текущим экономическим ростом, явно не влияют на смертность колонистов несколько веков назад.

В соответствии с логикой авторов статьи, увеличение смертности поселенцев должно негативно влиять на качество институтов, а снижение качества институтов, в свою очередь, должно сдерживать рост экономики. Следовательно, если предположения авторов верны, между смертностью поселенцев в прошлом и ВВП сегодня должна быть отрицательная корреляция. Именно такой характер связи мы и наблюдаем в уравнении в таблице 8.2 (коэффициент при переменной logmort значимый и отрицательный) и на рисунке 8.2 (который соответствует графику 1 в статье). Уравнение, в котором зависимая переменная регрессируется непосредственно на инструмент (а не на эндогенный регрессор), называется уравнением приведенной формы.

Таблица 8.2. Приведенная форма модели: зависимость логарифма подушевого ВВП (lgdp) от логарифма смертности колонистов (logmort). Обычный МНК

Приведенная форма: МНК, использованы наблюдения 1-64

Зависимая переменная: lgdp

Робастные оценки стандартных ошибок (с поправкой на гетероскедастичность), вариант HC1

	Коэффициент	Ст. ошибка	t-статистика	P-значение	
const	10,7076	0,383628	27,91	<0,0001	***
logmort	-0,569798	0,0730481	-7,800	<0,0001	***

Среднее зав. перемен	8,062237		Ст. откл. зав. перемен	1,043359
Сумма кв. остатков	36,86476		Ст. ошибка модели	0,771099
R-квадрат	0,462470		Испр. R-квадрат	0,453800
F(1, 62)	60,84495		P-значение (F)	8,79e-11
Лог. правдоподобие	-73,16000		Крит. Акаике	150,3200
Крит. Шварца	154,6378		Крит. Хеннана-Куинна	152,0210

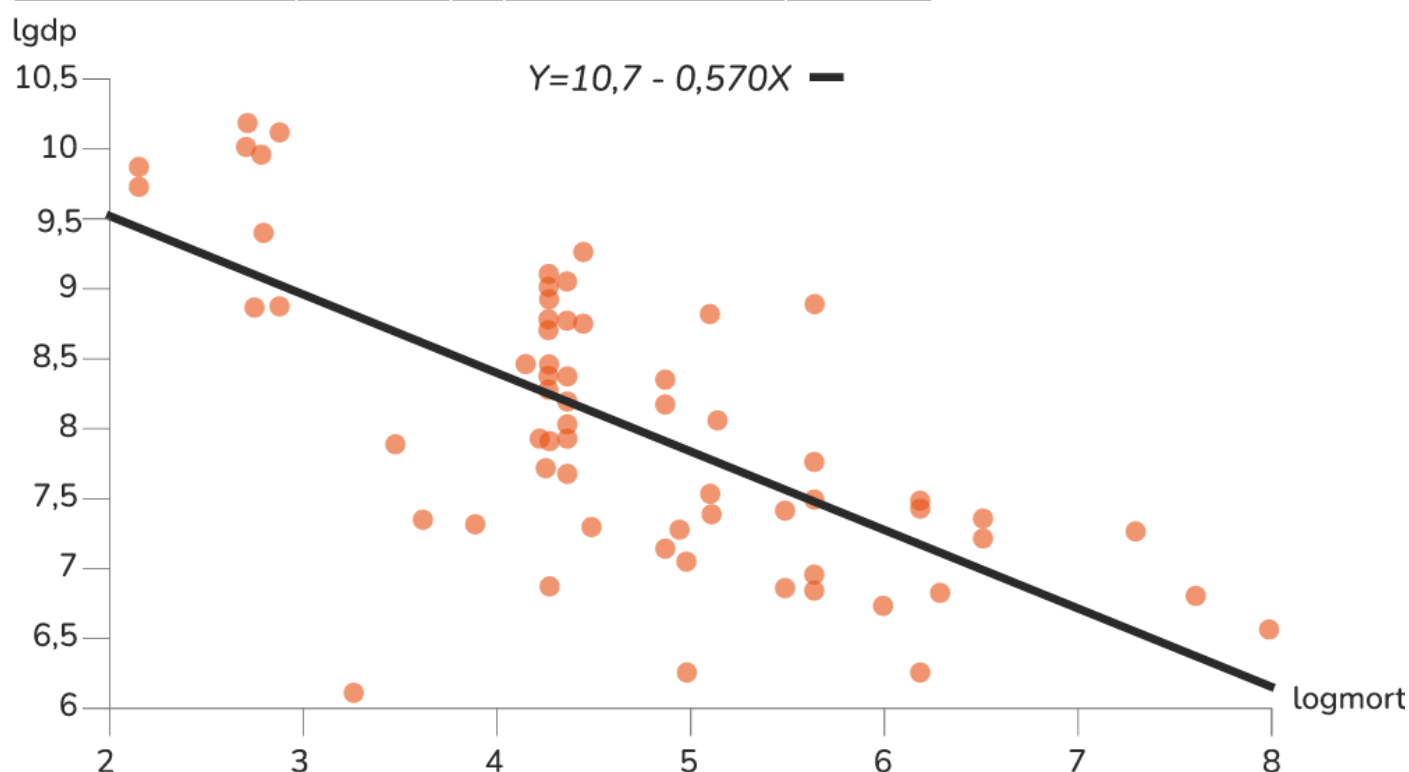


Рисунок 8.2. Приведенная форма модели: зависимость логарифма подушевого ВВП (lgdp) от логарифма смертности колонистов (logmort). Обычный МНК

Высказанные соображения приводят нас к использованию двухшагового МНК: на первом шаге следует оценить регрессию переменной $prot$ на инструмент $logmort$. На втором шаге следует оценить регрессию зависимой переменной $lgdp$ на предсказанные значения переменной интереса, полученные в результате оценки регрессии первого шага.

Результаты оценки регрессии первого шага представлены в таблице 8.3 и изображены графически на рисунке 8.3.

Как видно из таблицы и диаграммы ниже, смертность колонистов отрицательно коррелирует с качеством институтов, что соответствует рассуждениям авторов. F-статистика для теста на незначимость инструмента (в данном случае она же является F-статистикой для теста на незначимость уравнения) равна 23,8, что больше 10. Можно заключить, что инструмент является релевантным.

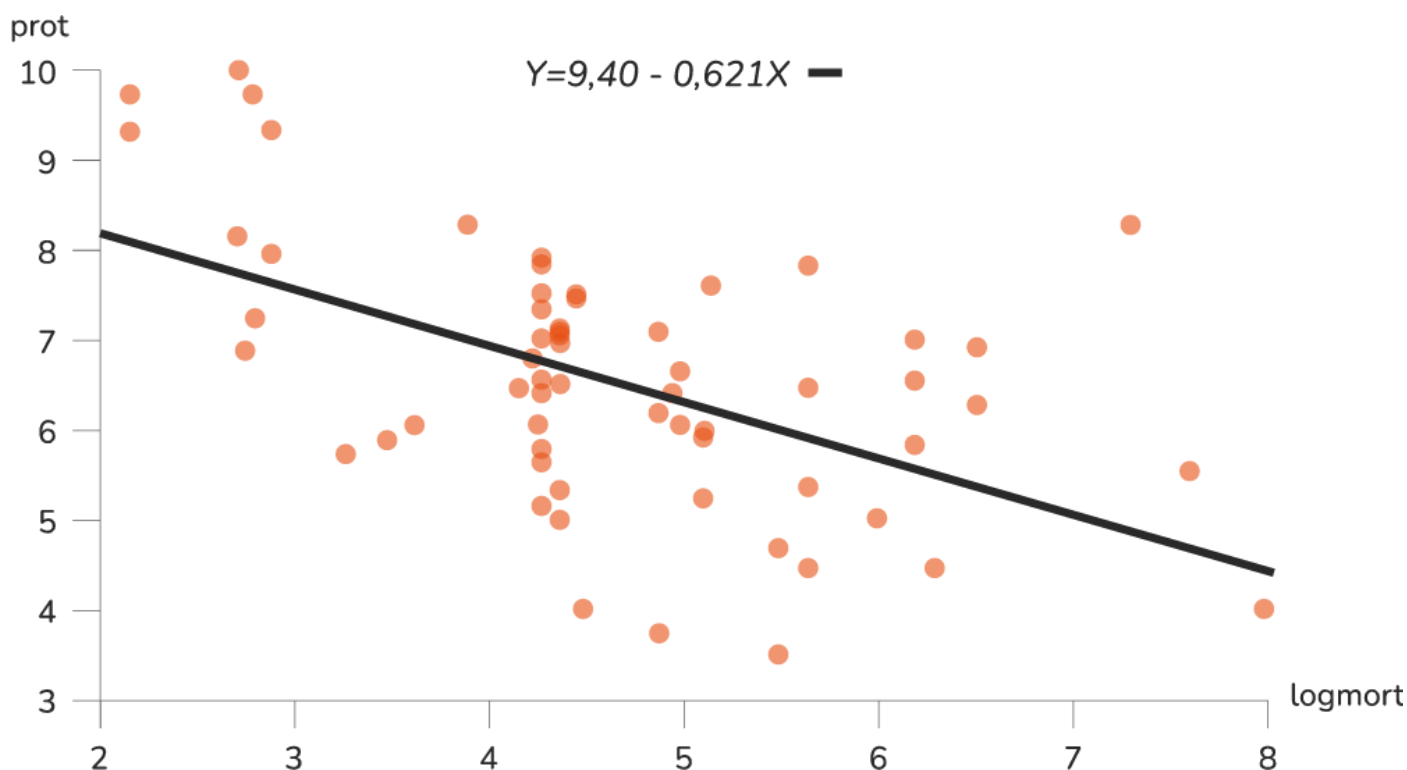


Рисунок 8.3. Регрессия первого шага двухшагового МНК: зависимость качества институтов (*prot*) от логарифма смертности колонистов (*logmort*). Обычный МНК

Таблица 8.3. Регрессия первого шага двухшагового МНК: зависимость качества институтов (*prot*) от логарифма смертности колонистов (*logmort*). Обычный МНК

first stage: МНК, использованы наблюдения 1-64

Зависимая переменная: *prot*

	Коэффициент	Ст. ошибка	t-статистика	P-значение	
const	9,40017	0,611645	15,37	<0,0001	***
logmort	-0,621318	0,127315	-4,880	<0,0001	***
Среднее зав. перемен	6,515625		Ст. откл. зав. перемен	1,468647	
Сумма кв. остатков	98,17445		Ст. ошибка модели	1,258356	
R-квадрат	0,277525		Испр. R-квадрат	0,265872	
F(1, 62)	23,81609		P-значение (F)	7,76e-06	
Лог. правдоподобие	-104,5037		Крит. Акаике	213,0074	
Крит. Шварца	217,3251		Крит. Хеннана-Куинна	214,7083	

Результаты оценивания регрессии второго шага представлены в таблице 8.4. Эффект воздействия качества институтов на реальный ВВП оказывается позитивным, значимым и довольно сильным. Стоит отметить, что его оценка увеличилась почти в два раза по сравнению с оценкой обычного МНК: с 0,52 до 0,92. Тест Хаусмана указывает на несостоятельность оценок обычного МНК и, следовательно, целесообразность использования двухшагового МНК.

Обратите внимание: в данном случае (в отличие от предыдущего пункта, с регрессией первого шага) использовались робастные стандартные ошибки. Это привело к некоторому изменению значения F-статистики, проверяющей релевантность инструмента. Однако эта F-статистика тоже больше 10, так что вывод о релевантности инструмента сохраняется вне зависимости от отсутствия или наличия поправки на гетероскедастичность.

Важно отметить, что 2МНК-оценка оказалась значительно больше по абсолютной величине, чем МНК-оценка. Это означает, что, скорее всего, преобладающей причиной эндогенности регрессора были ошибки измерения. Мы знаем, что в условиях ошибок измерения регрессора МНК смещает оценки коэффициента при этом регрессоре к нулю. 2МНК устранил эту проблему, и в результате оценка коэффициента оказалась больше.

Таблица 8.4. Зависимость логарифма подушевого ВВП (*lgdp*) от переменной качества институтов (*prot*). Двухшаговый МНК

Зависимая переменная: *lgdp*
Независимые переменные: *prot*

Инструменты: const logmort

Робастные оценки стандартных ошибок (с поправкой на гетероскедастичность), вариант HC1

	Коэффициент	Ст. ошибка	t-статистика	P-значение	
const	2,08689	1,12724	1,851	0,0689	*
prot	0,917080	0,169112	5,423	1,02e-06	***
Среднее зав. перемен	8,062237	Ст. откл. зав. перемен	1,043359		
Сумма кв. остатков	52,73844	Ст. ошибка модели	0,922291		
R-квадрат	0,540115	Испр. R-квадрат	0,532697		
F(1, 62)	29,40812	P-значение (F)	1,02e-06		
Лог. правдоподобие	-306,4709	Крит. Акаике	616,9418		
Крит. Шварца	621,2595	Крит. Хеннана-Куинна	618,6427		

Тест Хаусмана (Hausman) -

Нулевая гипотеза: МНК оценки состоятельны

Асимптотическая тестовая статистика: Хи-квадрат(1) = 22,3

p-значение = 2,36278e-006

Тест на слабые инструменты -

F-статистика для 1-го шага (1, 62) = 16,7

Значение < 10 может указывать на слабые инструменты

Авторы статьи предполагают, что климатические условия могут влиять на экономическое развитие, поэтому важно учесть этот фактор в модели. Переменная latitude как раз является замещающей переменной для климата. В таблице 8.5 представлены результаты оценивания 2МНК-регрессии с учетом этой контрольной переменной. Обратите внимание, что её включение в модель не сказывается на выводах по поводу влияния качества институтов на экономическое развитие: коэффициент при переменной prot остается значимым и положительным, а также не слишком сильно меняется по абсолютной величине.

Таблица 8.5. Зависимость логарифма подушевого ВВП (lgdp) от переменной качества институтов (prot) в случае добавления контрольной переменной. Двухшаговый МНК

Зависимая переменная: lgdp

Независимые переменные: prot

Инструменты: const logmort latitude

Робастные оценки стандартных ошибок (с поправкой на гетероскедастичность), вариант HC1

	Коэффициент	Ст. ошибка	t-статистика	P-значение	
const	1,93582	1,37215	1,411	0,1634	
prot	0,953286	0,228158	4,178	9,52e-05	***
latitude	-0,468469	1,18788	-0,3944	0,6947	
Среднее зав. перемен	8,062237	Ст. откл. зав. перемен	1,043359		
Сумма кв. остатков	56,55973	Ст. ошибка модели	0,962917		
R-квадрат	0,527875	Испр. R-квадрат	0,512395		
F(2, 61)	14,89629	P-значение (F)	5,40e-06		
Лог. правдоподобие	-304,7061	Крит. Акаике	615,4122		
Крит. Шварца	621,8889	Крит. Хеннана-Куинна	617,9637		

Тест Хаусмана (Hausman) -

Нулевая гипотеза: МНК оценки состоятельны

Асимптотическая тестовая статистика: Хи-квадрат(1) = 17,3

p-значение = 3,11377e-005

Тест на слабые инструменты -

F-статистика для 1-го шага (1, 61) = 10,0

Значение < 10 может указывать на слабые инструменты

2МНК-оценки позволяют заключить, что улучшение защиты прав собственности позитивно влияет на экономическое развитие.

В одном из заданий для самостоятельного решения вы сможете провести дополнительные расчеты, чтобы проверить устойчивость полученных авторами статьи результатов.

В заключение отметим, что хорошим подспорьем в поиске валидных инструментов может быть формальная теоретическая модель, описывающая взаимосвязь между анализируемыми переменными. Продемонстрируем это на следующем примере.

Пример 8.3. Система одновременных уравнений

Рассмотрим простую модель закрытой экономики, заданную следующей системой одновременных уравнений в структурной форме:

$$C_t = \beta_1 + \beta_2 Y_t + \varepsilon_t$$

$$I_t = \alpha_1 + \alpha_2 Y_t + u_t$$

$$Y_t = C_t + I_t + G_t$$

Y_t — ВВП в году t , C_t — совокупное потребление, I_t — совокупные инвестиции, G_t — государственные закупки.

$$E(\varepsilon_t) = 0, \text{var}(\varepsilon_t) = \sigma_\varepsilon^2, E(u_t) = 0, \text{var}(u_t) = \sigma_u^2, \text{var}(Y_t) = \sigma_Y^2, \text{cov}(\varepsilon_t, G_t) = 0, \text{cov}(u_t, G_t) = 0, \text{cov}(u_t, \varepsilon_t) = 0, \beta_2 > 0, \alpha_2 > 0, \alpha_2 + \beta_2 < 1.$$

Рассмотрите три способа оценки параметров функции потребления:

(а) обычный МНК,

(б) двухшаговый МНК, где в качестве инструмента для ВВП используются инвестиции,

(в) двухшаговый МНК, где в качестве инструмента для ВВП используются госзакупки.

Для каждого варианта вычислите предел по вероятности для оценки параметра β_2 (выразите этот предел через $\alpha_2, \beta_2, \sigma_\varepsilon^2, \sigma_u^2, \sigma_Y^2$). Для каждого варианта укажите, будет ли оценка параметра β_2 состоятельной или не состоятельной (для каждой несостоятельной оценки укажите, будет ли она завышена или занижена).

Решение:

Выразим Y_i через экзогенные переменные:

$$Y_i = \frac{\alpha_1 + \beta_1 + G_i + \varepsilon_i + u_i}{1 - \alpha_2 - \beta_2}$$

Отсюда получаем, что

$$\text{cov}(Y_i, \varepsilon_i) = \text{cov}\left(\frac{\alpha_1 + \beta_1 + G_i + \varepsilon_i + u_i}{1 - \alpha_2 - \beta_2}, \varepsilon_i\right) = \frac{\sigma_\varepsilon^2}{1 - \alpha_2 - \beta_2}$$

Здесь мы использовали нулевые ковариации, которые даны в условии.

Отметим, что по условию $\alpha_2 + \beta_2 < 1$, поэтому $\frac{\sigma_\varepsilon^2}{1 - \alpha_2 - \beta_2} > 0$. Поэтому:

$$\widehat{\beta}_2 = \frac{\widehat{\text{cov}}(C, Y)}{\widehat{\text{var}}(Y)} \xrightarrow{P} \frac{\text{cov}(C_i, Y_i)}{\text{var}(Y_i)} = \beta_2 + \frac{\text{cov}(Y_i, \varepsilon_i)}{\text{var}(Y_i)} =$$

$$\beta_2 + \frac{\frac{\sigma_\varepsilon^2}{1 - \alpha_2 - \beta_2} * 1}{\sigma_Y^2} > \beta_2$$

Таким образом, МНК-оценка несостоятельна и завышена.

(б) Перейдем теперь к 2МНК-оценке с инвестициями в качестве инструмента. Сразу обратим внимание на то, что в силу второго уравнения исходной системы инвестиции являются эндогенной переменной, а значит, вряд ли их использование в данном случае позволит получить состоятельную оценку коэффициента. Действительно:

$$\text{cov}(I_i, \varepsilon_i) = \text{cov}(\alpha_1 + \alpha_2 Y_i + u_i, \varepsilon_i) = \alpha_2 \text{cov}(Y_i, \varepsilon_i) = \alpha_2 \frac{\sigma_\varepsilon^2}{1 - \alpha_2 - \beta_2}$$

$$\text{cov}(I_i, Y_i) = \text{cov}(\alpha_1 + \alpha_2 Y_i + u_i, Y_i) = \alpha_2 \sigma_Y^2 + \text{cov}(u_i, Y_i) = \alpha_2 \sigma_Y^2 + \frac{\sigma_u^2}{1 - \alpha_2 - \beta_2}$$

$$\widehat{\beta}_2 = \frac{\widehat{\text{cov}}(C, I)}{\widehat{\text{cov}}(Y, I)} \xrightarrow{P} \frac{\text{cov}(C_i, I_i)}{\text{cov}(Y_i, I_i)} = \beta_2 + \frac{\text{cov}(I_i, \varepsilon_i)}{\text{cov}(Y_i, I_i)} =$$

$$\beta_2 + \frac{\frac{\alpha_2 \sigma_\varepsilon^2}{1 - \alpha_2 - \beta_2}}{\alpha_2 \sigma_Y^2 + \frac{\sigma_u^2}{1 - \alpha_2 - \beta_2}} > \beta_2$$

(в) Из условия задания и анализа исходной системы можно видеть, что госзакупки в нашем случае экзогенны, а значит, есть надежда на то, что они будут валидным инструментом:

$$\widehat{\beta}_2 = \frac{\widehat{\text{cov}}(C, G)}{\widehat{\text{cov}}(Y, G)} \xrightarrow{P} \frac{\text{cov}(C_i, G_i)}{\text{cov}(Y_i, G_i)} = \beta_2 + \frac{\text{cov}(G_i, \varepsilon_i)}{\text{cov}(Y_i, G_i)} = \beta_2.$$

Последнее равенство верно в силу того, что, во-первых, по условию $\text{Cov}(\varepsilon_i, G_i) = 0$, а во-вторых, $\text{cov}(Y_i, G_i) \neq 0$, потому что, как мы показали в самом начале решения, Y_i зависит от G_i .

2МНК-оценка, для которой государственные закупки используются в качестве инструмента, состоятельна.

1. Card, David (1995), "Using geographic variation in college proximity to estimate the return to schooling", in: Louis N. Christofides, E. Kenneth Grant and Robert Swidinsky, eds., Aspects of labour market behaviour: essays in honour of John Vanderkamp (University of Toronto Press, Toronto, Canada) pp. 201-222. [↵](#)
2. Acemoglu, Daron, Simon Johnson and James A. Robinson. The Colonial Origins of Comparative Development: An Empirical Investigation, American Economic Review, 2001, vol. 91(5) [↵](#)
3. Интересно отметить, что первое место в этом рейтинге занимает эконометрическая работа (Arellano, Bond, 1991), посвященная оценке динамических моделей на панельных данных. [↵](#)

Глава 9. Панельные данные

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

Если в выборке содержатся данные о нескольких объектах, каждый из которых наблюдается в течение нескольких моментов времени, то такие данные называют панельными (*panel data* или *longitudinal data*). Например, ежегодные данные о доходе и потреблении в 50 регионах некоторой страны за период с 1992 по 2011 годы.

В отличие от пространственных данных, на которых мы концентрировались в предыдущих главах, теперь для обозначения наблюдений нам будет удобно использовать не один, а два индекса:

x_{it} ,

где $i = 1, \dots, n$ — номер объекта (например, региона),

$t = 1, \dots, T$ — номер момента времени (например, номер года).

Есть несколько причин для использования панельных данных в прикладных исследованиях:

1. Большое количество наблюдений. Представьте, что вы проводите исследование, опираясь на информацию по странам мира. Если вы используете пространственные данные, то в вашем распоряжении, по всей видимости, будет менее 200 наблюдений. Ведь даже если вы включите в выборку все независимые государства-члены ООН, то их окажется меньше двухсот. А если оставить только те из них, по которым доступна достаточно полная статистическая информация, список окажется ещё короче. С другой стороны, применив панельные данные, вы будете иметь возможность использовать гораздо больше точек. Например, получив информацию о 100 странах за 10 лет, вы сможете строить регрессии по $100 \times 10 = 1000$ наблюдений.

Количество доступных наблюдений зависит от того, имеете ли вы дело со сбалансированной или несбалансированной панелью. Панель называется **сбалансированной**, если существует наблюдение для каждого объекта и для каждого момента времени. В этом случае общее число наблюдений равно $n \times T$. Когда в данных есть пропуски, панель называется **несбалансированной**. В этом случае общее число наблюдений меньше, чем $n \times T$, однако все равно может оставаться достаточно большим. Если возникновение пропусков является экзогенным, то для несбалансированных панелей можно использовать те же методы оценивания, что и для сбалансированных.

2. Возможность отслеживать динамику для множества объектов. Использование панельных данных позволяет анализировать распределение тех или иных эффектов во времени. Например, постепенное изменение потребления сигарет в некоторой стране после принятия антитабачных законов в ряде её регионов.

3. Дополнительный способ устранения эндогенности. Пожалуй, самым главным мотивом для использования панельных данных является возможность получить состоятельные оценки коэффициентов при интересующих нас переменных в условиях, когда на пространственной выборке это невозможно. Такой шанс появляется за счет **учета неоднородности моделируемых объектов**.

Чтобы понять, как указанная неоднородность может затруднять оценивание, рассмотрим пример. Представим, что нас интересует ответ на такой вопрос: влияет ли закон, разрешающий гражданам носить с собой личное огнестрельное оружие, на уровень преступности? Ответ на него, действительно, вовсе не очевиден. Сторонники закона утверждают, что его введение позволяет снизить преступность, так как гражданские лица получают шанс защититься от злоумышленников. Их оппоненты возражают, что в результате введения такого закона преступность, наоборот, вырастет из-за избыточного количества огнестрельного оружия на руках у населения и его спонтанного использования.

Пусть мы располагаем панельными данными о регионах некоторой страны, и уровень преступности в них описывается следующим уравнением:

$$y_{it} = \beta x_{it} + \mu_i + \varepsilon_{it} \quad (9.1)$$

Здесь y_{it} — уровень преступности в регионе i в год t ;

x_{it} — бинарная переменная, которая равна единице, если в регионе i в год t введен закон, разрешающий гражданам носить личное огнестрельное оружие, и равная нулю в противном случае;

ε_{it} — это, как обычно, случайные ошибки модели;

μ_i — ненаблюдаемая переменная, характеризующая специфические особенности каждого из регионов. Например, культурные или институциональные особенности, которые трудно поддаются измерению. Так как все регионы разные, почти нет шансов полностью учесть их специфику в наблюдаемых контрольных переменных. Поэтому такой фактор в модели наверняка останется. Обратите внимание, что у этой переменной нет индекса t , а есть только индекс i . Так мы подчеркиваем, что указанные особенности не меняются во времени (или меняются очень медленно, поэтому в рамках исследования могут считаться постоянными). Присутствие переменной μ_i порождает следующую дилемму:

- С одной стороны, мы не можем включить её в модель непосредственно, так как она не является наблюдаемой.
- С другой стороны, если она коррелирована с интересующей нас переменной x_{it} , то её невключение приведет к несостоятельности оценки коэффициента β из-за пропуска существенной переменной (см. главу 7).

Таким образом, неоднородность объектов часто становится причиной эндогенности регрессоров. К счастью, панельные данные дают возможность применить простой и хорошо работающий способ решения этой проблемы — использование моделей с фиксированными эффектами.

Именно этот класс моделей рассматривается в параграфах 9.1–9.4. Затем в параграфах 9.5–9.6 обсуждается альтернативный подход к оцениванию — модель со случайными эффектами. В конце главы, в параграфе 9.7, обобщена информация о

спецификационных тестах, которые помогут выбрать наиболее подходящий в каждом случае метод работы с панельными данными.

[9.1. Модель с фиксированными эффектами](#)

[9.2. Модель с фиктивными переменными](#)

[9.3. Внутригрупповое преобразование](#)

[9.4. Модель в первых разностях](#)

[9.5. Модель со случайными эффектами](#)

[9.6. Доступный ОМНК для оценивания модели со случайными эффектами](#)

[9.7. Спецификационные тесты](#)

[Задания для самостоятельного решения](#)

9.1. Модель с фиксированными эффектами

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

В рамках этого подхода индивидуальные особенности каждого объекта μ_i рассматриваются в качестве неизвестных исследователю (и ненаблюдаемых) параметров — так называемых фиксированных эффектов.

Предпосылки модели с фиксированными эффектами (случай парной регрессии):

1. Модель линейна по параметрам:

$$y_{it} = \beta x_{it} + \mu_i + \varepsilon_{it}, \quad i = 1, 2, \dots, n, \quad t = 1, 2, \dots, T$$

2. Наблюдения $\{(x_{i1}, x_{i2}, \dots, x_{iT}, \varepsilon_{i1}, \varepsilon_{i2}, \dots, \varepsilon_{iT}), i = 1, 2, \dots, n\}$ независимы и одинаково распределены

3. x_{it} и ε_{it} имеют ненулевые конечные четвертые моменты распределения $E(x_{it}^4) < \infty$, $E(\varepsilon_{it}^4) < \infty$

4. Случайные ошибки имеют нулевое условное математическое ожидание: $E(\varepsilon_{it} | x_{i1}, x_{i2}, \dots, x_{iT}, \mu_i) = 0$

Легко видеть, что эти предположения очень похожи на предпосылки линейной регрессионной модели со стохастическим регрессором из главы 6 с учетом добавления в уравнение ненаблюдаемых фиксированных эффектов и расширения формулировки на несколько периодов времени.

Отметим две важных детали, которые делают модель с фиксированными эффектами максимально близкой к реальности:

- Вторая предпосылка требует, чтобы значения регрессоров, относящиеся к разным объектам, были независимы друг от друга. Однако важно подчеркнуть, что она допускает наличие зависимости между значениями регрессоров, относящимися к **одному объекту**, но **разным моментам времени**: например, она допускает, что x_{i3} может быть коррелирован с x_{i2} , а тот, в свою очередь, может быть коррелирован с x_{i1} . Иными словами, будущие значения регрессора для данного объекта могут зависеть от его прошлых значений. Это реалистичное предположение. Например, потребление табака в данном регионе сегодня наверняка связано с его потреблением в прошлом. Аналогично, инфляция в России сегодня, скорее всего, влияет на будущую российскую инфляцию.
- Четвертая предпосылка требует, чтобы регрессор был экзогенен в том смысле, что он не должен быть связан со случайной ошибкой модели. Однако она допускает наличие корреляции между значением регрессора x_{it} и фиксированным эффектом μ_i . Это тоже реалистичная предпосылка. В рамках нашего примера про ношение оружия ясно, что культурные особенности данного региона (которые как раз и характеризуются его фиксированным эффектом) могут влиять на решение по поводу закона о ношении оружия в этом регионе (то есть на значение x_{it}).

Модель может быть сформулирована и для случая множественной регрессии. Тогда придется добавить предпосылку, требующую отсутствия чистой мультиколлинеарности. В остальном все предположения аналогичны случаю парной регрессии, только запись становится гораздо более громоздкой:

Предпосылки модели с фиксированными эффектами (случай множественной регрессии):

1. Модель линейна по параметрам:

$$y_{it} = \beta_1 x_{it}^{(1)} + \beta_2 x_{it}^{(2)} + \dots + \beta_k x_{it}^{(k)} + \mu_i + \varepsilon_{it},$$

$$i = 1, 2, \dots, n, \quad t = 1, 2, \dots, T.$$

2. Наблюдения

$\{(x_{i1}^{(1)}, x_{i2}^{(1)}, \dots, x_{iT}^{(1)}, x_{i1}^{(2)}, x_{i2}^{(2)}, \dots, x_{iT}^{(2)}, \dots, x_{i1}^{(k)}, x_{i2}^{(k)}, \dots, x_{iT}^{(k)}, \varepsilon_{i1}, \varepsilon_{i2}, \dots, \varepsilon_{iT}), i = 1, 2, \dots, n\}$ независимы и одинаково распределены.

3. $x_{it}^{(1)}, x_{it}^{(2)}, \dots, x_{it}^{(k)}$ и ε_{it} имеют ненулевые конечные четвертые моменты.

4. Случайные ошибки имеют нулевое условное математическое ожидание:

$$E(\varepsilon_{it} | x_{i1}^{(1)}, x_{i2}^{(1)}, \dots, x_{iT}^{(1)}, x_{i1}^{(2)}, x_{i2}^{(2)}, \dots, x_{iT}^{(2)}, \dots, x_{i1}^{(k)}, x_{i2}^{(k)}, \dots, x_{iT}^{(k)}, \mu_i) = 0.$$

5. С вероятностью 1 в модели отсутствует чистая мультиколлинеарность

Рассмотрим три подхода, которые позволяют получить состоятельные оценки коэффициентов при регрессорах в модели с фиксированными эффектами: модель с фиктивными переменными, внутригрупповое преобразование и модель в первых разностях.

9.2. Модель с фиктивными переменными

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

Естественным подходом к решению проблемы ненаблюдаемых фиксированных эффектов является добавление фиктивных переменных, характеризующих эти эффекты. Например, если в вашем распоряжении есть данные по регионам России, можно добавить в уравнение фиктивную переменную для каждого из регионов¹.

Зададим эти фиктивные переменные вот так:

$d_i^{(2)}$ — фиктивная переменная, которая равна единице, если наблюдение относится ко второму объекту ($i = 2$), и равна нулю в противном случае;

$d_i^{(3)}$ — фиктивная переменная, которая равна единице, если наблюдение относится к третьему объекту ($i = 3$), и равна нулю в противном случае;

...

$d_i^{(n)}$ — фиктивная переменная, которая равна единице, если наблюдение относится к объекту номер n ($i = n$), и равна нулю в противном случае.

Покажем, как эти фиктивные переменные могут быть применены для оценки коэффициентов в уравнении

$$y_{it} = \beta x_{it} + \mu_i + \varepsilon_{it}. \quad (9.1)$$

Для этого перепишем уравнение (9.1) следующим образом:

$$y_{it} = \beta x_{it} + \mu_1 + (\mu_2 - \mu_1)d_i^{(2)} + (\mu_3 - \mu_1)d_i^{(3)} + \dots + (\mu_n - \mu_1)d_i^{(n)} + \varepsilon_{it} \quad (9.2)$$

Легко проверить, что уравнение (9.2) эквивалентно уравнению (9.1). Действительно, если записать его для первого объекта (для $i = 1$), то все фиктивные переменные окажутся равными нулю, и мы получим:

$$y_{1t} = \beta x_{1t} + \mu_1 + \varepsilon_{1t}.$$

Если же записать его, например, для второго объекта (для $i = 2$), то окажется, что $d_i^{(2)} = 1$, $d_i^{(3)} = d_i^{(4)} = \dots = d_i^{(n)} = 0$. Следовательно, уравнение (9.2) примет вид:

$$y_{2t} = \beta x_{2t} + \mu_1 + (\mu_2 - \mu_1) + \varepsilon_{2t},$$

$$y_{2t} = \beta x_{2t} + \mu_2 + \varepsilon_{2t}.$$

Аналогично для всех остальных объектов. Таким образом, чтобы учесть фиксированные эффекты, достаточно добавить в модель константу и $(n - 1)$ фиктивную переменную². Эти фиктивные переменные мы всегда можем задать, то есть они наблюдаемы. Следовательно, проблема смещения из-за пропуска ненаблюдаемой существенной переменной (описанная в самом начале этой главы) решается, и обычный МНК снова будет давать состоятельные оценки коэффициентов при интересующих нас переменных.

Если переписать уравнение (9.2), используя более привычные обозначения, то можно представить его следующим образом³:

$$y_{it} = \beta_0 + \beta_1 x_{it} + \beta_2 d_i^{(2)} + \beta_3 d_i^{(3)} + \dots + \beta_n d_i^{(n)} + \varepsilon_{it} \quad (9.3)$$

Это так называемая модель с фиктивными переменными (*least squares dummy variables model*, *LSDV-модель*).

R-квадрат в этой модели обычно называют $LSDV-R^2$. Обычно на практике он оказывается сравнительно высоким за счет того, что большое количество фиктивных переменных объясняет значительную долю дисперсии зависимой переменной.

В рамках данного подхода легко осуществить тест, который ответит на вопрос, нужно ли учитывать фиксированные эффекты в уравнении. Иными словами, позволит сравнить модель с фиксированными эффектами и обычную регрессию, не учитывающую специфические особенности отдельных объектов (последняя называется регрессией пула, *pooled regression*). Для этого достаточно тестировать гипотезу

$$H_0: \beta_2 = \beta_3 = \dots = \beta_n = 0.$$

Альтернативная гипотеза состоит в том, что хотя бы один из коэффициентов отличается от нуля

Проверку можно осуществить при помощи стандартного теста на сравнение «короткой» и «длинной» регрессий. Если эта гипотеза отвергается, то следует использовать модель с фиксированными эффектами. В противном случае можно заключить, что все объекты в выборке одинаковые, и ограничиться использованием *pooled regression*.

Описанный способ оценивания демонстрирует преимущество панельных данных по сравнению с пространственными данными. Действительно, на пространственных данных оценить уравнение (9.3) при помощи МНК было бы невозможно, так как в этом случае в вашем распоряжении было бы всего n наблюдений. А в уравнении (9.3) аж $(n + 1)$ неизвестный параметр.

Есть у этого подхода и существенное ограничение. Модель с фиктивными переменными не позволяет идентифицировать коэффициенты при переменных, которые не меняются во времени. При попытке добавить такие переменные в модель в ней возникнет чистая мультиколлинеарность из-за того, что они окажутся линейно зависимы со множеством фиктивных переменных для отдельных объектов. Например, оценивая при помощи модели с фиксированными эффектами уравнение для заработной платы работника, вы сможете включить в него возраст, но не сможете включить бинарную переменную, характеризующую расу работника. Ведь раса работника, в отличие от его возраста, со временем не меняется и де-факто уже учтена в его фиксированном эффекте.

Двунаправленная модель с фиксированными эффектами

Модель (9.1) может быть обобщена следующим образом:

$$y_{it} = \beta x_{it} + \mu_i + \gamma_t + \varepsilon_{it}$$

Здесь γ_t — фиксированные эффекты различных периодов времени. Такие переменные могут быть полезны, если вы хотите учесть общие для всех объектов структурные изменения, которые происходят с течением времени. Скажем, если ваша выборка представляет собой данные по регионам России за 20 лет, то фиксированные эффекты объектов μ_i будут учитывать специфические особенности каждого из регионов, а временные эффекты γ_t — особенности различных лет, например, влияние на зависимую переменную экономических подъемов и спадов, характерных для экономики страны в целом.

Модель, которая учитывает оба типа эффектов, называется двунаправленной моделью (модель, учитывающая только один тип, соответственно, однонаправленной). Оценить её параметры можно, добавив в уравнение (9.3) ещё $(T - 1)$ фиктивную переменную:

$$y_{it} = \beta_0 + \beta_1 x_{it} + \beta_2 d_i^{(2)} + \dots + \beta_n d_i^{(n)} + \theta_2 td_i^{(2)} + \dots + \theta_T td_i^{(T)} + \varepsilon_{it}.$$

Здесь $td_i^{(2)}$ — бинарная переменная, которая равна единице, если наблюдение относится ко второму периоду времени, и равна нулю в противном случае;

$td_i^{(3)}$ — бинарная переменная, которая равна единице, если наблюдение относится к третьему периоду времени, и равна нулю в противном случае;

...

$td_i^{(T)}$ — бинарная переменная, которая равна единице, если наблюдение относится к периоду времени T , и равна нулю в противном случае.

Пример 9.1. Отдача от посещения лекций

Три сотни студентов изучали курс математического анализа, который длился два семестра. В файле [Attendance](#) доступны следующие данные о каждом из них:

performance_{it} — результат по курсу i -го студента в семестре t ;

$t = 1$ соответствует первому семестру, $t = 2$ соответствует второму семестру. Результат семестра измерен в баллах по шкале от 0 до 100.

attendance_{it} — количество лекций, посещенных i -м студентом в семестре t .

Результат студента по курсу описывается моделью:

$$\text{performance}_{it} = \beta * \text{attendance}_{it} + \mu_i + \gamma_t + \varepsilon_{it}$$

Здесь μ_i — переменная, которая характеризует индивидуальные особенности i -го студента. Например, уровень его мотивации и школьной математической подготовки.

Переменная γ_t , в свою очередь, характеризует особенности семестра t . Например, если $\gamma_2 > \gamma_1$, это будет свидетельствовать о том, что результаты студентов во втором семестре в среднем выше, чем в первом (при равной мотивации и равном количестве посещенных лекций), то есть о том, что второй семестр проще первого.

Исследовательский вопрос, на который мы попытаемся ответить: влияет ли посещение лекций на результат студента в семестре, или же этот результат полностью определяется индивидуальными особенностями студента?

В терминах модели этот вопрос можно переформулировать так: отличается ли от нуля коэффициент при переменной **attendance**?

Для ответа на этот вопрос оценим параметры трех уравнений.

Уравнение 1. Объединенная регрессия (регрессия пула, *pooled regression*):

$$\widehat{\text{performance}}_{it} = \widehat{\beta}_0 + \widehat{\beta}_1 * \text{attendance}_{it}.$$

Уравнение 2. Модель с фиксированными эффектами для студентов. Так как в нашем распоряжении есть данные про 300 студентов, необходимо добавить в уравнение 299 фиктивных переменных:

$$\widehat{\text{performance}}_{it} = \widehat{\beta}_0 + \widehat{\beta}_1 * \text{attendance}_{it} + \widehat{\beta}_2 d_i^{(2)} + \dots + \widehat{\beta}_{300} d_i^{(300)}.$$

Уравнение 3. Двунаправленная модель с фиксированными эффектами. Так как нам доступны данные за два периода, следует добавить в уравнение ещё одну фиктивную переменную для второго периода времени:

$$\widehat{\text{performance}}_{it} = \widehat{\beta}_0 + \widehat{\beta}_1 * \text{attendance}_{it} + \widehat{\beta}_2 d_i^{(2)} + \dots + \widehat{\beta}_{300} d_i^{(300)} + \widehat{\theta}_2 td_i^{(2)}.$$

Результаты представлены в таблице 9.1. Первый столбец соответствует уравнению 1, второй — уравнению 2, третий — уравнению 3.

Если взглянуть на первый столбец, можно обнаружить, что коэффициент при переменной *attendance* значим на однопроцентном уровне и отрицателен: он равен -0,989. Таким образом, из этого уравнения, казалось бы, можно заключить, что посещение каждой лекции в среднем снижает результат по курсу математического анализа на один балл.

Чтобы проверить, не вызван ли этот неожиданный вывод смещением оценки коэффициента из-за пропуска существенной переменной, посмотрим на второй столбец, где представлены результаты оценки модели с фиксированными эффектами. Прежде всего отметим, что гипотеза о равенстве нулю коэффициентов при всех фиктивных переменных фиксированных эффектов уверенно отвергается при любом разумном уровне значимости (см. значение F-статистики и соответствующее Р-значение внизу

второго столбца). Следовательно, мы должны сделать вывод о том, что предпочтительна модель с фиксированными эффектами для отдельных студентов. Иными словами, индивидуальные особенности каждого из студентов играют важную роль в определении его результатов по курсу.

Таблица 9.1. Три спецификации моделей для результатов семестра

	(1)	(2)	(3)
Константа	62,639***	34,163***	34,219***
	(2,420)	(0,509)	(0,528)
attendance	-0,989***	1,119***	1,116***
	(0,219)	(0,051)	(0,051)
d_2		-3,669***	-3,666***
		(0,051)	(0,051)
d_3		11,491***	11,492***
		(0,025)	(0,026)
...	...	...	...
d_299		26,809***	26,808***
		(0,025)	(0,026)
d_300		10,937***	10,932***
		(0,102)	(0,102)
td_2			-0,062
			(0,185)
Число наблюдений	600	600	600
R ²	0,046	0,993	0,993
F-статистика для проверки незначимости индивидуальных эффектов	—	138,5 (0,000)	137,8 (0,000)
F-статистика для проверки незначимости временных эффектов	—	—	0,2 (0,639)

В скобках под оценками коэффициентов указаны робастные стандартные ошибки. В скобках рядом с F-статистиками указаны P-значения для проверки соответствующей гипотезы. *** обозначает значимость на 1-процентном уровне.

Коэффициент при переменной *attendance* во втором столбце снова значим на однопроцентном уровне, но теперь оказался положительным: он равен 1,119. То есть посещение каждой лекции в среднем увеличивает результат студента по курсу примерно на 1,1 балла. Скажем, посещение десяти лекций добавит к результату семестра примерно 11 баллов. Учитывая, что суммарное количество баллов за семестр может принимать значения от 0 до 100, можно видеть, что лекции не играют решающей роли в оценке. Тем не менее, их вклад на самом деле является положительным, а не отрицательным (как мы могли бы ошибочно заключить, если бы ограничились моделью без фиксированных эффектов из первого столбца). Этот пример демонстрирует, как важно учитывать неоднородность моделируемых объектов для получения корректной оценки.

Кстати, отметим, что радикальное увеличение коэффициента R-квадрат в модели с фиксированными эффектами по сравнению с первой моделью говорит в пользу того, что индивидуальные особенности студентов играют решающую роль в их результате по курсу.

В третьем столбце рассматривается двусторонняя модель, то есть модель с добавлением ещё и фиктивной переменной для временного периода. Если бы у нас было больше временных периодов, то и таких бинарных переменных было бы больше одной. Формальный тест не отвергает гипотезу о равенстве нулю коэффициента при бинарной переменной времени. То есть данные не противоречат тому, что сложность двух семестров курса была примерно одинаковой и, следовательно, номер семестра не

влияет на результаты студентов. Поэтому в данном случае можно сделать выбор в пользу модели из второго столбца (впрочем, коэффициент при интересующей нас переменной примерно одинаков для второго и третьего столбцов).

-
1. Точнее, для всех регионов, кроме одного, чтобы избежать ловушки фиктивных переменных, которая описана в главе 4 [↪](#)
 2. Ещё раз подчеркнем, что если добавить к константе не $(n - 1)$, а n фиктивных переменных, то возникнет чистая мультиколлинеарность. [↪](#)
 3. Здесь мы обозначили $\beta_0 = \mu_1$, $\beta_i = (\mu_i - \mu_1)$, $i = 2, 3, \dots, n$. [↪](#)

9.3. Внутригрупповое преобразование

Оценка параметров модели с фиктивными переменными в случае большого числа объектов в выборке является довольно затратной в вычислительном смысле. Представьте, например, что уравнение (9.3) оценивается по данным о тысяче объектов. В этом случае в матрице регрессоров X будет 1001 столбец. Поэтому матрица $X'X$ имеет размер 1001 на 1001. Следовательно, для вычисления вектора МНК-оценок $\hat{\beta} = (X'X)^{-1}X'y$ потребуется вычислить матрицу, обратную к матрице 1001 на 1001. Если для современных компьютеров это не такая уж и сложная задача, то всего несколько десятилетий назад она была почти не решаемой.

Чтобы преодолеть эту трудность, было разработано так называемое внутригрупповое преобразование данных. Оценка, полученная на основе этого преобразования, называется внутригрупповой оценкой (*within estimator*). Она позволяет получить в точности ту же самую оценку коэффициента при регрессоре, что и LSDV-модель, но с меньшими вычислительными сложностями. Упрощение достигается за счет того, что в результате этого преобразования удается обойтись без вычисления величины отдельных фиксированных эффектов.

Придуманый ради экономии вычислительных мощностей способ оказался удобным, поэтому в прикладных исследованиях под оценкой модели с фиксированными эффектами по умолчанию понимают именно эту оценку.

Чтобы понять, как она устроена, снова вернемся к исходному уравнению (9.1):

$$y_{it} = \beta x_{it} + \mu_i + \varepsilon_{it} \quad (9.1)$$

Запишем его отдельно для каждого периода времени:

$$y_{i1} = \beta x_{i1} + \mu_i + \varepsilon_{i1}$$

$$y_{i2} = \beta x_{i2} + \mu_i + \varepsilon_{i2}$$

$$y_{i3} = \beta x_{i3} + \mu_i + \varepsilon_{i3}$$

...

$$y_{iT} = \beta x_{iT} + \mu_i + \varepsilon_{iT}$$

Сложим все эти уравнения друг с другом:

$$\sum_{t=1}^T y_{it} = \beta \sum_{t=1}^T x_{it} + T \cdot \mu_i + \sum_{t=1}^T \varepsilon_{it}.$$

Разделим результат на T :

$$\bar{y}_i = \beta \bar{x}_i + \mu_i + \bar{\varepsilon}_i. \quad (9.4)$$

Здесь мы воспользовались следующими обозначениями в средних по времени:

$$\bar{y}_i = \frac{\sum_{t=1}^T y_{it}}{T}, \quad \bar{x}_i = \frac{\sum_{t=1}^T x_{it}}{T}, \quad \bar{\varepsilon}_i = \frac{\sum_{t=1}^T \varepsilon_{it}}{T}.$$

Осталось вычесть из уравнения (9.1) уравнение (9.4):

$$y_{it} - \bar{y}_i = \beta x_{it} - \beta \bar{x}_i + \mu_i - \mu_i + \varepsilon_{it} - \bar{\varepsilon}_i,$$

$$(y_{it} - \bar{y}_i) = \beta (x_{it} - \bar{x}_i) + (\varepsilon_{it} - \bar{\varepsilon}_i) \quad (9.5).$$

Обозначим отклонения от средних по времени следующим образом:

$$\tilde{y}_{it} = y_{it} - \bar{y}_i, \quad \tilde{x}_{it} = x_{it} - \bar{x}_i, \quad \tilde{\varepsilon}_{it} = \varepsilon_{it} - \bar{\varepsilon}_i.$$

В этом случае равенство (9.5) примет вид:

$$\tilde{y}_{it} = \beta \tilde{x}_{it} + \tilde{\varepsilon}_{it}.$$

Мы снова устранили ненаблюдаемые фиксированные эффекты из модели. Теперь для оценки параметра β нам достаточно вычислить МНК-оценку в обычной парной регрессии без константы (см. задание 8 к главе 2):

$$\tilde{\beta} = \frac{\sum_{i=1}^n \sum_{t=1}^T \tilde{x}_{it} \tilde{y}_{it}}{\sum_{i=1}^n \sum_{t=1}^T (\tilde{x}_{it})^2}.$$

Если для исходного уравнения (9.1) выполнены все предпосылки модели с фиксированными эффектами, то для переменных преобразованного уравнения (9.5) будут выполнены предпосылки регрессионной модели со стохастическим регрессором из главы 6. Отсюда ясно, что МНК-оценка коэффициента в таком уравнении будет состоятельной и асимптотически нормальной. Для наглядности докажем состоятельность этой оценки непосредственно:

Теорема о состоятельности within-оценки

Если выполнены предпосылки 1-5 модели с фиксированными эффектами, то $\tilde{\beta}$ будет состоятельной оценкой параметра β при $n \rightarrow \infty$.

Доказательство:

В силу предпосылки о конечных четвертых моментах распределения, существуют конечные пределы по вероятности для следующих выражений (детальное обоснование для этого факта может быть проведено аналогично доказательству состоятельности обычных МНК-оценок в главе 6):

$$\frac{1}{n} \sum_{i=1}^n \sum_{t=1}^T \tilde{x}_{it} \tilde{\varepsilon}_{it} \xrightarrow{p} E \sum_{t=1}^T \tilde{x}_{it} \tilde{\varepsilon}_{it} < \infty,$$

$$\frac{1}{n} \sum_{i=1}^n \sum_{t=1}^T (\tilde{x}_{it})^2 \xrightarrow{p} E \sum_{t=1}^T (\tilde{x}_{it})^2 < \infty.$$

Опираясь на этот факт, можно вычислить предел по вероятности для within-оценки:

$$\begin{aligned}\tilde{\beta} &= \frac{\frac{1}{n} \sum_{i=1}^n \sum_{t=1}^T \tilde{x}_{it} \tilde{y}_{it}}{\frac{1}{n} \sum_{i=1}^n \sum_{t=1}^T (\tilde{x}_{it})^2} = \frac{\frac{1}{n} \sum_{i=1}^n \sum_{t=1}^T \tilde{x}_{it} (\beta \tilde{x}_{it} + \tilde{\varepsilon}_{it})}{\frac{1}{n} \sum_{i=1}^n \sum_{t=1}^T (\tilde{x}_{it})^2} = \\ &= \frac{\beta \frac{1}{n} \sum_{i=1}^n \sum_{t=1}^T (\tilde{x}_{it})^2 + \frac{1}{n} \sum_{i=1}^n \sum_{t=1}^T \tilde{x}_{it} \tilde{\varepsilon}_{it}}{\frac{1}{n} \sum_{i=1}^n \sum_{t=1}^T (\tilde{x}_{it})^2} \xrightarrow{p} \\ &\xrightarrow{p} \frac{\beta E \sum_{t=1}^T (\tilde{x}_{it})^2 + E \sum_{t=1}^T \tilde{x}_{it} \tilde{\varepsilon}_{it}}{E \sum_{t=1}^T (\tilde{x}_{it})^2} = \beta + \frac{E \sum_{t=1}^T \tilde{x}_{it} \tilde{\varepsilon}_{it}}{E \sum_{t=1}^T (\tilde{x}_{it})^2} = \beta\end{aligned}$$

В последнем равенстве мы воспользовались тем, что в силу экзогенности регрессора $E \sum_{t=1}^T \tilde{x}_{it} \tilde{\varepsilon}_{it} = 0$.

Полученная модель называется внутригрупповой регрессионной моделью (within-group regression), так как объясняет вариацию зависимой переменной вокруг среднего значения для группы наблюдений, относящихся к разным моментам времени, но к одному объекту. Внутригрупповое преобразование для множественной регрессии осуществляется аналогичным образом.

Коэффициент R-квадрат, посчитанный для такой модели, называется within- R^2 . Он не совпадает с LSDV- R^2 и не может быть сопоставлен с ним, так как рассчитывается на основе общей суммы квадратов остатков для преобразованной переменной (а не для исходной). Поэтому, приводя коэффициент детерминации для вашей модели, не забывайте уточнять, о каком именно R-квадрате идет речь.

В отличие от R-квадратов, оценки коэффициентов при переменных, полученные в ходе применения внутригруппового преобразования, в точности совпадают с оценками LSDV-модели. В этом смысле два указанных подхода эквивалентны.

В случае выполнения предпосылок 1-5 модели с фиксированными эффектами within-оценка асимптотически нормальна, что позволяет тестировать гипотезы относительно соответствующих коэффициентов и строить доверительные интервалы для них. Но при вычислении стандартных ошибок для модели с фиксированными эффектами следует принимать во внимание не только возможную гетероскедастичность, но и то, что случайные ошибки, относящиеся к одному объекту (но к разным моментам времени), могут быть коррелированы между собой. Оценка стандартных ошибок, состоятельная в условиях такой кластеризации, была предложена Ареллано, поэтому соответствующие стандартные ошибки называются стандартные ошибки в форме Ареллано (Arellano standard errors) или состоятельные в условиях кластеризации стандартные ошибки (cluster-robust standard errors).

Пример 9.2. Отдача от посещения лекций (продолжение)

Вернемся к нашему примеру с оценкой пользы от посещения лекций (см. пример 9.1). Оценим теперь коэффициент при переменной *attendance* при помощи внутригруппового преобразования. Ниже представлена стандартная выдача эконометрического пакета для этого случая.

Модель 1: Фиксированные эффекты, использовано наблюдений - 600
Включено 300 пространственных объектов
Длина временного ряда = 2
Зависимая переменная: performance
Робастные стандартные ошибки

	Коэффициент	Ст. ошибка	t-статистика	P-значение	
const	40,1209	0,384245	104,4	2,61e-237	***
attendance	1,11873	0,0359668	31,10	9,90e-096	***
Сумма кв. остатков	759,9752	Ст. ошибка модели	1,594278		
LSDV R-squared	0,993163	within-R-квадрат	0,786506		

Обратите внимание, что интересующий нас коэффициент оказался равен 1,12. Он совпадает с коэффициентом при переменной *attendance*, который мы получили при помощи МНК с фиктивными переменными. Чтобы убедиться в этом, посмотрите на второй столбец в таблице 9.1 в примере 9.1. Таким образом, мы убедились на примере, что два этих метода оценивания приводят к одинаковому итогу. Стандартные ошибки для коэффициентов несколько отличаются. Это связано с тем, что в последнем случае использованы стандартные ошибки в форме Ареллано (Arellano standard errors), которые учитывают панельную структуру данных.

В последней таблице также приведены два варианта коэффициентов R-квадрат: LSDV-R-квадрат и within-R-квадрат. Первый из них — это тот же самый R-квадрат, что и во втором уравнении из таблицы 9.1. Он, как это обычно и бывает, больше, чем within-R-квадрат.

9.4. Модель в первых разностях

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

Представим для начала, что в нашем распоряжении есть данные только за два периода времени $t = 1, 2$. В этом случае модель (9.1) можно записать для каждого из них:

$$y_{i1} = \beta x_{i1} + \mu_i + \varepsilon_{i1}$$

$$y_{i2} = \beta x_{i2} + \mu_i + \varepsilon_{i2}$$

Вычтем из второго уравнения первое:

$$y_{i2} - y_{i1} = \beta (x_{i2} - x_{i1}) + (\varepsilon_{i2} - \varepsilon_{i1})$$

Обозначим $\Delta y_i = y_{i2} - y_{i1}$, $\Delta x_i = x_{i2} - x_{i1}$, $u_i = \varepsilon_{i2} - \varepsilon_{i1}$. С учетом этих обозначений наше новое уравнение примет вид:

$$\Delta y_i = \beta \Delta x_i + u_i$$

Обратите внимание, что в результате трюка с переходом к первым разностям мы снова устранили фиксированные эффекты из модели. В этом и есть основная идея модели в первых разностях.

Следовательно, в новой модели состоятельную оценку коэффициента при переменной можно получить при помощи МНК. Её формула будет соответствовать обычной МНК-оценке в парной регрессии без константы:

$$\widehat{\beta}_{FD} = \frac{\sum_{i=1}^n \Delta x_i \Delta y_i}{\sum_{i=1}^n (\Delta x_i)^2}.$$

Для преобразованной модели в первых разностях оказываются выполненными все предпосылки модели со стохастическими регрессорами из главы 6, что гарантирует состоятельность и асимптотическую нормальность оценки $\widehat{\beta}_{FD}$. Проверим путем непосредственных вычислений, что эта оценка действительно является состоятельной:

$$\begin{aligned} \widehat{\beta}_{FD} &= \frac{\sum_{i=1}^n \Delta x_i \Delta y_i}{\sum_{i=1}^n (\Delta x_i)^2} = \frac{\frac{1}{n} \sum_{i=1}^n \Delta x_i \Delta y_i}{\frac{1}{n} \sum_{i=1}^n (\Delta x_i)^2} \xrightarrow{p} \frac{E \Delta x_i \Delta y_i}{E (\Delta x_i)^2} = \\ &= \frac{E \Delta x_i (\beta \Delta x_i + u_i)}{E (\Delta x_i)^2} = \frac{\beta E (\Delta x_i)^2 + E \Delta x_i u_i}{E (\Delta x_i)^2} = \beta + \frac{E \Delta x_i u_i}{E (\Delta x_i)^2} = \beta \end{aligned}$$

В последнем переходе мы используем тот факт, что $E \Delta x_i u_i = 0$ в силу предпосылки об экзогенности регрессора.

В рамках этого подхода тоже невозможно оценить коэффициент при переменной, которая не меняется во времени.

Действительно, представьте, что для наблюдений выполнено равенство: $x_{i2} = x_{i1}$. Тогда для всех наблюдений $\Delta x_i = 0$.

Следовательно, в этом случае $\sum_{i=1}^n (\Delta x_i)^2 = 0$ и вычислить $\widehat{\beta}_{FD} = \frac{\sum_{i=1}^n \Delta x_i \Delta y_i}{\sum_{i=1}^n (\Delta x_i)^2}$ просто не получится.

Поэтому для работоспособности модели с первыми разностями требуется, чтобы хотя бы для некоторых наблюдений значение регрессора менялось во времени. В нашем примере с законом о ношении оружия это означает, что хотя бы для некоторых регионов в первом периоде закон должен отсутствовать, а во втором периоде он должен быть принят в них же.

Разумеется, вовсе не обязательно ограничиваться моделью с единственной объясняющей переменной и всего с двумя периодами. Рассмотренный нами подход может быть обобщен и на случай множественной регрессии, и на произвольное число временных периодов. Для этого рассмотрим модель множественной регрессии с фиксированными эффектами:

$$y_{it} = \sum_{j=1}^k \beta_j * x_{it}^{(j)} + \mu_i + \varepsilon_{it}, \quad i = 1, 2, \dots, n, \quad t = 1, 2, \dots, T$$

Вычтем из уравнения для периода t уравнение для периода $t - 1$:

$$\Delta y_{it} = \sum_{j=1}^k \beta_j * \Delta x_{it}^{(j)} + u_{it}, \quad i = 1, 2, \dots, n, \quad t = 2, 3, \dots, T$$

Здесь $\Delta y_{it} = y_{it} - y_{it-1}$, $\Delta x_{it} = x_{it} - x_{it-1}$, $u_{it} = \varepsilon_{it} - \varepsilon_{it-1}$. В преобразованной модели снова устранены ненаблюдаемые фиксированные эффекты, что решает проблему, описанную в самом начале данной главы. Следовательно, параметры в преобразованной модели можно состоятельно оценить при помощи МНК.

Если в вашем распоряжении имеются данные только за два временных периода ($T = 2$), то оценка модели в первых разностях численно равна within-оценке. В случае большего числа периодов эти оценки могут не совпадать.

9.5. Модель со случайными эффектами

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

Альтернативным подходом к моделированию на панельных данных является модель со случайными эффектами.

Предпосылки модели со случайными эффектами:

1. Модель линейна по параметрам:

$$y_{it} = \beta_1 x_{it}^{(1)} + \beta_2 x_{it}^{(2)} + \dots + \beta_k x_{it}^{(k)} + \mu_i + \varepsilon_{it},$$

$$i = 1, 2, \dots, n, \quad t = 1, 2, \dots, T.$$

2. Наблюдения

$$\left\{ \left(x_{i1}^{(1)}, x_{i1}^{(2)}, \dots, x_{iT}^{(1)}, x_{iT}^{(2)}, \dots, x_{iT}^{(k)}, x_{i1}^{(k)}, x_{i2}^{(k)}, \dots, x_{iT}^{(k)}, \varepsilon_{i1}, \varepsilon_{i2}, \dots, \varepsilon_{iT} \right), \quad i = 1, 2, \dots, n, \quad t = 1, 2, \dots, T \right\}$$

независимы и одинаково распределены.

3. $x_{it}^{(1)}, x_{it}^{(2)}, \dots, x_{it}^{(k)}$ и ε_{it} имеют ненулевые конечные четвертые моменты.

4. Случайные ошибки имеют нулевое условное матожидание:

$$E(\varepsilon_{it} | x_{i1}^{(1)}, x_{i2}^{(1)}, \dots, x_{iT}^{(1)}, x_{i1}^{(2)}, x_{i2}^{(2)}, \dots, x_{iT}^{(2)}, \dots, x_{i1}^{(k)}, x_{i2}^{(k)}, \dots, x_{iT}^{(k)}, \mu_i) = 0.$$

5. С вероятностью единица в модели отсутствует чистая мультиколлинеарность.

$$6. E(\mu_i | x_{i1}^{(1)}, x_{i2}^{(1)}, \dots, x_{iT}^{(1)}, x_{i1}^{(2)}, \dots, x_{iT}^{(2)}, \dots, x_{i1}^{(k)}, \dots, x_{iT}^{(k)}) = E(\mu_i) = 0$$

Сравнив данный набор условий с предпосылками модели с фиксированными эффектами, легко видеть, что эта модель отличается всего одним дополнительным пунктом: №6 (остальные предпосылки в точности совпадают). Из него следует, что регрессоры не должны быть коррелированы с ненаблюдаемыми эффектами μ_i .

Таким образом, называя величины μ_i случайными эффектами, в прикладных исследованиях предполагают их некоррелированность с регрессорами.

Обозначим $v_{it} = \mu_i + \varepsilon_{it}$. В этом случае исходное уравнение можно переписать следующим образом:

$$y_{it} = \beta_1 x_{it}^{(1)} + \beta_2 x_{it}^{(2)} + \dots + \beta_k x_{it}^{(k)} + v_{it}.$$

В этом уравнении регрессоры экзогенны, так как они не коррелированы ни с одной из компонент v_{it} .

Так как регрессоры в модели со случайными эффектами экзогенны, параметры этой модели могут быть состоятельно оценены обычным МНК. Однако, как правило, на практике для этого используется не обычный МНК, а доступный обобщенный МНК. Дело в том, что МНК-оценки в этой модели будут хоть и состоятельными, но неэффективными. Поэтому применение доступного обобщенного МНК позволяет получить более точные результаты.

Читателя, заинтересованного в технических деталях получения оценки доступного ОМНК, мы приглашаем обратиться к параграфу 9.6. Здесь же мы отметим несколько важных с прикладной точки зрения соображений о модели со случайными эффектами:

- Преимуществом модели со случайными эффектами является возможность идентификации коэффициентов при переменных, которые не меняются во времени. Как вы помните, для модели с фиксированными эффектами это было невозможно.
- Недостатком же являются более жесткие предпосылки: требование некоррелированности регрессоров и случайных эффектов. Выполнение этой предпосылки можно тестировать (см. параграф 9.7).
- Так как параметры модели оцениваются не обычным МНК, а доступным ОМНК, то R-квадрат для этой модели не определен.

Пример 9.3. Отдача от посещения лекций (продолжение)

Вернемся к нашему примеру с оценкой пользы от посещения лекций (см. примеры 9.1 и 9.2). Оценим теперь интересующий нас параметр при помощи модели со случайными эффектами. Соответствующие результаты представлены ниже.

Модель 1: Случайные эффекты (GLS), использовано наблюдений - 600
Включено 300 пространственных объектов
Длина временного ряда = 2
Зависимая переменная: performance

	Коэффициент	Ст. ошибка	z	P-значение	
const	40,5705	0,871857	46,53	0,0000	***
attendance	1,07664	0,0371079	29,01	4,40e-185	***

Как и в случае использования модели с фиксированными эффектами, мы можем заключить, что посещение лекций полезно для овладения курсом. В соответствии с полученной оценкой, посещение одной дополнительной лекции увеличивает результат по курсу примерно на 1,08 балла.

Добавим фиктивную переменную временного периода для учета возможных временных эффектов. Результаты для соответствующей спецификации представлены ниже. Легко видеть, что соответствующая переменная не является статистически значимой. К тому же её добавление не оказывает существенного влияния на оценку коэффициента при интересующей нас переменной *attendance*. Поэтому в нашем случае нет необходимости в учете временных эффектов.

Модель 2: Случайные эффекты (GLS), использовано наблюдений - 600

Включено 300 пространственных объектов

Длина временного ряда = 2

Зависимая переменная: *performance*

	Коэффициент	Ст. ошибка	z	P-значение	
const	40,6545	0,882711	46,06	0,0000	***
attendance	1,07293	0,0376277	28,51	7,75e-179	***
dt_2	-0,0886003	0,146151	-0,6062	0,5444	

9.6. Доступный ОМНК для оценивания модели со случайными эффектами

Чтобы понять, как устроена ковариационная матрица вектора случайных ошибок Ω в модели со случайными эффектами, вычислим её элементы. Для этого определим дисперсии ε_{it} и случайных эффектов следующим образом: $\text{var}(\varepsilon_{it}) = \sigma_\varepsilon^2$, $\text{var}(\mu_i) = \sigma_\mu^2$. Элементами ковариационной матрицы Ω являются коэффициенты ковариации:

$$\text{cov}(v_{it}, v_{jp}), \quad i, j = 1, 2, \dots, n, \quad t, p = 1, 2, \dots, T.$$

С учетом наших обозначений:

$$\begin{aligned} \text{cov}(v_{it}, v_{jp}) &= \text{cov}(\mu_i + \varepsilon_{it}, \mu_j + \varepsilon_{jp}) = \\ &= \text{cov}(\mu_i, \mu_j) + \text{cov}(\mu_i, \varepsilon_{jp}) + \text{cov}(\varepsilon_{it}, \mu_j) + \text{cov}(\varepsilon_{it}, \varepsilon_{jp}) = \\ &= \text{cov}(\mu_i, \mu_j) + \text{cov}(\varepsilon_{it}, \varepsilon_{jp}) \end{aligned}$$

Здесь в последнем переходе мы использовали предпосылку №2 модели со случайными эффектами, в силу которой случайные ошибки ε_{it} не коррелированы с μ_i .

Если $i = j$ и $t = p$, то

$$\text{cov}(v_{it}, v_{jp}) = \text{cov}(\mu_i, \mu_i) + \text{cov}(\varepsilon_{it}, \varepsilon_{it}) = \sigma_\mu^2 + \sigma_\varepsilon^2 = \sigma_v^2.$$

Поэтому на главной диагонали ковариационной матрицы вектора случайных ошибок стоят суммы $\sigma_v^2 = \sigma_\mu^2 + \sigma_\varepsilon^2$.

Если $i = j$ и $t \neq p$, то

$$\text{cov}(v_{it}, v_{jp}) = \text{cov}(\mu_i, \mu_i) + \text{cov}(\varepsilon_{it}, \varepsilon_{ip}) = \sigma_\mu^2 + 0 = \sigma_\mu^2.$$

Если же $i \neq j$, то

$$\text{cov}(v_{it}, v_{jp}) = \text{cov}(\mu_i, \mu_j) + \text{cov}(\varepsilon_{it}, \varepsilon_{jp}) = 0 + 0 = 0.$$

Таким образом, ковариационная матрица вектора случайных ошибок представляет собой блочную матрицу размера n на n блоков:

$$\Omega = \begin{pmatrix} \Sigma & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \Sigma & \dots & \mathbf{0} \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \Sigma \end{pmatrix}$$

Здесь $\mathbf{0}$ обозначает нулевую квадратную матрицу размера T на T , а Σ — квадратную матрицу размера t на t , на главной диагонали которой стоят числа $\sigma_\mu^2 + \sigma_\varepsilon^2$, а вне главной диагонали — числа σ_μ^2 :

$$\Sigma = \begin{pmatrix} \sigma_\mu^2 + \sigma_\varepsilon^2 & \sigma_\mu^2 & \dots & \sigma_\mu^2 \\ \sigma_\mu^2 & \sigma_\mu^2 + \sigma_\varepsilon^2 & \dots & \sigma_\mu^2 \\ \vdots & & \ddots & \vdots \\ \sigma_\mu^2 & \sigma_\mu^2 & \dots & \sigma_\mu^2 + \sigma_\varepsilon^2 \end{pmatrix} = \begin{pmatrix} \sigma_v^2 & \sigma_\mu^2 & \dots & \sigma_\mu^2 \\ \sigma_\mu^2 & \sigma_v^2 & \dots & \sigma_\mu^2 \\ \vdots & & \ddots & \vdots \\ \sigma_\mu^2 & \sigma_\mu^2 & \dots & \sigma_v^2 \end{pmatrix}$$

Представим, например, что в выборке имеются данные про три объекта ($n = 3$), и про каждый из них доступна информация за два периода времени $T = 2$. В этом случае

$$\Sigma = \begin{pmatrix} \sigma_\mu^2 + \sigma_\varepsilon^2 & \sigma_\mu^2 \\ \sigma_\mu^2 & \sigma_\mu^2 + \sigma_\varepsilon^2 \end{pmatrix}.$$

И, следовательно, ковариационная матрица вектора случайных ошибок будет иметь вид:

$$\Omega = \begin{pmatrix} \sigma_\mu^2 + \sigma_\varepsilon^2 & \sigma_\mu^2 & 0 & 0 & 0 & 0 \\ \sigma_\mu^2 & \sigma_\mu^2 + \sigma_\varepsilon^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_\mu^2 + \sigma_\varepsilon^2 & \sigma_\mu^2 & 0 & 0 \\ 0 & 0 & \sigma_\mu^2 & \sigma_\mu^2 + \sigma_\varepsilon^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_\mu^2 + \sigma_\varepsilon^2 & \sigma_\mu^2 \\ 0 & 0 & 0 & 0 & \sigma_\mu^2 & \sigma_\mu^2 + \sigma_\varepsilon^2 \end{pmatrix}.$$

С учетом этого результата процедура реализации доступного ОМНК для оценки параметров модели со случайными эффектами устроена так:

1. Находим оценки $\hat{\sigma}_v^2 = \hat{\sigma}_\varepsilon^2 + \hat{\sigma}_\mu^2$ и $\hat{\sigma}_\mu^2$. Зная их, вычисляем оценку ковариационной матрицы вектора случайных ошибок $\hat{\Omega}$.
2. Находим оценку вектора коэффициентов модели со случайными эффектами при помощи доступного ОМНК (см. параграф 5.5):

$$\widehat{\beta^{\text{RE}}} = \begin{pmatrix} \widehat{\beta_1^{\text{RE}}} \\ \widehat{\beta_2^{\text{RE}}} \\ \dots \\ \widehat{\beta_k^{\text{RE}}} \end{pmatrix} = (X^T \widehat{\Omega}^{-1} X)^{-1} X^T \widehat{\Omega}^{-1} y$$

Поясним детали первого пункта описанной процедуры. Оценка ковариационной матрицы может быть вычислена по формуле:

$$\widehat{\Omega} = \begin{pmatrix} \widehat{\Sigma} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \widehat{\Sigma} & & \mathbf{0} \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \widehat{\Sigma} \end{pmatrix}, \text{ где } \widehat{\Sigma} = \begin{pmatrix} \widehat{\sigma}_v^2 & \widehat{\sigma}_\mu^2 & \dots & \widehat{\sigma}_\mu^2 \\ \widehat{\sigma}_\mu^2 & \widehat{\sigma}_v^2 & & \widehat{\sigma}_\mu^2 \\ \vdots & & \ddots & \vdots \\ \widehat{\sigma}_\mu^2 & \widehat{\sigma}_\mu^2 & \dots & \widehat{\sigma}_v^2 \end{pmatrix}.$$

Чтобы определить $\widehat{\Omega}$, нужно получить состоятельные оценки $\widehat{\sigma}_v^2$ и $\widehat{\sigma}_\mu^2$. Это можно сделать по следующим формулам:

$$\widehat{\sigma}_v^2 = \frac{1}{nT-k} \sum_{i=1}^n \sum_{t=1}^T e_{it}^2 \quad (9.6),$$

$$\widehat{\sigma}_\mu^2 = \frac{1}{nT \cdot \frac{T-1}{2} - k} \sum_{i=1}^n \sum_{t=1}^{T-1} \sum_{s=t+1}^T e_{it} e_{is} \quad (9.7).$$

Здесь e_{it} — остатки, полученные в ходе оценки параметров модели со случайными эффектами при помощи обычного МНК (pooled regression).

9.7. Спецификационные тесты

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

Мы обсудили три подхода к оцениванию на панельных данных: регрессия пула, модель с фиксированными эффектами и модель со случайными эффектами. Следовательно, в каждом конкретном случае важно уметь выбирать наилучший подход. Для этого достаточно научиться попарно сравнивать между собой три доступных варианта.

Мы уже умеем выбирать между регрессией пула и моделью с фиксированными эффектами. Для этого используется специальная версия теста на сравнение «короткой» и «длинной» регрессий, которую мы рассмотрели в параграфе 9.2.

Теперь нам нужно научиться делать выбор между моделями со случайными эффектами и с фиксированными эффектами, а также между моделью со случайными эффектами и регрессией пула.

Для первого из указанных выборов используется тест Хаусмана, уже знакомый нам по теме «Инструментальные переменные». Он позволяет проверить выполнение предположения №6 модели со случайными эффектами (так как именно это предположение отличает её от модели с фиксированными эффектами). В рамках теста сравниваются оценки модели с фиксированными эффектами, полученные при помощи внутригруппового преобразования, и оценки модели со случайными эффектами, полученные при помощи доступного ОМНК. Нулевая гипотеза теста Хаусмана в данном случае состоит в том, что оценки доступного ОМНК являются состоятельными. Альтернативная гипотеза — в том, что они несостоятельны.

Тестовая статистика теста Хаусмана имеет вид:

$$(\widehat{\beta}_{FE} - \widehat{\beta}_{RE})' \left(\widehat{V}(\widehat{\beta}_{FE}) - \widehat{V}(\widehat{\beta}_{RE}) \right)^{-1} (\widehat{\beta}_{FE} - \widehat{\beta}_{RE})$$

$\widehat{\beta}_{FE}$ — вектор оценок коэффициентов для модели с фиксированными эффектами, $\widehat{V}(\widehat{\beta}_{FE})$ — оценка ковариационной матрицы этого вектора. Аналогичные величины с индексом RE соответствуют модели со случайными эффектами. Если нулевая гипотеза верна, то данная статистика имеет распределение Хи-квадрат с k степенями свободы.

Если нулевая гипотеза отвергается, то следует заключить, что оценки доступного ОМНК несостоятельны, и сделать выбор в пользу модели с фиксированными эффектами. В противном случае предпочтительной является модель со случайными эффектами.

Тест Бреуша — Пагана позволяет сделать выбор между моделью со случайными эффектами и обычной регрессией пула, которая не учитывает никаких индивидуальных эффектов. Нулевая гипотеза этого теста состоит в том, что дисперсия случайных эффектов σ_μ^2 равна нулю (то есть в том, что все объекты однородны). Если это так, то в применении доступного ОМНК нет необходимости, и можно ограничиться обычным МНК без учета случайных эффектов. Тестовая статистика для этого теста имеет вид:

$$\frac{nT}{2(T-1)} \left(\frac{\sum_{i=1}^n \left(\sum_{t=1}^T e_{it} \right)^2}{\sum_{i=1}^n \left(\sum_{t=1}^T e_{it}^2 \right)} - 1 \right)^2.$$

Здесь e_{it} — остатки регрессии, оцененной обычным МНК. Если верна нулевая гипотеза, то эта статистика имеет распределение Хи-квадрат с одной степенью свободы. Отвержение нулевой гипотезы говорит в пользу модели со случайными эффектами. Не отвержение — в пользу обычной регрессии пула.

Посмотрим, как работают эти тесты в нашем примере про результаты по курсу математического анализа.

Пример 9.4. Отдача от посещения лекций (окончание)

В таблице 9.2 представлены итоги оценивания отдачи от посещения лекции тремя способами: при помощи обычного МНК (регрессия пула), при помощи внутригруппового преобразования (модель с фиксированными эффектами) и при помощи доступного ОМНК (модель со случайными эффектами). Мы получили эти результаты в ходе решения примеров 9.1–9.3 в предыдущих параграфах.

Таблица 9.2. Сводные результаты оценивания отдачи от посещения лекций

Метод оценивания	МНК	Фиксированные эффекты	Случайные эффекты
Коэффициент при переменной <i>attendance</i>	–0,989*** (0,219)	1,119*** (0,036)	1,073*** (0,037)

*Примечание: в скобках под оценками коэффициентов указаны стандартные ошибки. В первом случае использованы состоятельные в условиях гетероскедастичности стандартные ошибки, во втором случае — стандартные ошибки в форме Ареллано. Символ «***» обозначает значимость на однопроцентном уровне.*

Из таблицы легко видеть, что оценки отдачи от посещения при использовании разных подходов могут заметно отличаться друг от друга. Следовательно, важно осуществить спецификационные тесты для выбора подходящей модели.

МНК против фиксированных эффектов:

Расчетное значение F-статистики для проверки гипотезы об отсутствии фиксированных эффектов равно 138,5, что значительно больше критического значения $F(299, 299)$ при однопроцентном уровне значимости. Поэтому и соответствующее Р-значение меньше 0,01. Следовательно, нулевая гипотеза отвергается при уровне значимости 1%, и мы делаем выбор в пользу модели с фиксированными эффектами.

МНК против случайных эффектов:

Расчетное значение статистики Хи-квадрат (1) составляет 191,1, что заметно больше критического значения из таблиц распределения Хи-квадрат для одной степени свободы и уровня значимости 1%. Соответственно, и Р-значение для проверки гипотезы о том, что дисперсия случайных эффектов равна нулю, заметно меньше, чем 0,01. Таким образом, мы отвергаем нулевую гипотезу и делаем выбор в пользу модели со случайными эффектами.

Случайные эффекты против фиксированных эффектов:

Расчетное значение тестовой статистики для проверки гипотезы о состоятельности ОМНК-оценок равно 136,8. Это тоже больше критического значения из таблиц распределения Хи-квадрат для одной степени свободы (в этом тесте число степеней свободы равно количеству регрессоров в модели, в нашем случае это как раз единица) и уровня значимости 1%. Соответственно, и Р-значение для данной гипотезы меньше, чем 0,01. Мы должны заключить, что оценки модели со случайными эффектами несостоятельны, и сделать выбор в пользу модели с фиксированными эффектами.

Подводя итоги всех тестов, мы можем утверждать, что наилучшим выбором является модель с фиксированными эффектами. Таким образом, посещение лекций положительно влияет на сумму баллов по курсу математического анализа.

В заключение отметим, что, если результаты спецификационных тестов противоречивы, следует отдавать предпочтение модели с фиксированными эффектами. Это объясняется тем, что её предпосылки являются наиболее слабыми. Следовательно, она будет давать состоятельные результаты в большем числе случаев, чем альтернативные модели.

Глава 10. Модели бинарного выбора

Модели бинарного выбора полезны, если вы хотите оценить вероятность наступления некоторого события и определить, от чего эта вероятность зависит, или если вы хотите выяснить, какие факторы и каким образом влияют на решения, принимаемые индивидом. Например:

- Какова вероятность наступления рецессии в экономике в следующем году? Какие факторы влияют на эту вероятность?
- Почему одни выпускники школы принимают решение о поступлении в вуз, а другие — нет? Какие факторы влияют на это решение?
- Вернет ли этот заемщик долг или нет?

[10.1. Линейная модель вероятности](#)

[10.2. Логит-модель: введение](#)

[10.3. Логит-модель: оценивание и тестирование гипотез](#)

[10.4. Пробит-модель](#)

[Задания для самостоятельного решения](#)

10.1. Линейная модель вероятности

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

На первый взгляд, если мы столкнулись с одним из вопросов, сформулированных в начале параграфа, мы можем легко обойтись тем инструментарием, который нам уже знаком: взять в качестве зависимой переменной бинарную переменную. В примере с заемщиками мы могли бы собрать данные про n индивидов, взявших кредит в банке, и в качестве зависимой переменной выбрать бинарную переменную, равную единице для тех, кто вернул долг, и нулю для тех, кто не вернул.

В такой модели мы можем считать, что вероятность наступления события (например, возвращения кредита) линейно зависит от некоторого фактора x (например, от заработной платы индивида). Тогда такую модель называют линейной моделью вероятности и записывают следующим образом:

$$p_i = P(y_i = 1) = \beta_1 + \beta_2 x_i,$$

где Y_i — переменная, которая равна единице, если событие наступило, и нулю в противном случае, x_i — фактор, влияющий на вероятность наступления данного события, а $p_i = P(y_i = 1)$ — вероятность того, что событие наступит.

Рассмотрим данную модель на примере влияния времени, затраченного на подготовку к зачету, на вероятность его успешной сдачи. Пусть y_i — переменная, которая равна единице, если i -й студент сдал зачет, и нулю в противном случае, x_i — время, затраченное на подготовку i -м студентом (в часах), а $p_i = P(y_i = 1)$ — вероятность того, что зачет будет сдан успешно.

Представим, что мы собрали данные о двух тысячах студентов, оценили параметры уравнения при помощи обычного МНК и получили следующие результаты (см. таблицу 10.1).

Таблица 10.1. Результаты оценки вероятности сдачи зачета при помощи линейной модели вероятности

Зависимая переменная: y			
	Коэффициент	Ст. ошибка	t-статистика
Константа	-0,30	0,05	-6,00
x	0,10	0,02	5,00

В соответствии с результатами оценивания из таблицы 10.1. можно записать уравнение для предсказанной вероятности сдачи зачета:

$$\hat{p} = -0,3 + 0,1x_i.$$

Это уравнение следует интерпретировать так: если x увеличивается на единицу, то вероятность наступления события увеличивается на 0.1. То есть один дополнительный час подготовки увеличивает вероятность сдать зачет на 0,1 (на 10 процентных пунктов). Скажем, если студент потратил на подготовку 9 часов, то вероятность сдать зачет будет равна

$$\hat{p} = -0,3 + 0,1 * 9 = 0,6$$

На практике линейная модель вероятности используется сравнительно редко, так как обладает недостатками. Главный из них — сложности с интерпретацией результатов: в такой модели предсказанные значения вероятности могут быть отрицательными или превышать единицу, что заведомо не соответствует действительности. В нашем примере, если $x = 2$, то $\hat{p} = -0,1 < 0$, а если $x = 20$, то $\hat{p} = 1,7 > 1$.

Кроме того, нереалистичным часто выглядит и предположение о том, что вероятность успеха зависит от объясняющей переменной линейно. Вряд ли увеличение времени подготовки с 20 до 30 часов должно приводить к такому же изменению вероятности сдачи зачета, как и увеличение с 1000 до 1010 часов.

Чтобы преодолеть эту проблему, в прикладных исследованиях вместо линейной модели вероятности обычно используют одну из двух альтернатив: логит-модель или пробит-модель. Мы рассмотрим их в последующих параграфах этой главы.

10.2. Логит-модель: введение

В рамках логит-анализа для описания вероятности наступления события используется логистическая функция:

$$F(z_i) = \frac{1}{1+e^{-z_i}}.$$

График этой функции представлен на рисунке 10.1. Она всегда принимает значения в пределах от нуля до единицы, что позволяет преодолеть основной недостаток линейной модели вероятности, упомянутый в предыдущем параграфе.

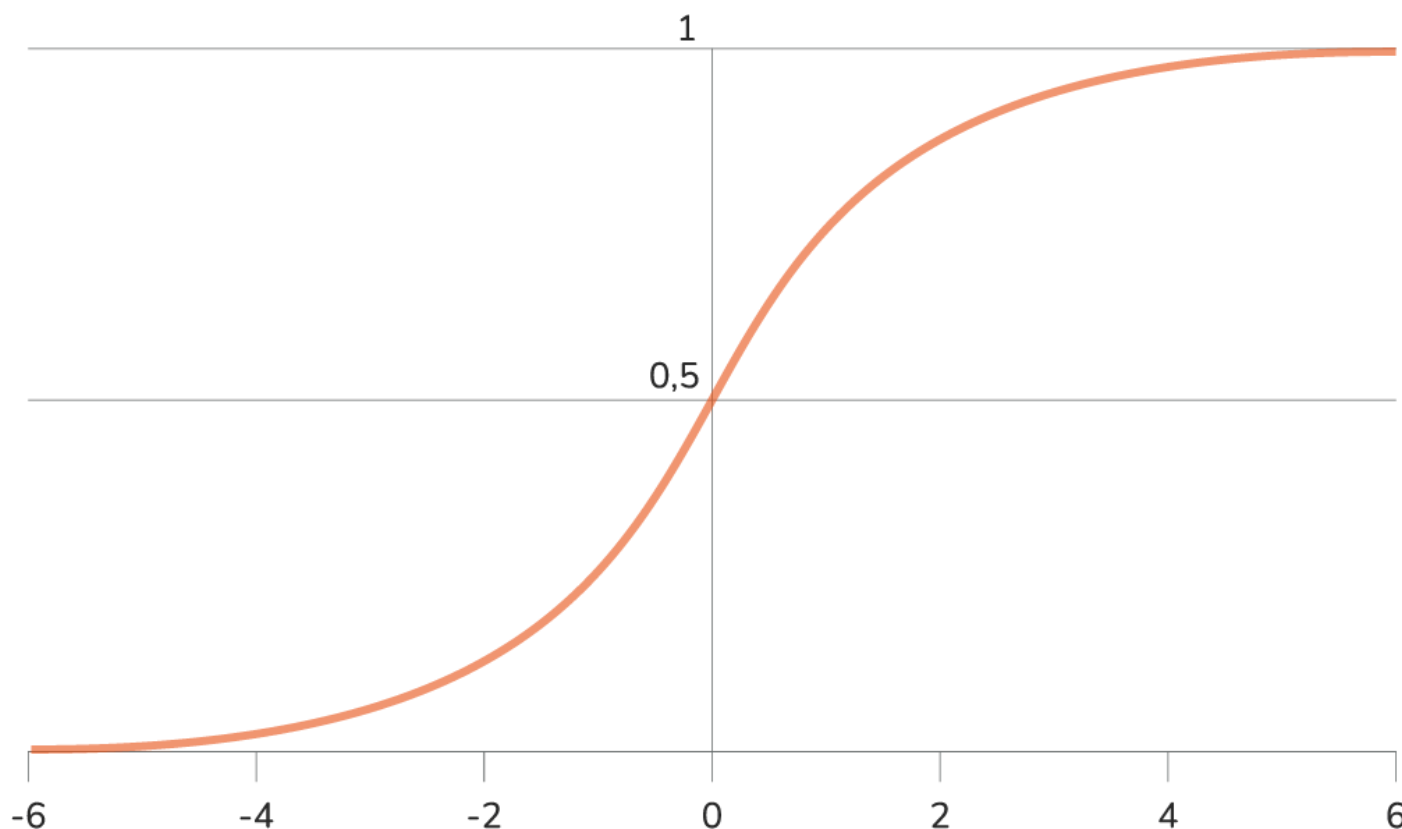


Рисунок 10.1. График логистической функции

Если мы анализируем случай парной взаимосвязи (то есть случай, когда вероятность наступления события зависит от единственного фактора x), то логит-модель может быть записана так:

$$P(y_i = 1) = F(z_i) = \frac{1}{1+e^{-z_i}}, \text{ где } z_i = \beta_1 + \beta_2 x_i.$$

$$P(y_i = 1) = \frac{1}{1+e^{-(\beta_1 + \beta_2 x_i)}}$$

Для оценивания такой модели применить обычный МНК не получится, так как параметры входят в уравнение нелинейно. Кроме того, невозможно свести модель к линейной по параметрам подобно тому, как мы делали это в главе 4 при помощи перехода к логарифмам. Поэтому для оценки логит-модели используется метод максимального правдоподобия (ММП).

С деталями процедуры оценивания можно познакомиться в параграфе 10.3. Здесь мы доверим эту работу эконометрическому пакету, а сами сосредоточимся на том, как можно интерпретировать результаты логит-модели. Для этого вернемся к нашему примеру про вероятность успешной сдачи зачета в зависимости от времени, затраченного на подготовку (см. таблицу 10.2).

Таблица 10.2. Результаты оценки вероятности сдачи зачета при помощи логит-модели

Зависимая переменная: y			
	Коэффициент	Ст. ошибка	z
const	-9,00	0,50	-18,00
x	0,50	0,04	10,00

В соответствии с таблицей 10.2 оценённая вероятность сдачи зачета имеет следующий вид:

$$\hat{P}(y_i = 1) = \frac{1}{1+e^{-(-9+0,5x_i)}}$$

Как можно интерпретировать подобные результаты?

Во-первых, можно оценить вероятность наступления интересующего нас события в тех или иных условиях. Например, если студент затратил на подготовку 15 часов, то какова вероятность сдать зачет? Для ответа на вопрос достаточно подставить в формулу $x_i = 15$ и вычислить вероятность. В нашем случае она равна 0,18. То есть для студента, который готовился 15 часов, вероятность сдать зачет составляет 18%.

Во-вторых, можно интерпретировать результаты в терминах изменения зависимой переменной в результате изменения регрессора. Для этого следует найти так называемый **предельный эффект** изменения регрессора, то есть вычислить, на сколько меняется вероятность наступления события при небольшом изменении переменной x . Для этого посчитаем производную вероятности по x :

$$\frac{dP(y_i=1)}{dx} = \frac{e^{-(\beta_1+\beta_2x)}}{(1+e^{-(\beta_1+\beta_2x)})^2} * \beta_2$$

В нашем примере для студента, который готовился к зачету 15 часов, оценка предельного эффекта составит:

$$\frac{d\hat{P}(y_i=1)}{dx} = \frac{e^{-(-9+0,5*15)}}{(1+e^{-(-9+0,5*15)})^2} * 0,5 = 0,07$$

То есть дополнительный час подготовки для нашего студента увеличит вероятность сдачи зачета примерно на 7 процентных пунктов.

Геометрически предельный эффект (как и любая производная) характеризует наклон функции. Так как у логистической функции разный наклон в разных точках (см. рис. 10.1), то и предельный эффект для разных значений регрессора будет отличаться. Скажем, для студента, который затратил на подготовку не 15 часов, а 100 часов, предельный эффект будет гораздо меньше:

$$\frac{d\hat{P}(y_i=1)}{dx} = \frac{e^{-(-9+0,5*100)}}{(1+e^{-(-9+0,5*100)})^2} * 0,5 = 0,00000000000000000008$$

В нашем примере это, по всей видимости, означает, что студент, который готовится уже целых 100 часов, и так знает всё, что нужно, и имеет вероятность сдачи зачета, близкую к ста процентам. Поэтому от ещё одного дополнительного часа подготовки толку практически не будет.

С прикладной точки зрения непостоянство предельного эффекта порождает некоторую сложность: не очень понятно, в какой именно точке его считать. На практике для этого обычно используется один из двух вариантов:

1. предельный эффект для среднего по выборке. В нашем примере работает следующим образом: вычисляем среднее по выборке время подготовки к зачету $\bar{x}$, а затем считаем предельный эффект в точке $\bar{x}$;
2. средний предельный эффект: вычисляем предельный эффект для каждого студента, затем считаем среднее значение из n предельных эффектов.

Говоря про интерпретацию результатов логит-модели, следует отдельно упомянуть интерпретацию коэффициентов при фиктивных переменных. Представим, например, что в нашей истории про вероятность сдачи зачета регрессором теперь является бинарная переменная. Пусть, например, x — переменная, которая равна единице для тех студентов, которые в прошлом в школе учились в математическом классе (и равна нулю для всех остальных). Представим для определённости, что, оценив логит-модель, мы получили вот такие результаты:

$$\hat{P}(y_i = 1) = \frac{1}{1+e^{-(-2,0+2,0*x)}}$$

Использовать здесь для интерпретации подход с вычислением предельного эффекта — это не слишком удачная идея, так как предельный эффект (как и любая производная) показывает изменение функции при бесконечно малом изменении аргумента. Однако в случае бинарной переменной аргумент не может меняться на бесконечно малую величину: он равен либо нулю, либо единице. Вместо этого удобно вычислить изменение вероятности наступления события в результате изменения регрессора с 0 до 1.

В нашем примере с бинарной переменной математического класса можно посчитать вероятность сдать зачёт для студента не из математического класса (т. е. для студента, для которого x равен 0). Для этого надо просто в формулу вероятности подставить x , равный 0 (получится $\frac{1}{1+e^2}$). А потом можно посчитать вероятность наступления этого события для студента из математического класса (т. е. для студента, для которого x равен 1). Для этого надо подставить 1 в выражение для вероятности (получится $\frac{1}{1+e^0}$). Потом следует просто сравнить эти две вероятности:

$$\hat{p}(x = 1) - \hat{p}(x = 0) = \frac{1}{1+e^0} - \frac{1}{1+e^2} = 0,38$$

Это будет значить, что для выпускников из математической школы по сравнению с выпускниками других школ вероятность получить зачёт на 0,38 больше (т. е. на 38 процентных пунктов больше).

10.3. Логит-модель: оценивание и тестирование гипотез

Вероятность наступления события в логит-модели описывается функцией:

$$P(y_i = 1) = F(z_i) = \frac{1}{1 + e^{-(\beta_1 + \beta_2 z_i)}}$$

Следовательно, вероятность ненаступления события имеет вид:

$$\begin{aligned} P(y_i = 0) &= 1 - P(y_i = 1) = \\ &= 1 - F(z_i) = 1 - \frac{1}{1 + e^{-(\beta_1 + \beta_2 z_i)}} \end{aligned}$$

С учетом этих соображений можем записать функцию правдоподобия:

$$\begin{aligned} L(y_1, \dots, y_n) &= \prod_{y_i=1} P(y_i = 1) * \prod_{y_i=0} P(y_i = 0) = \\ &= \prod_{y_i=1} \frac{1}{1 + e^{-(\beta_1 + \beta_2 z_i)}} * \prod_{y_i=0} \left(1 - \frac{1}{1 + e^{-(\beta_1 + \beta_2 z_i)}}\right) = \\ &= \prod_i \left(\frac{1}{1 + e^{-(\beta_1 + \beta_2 z_i)}}\right)^{y_i} * \left(1 - \frac{1}{1 + e^{-(\beta_1 + \beta_2 z_i)}}\right)^{1-y_i} \end{aligned}$$

Логарифмируя это выражение, получаем логарифм функции правдоподобия:

$$\begin{aligned} \ln L(y_1, \dots, y_n) &= \\ &= \sum_{i=1}^n y_i * \ln \frac{1}{1 + e^{-(\beta_1 + \beta_2 z_i)}} + \sum_{i=1}^n (1 - y_i) * \ln \left(1 - \frac{1}{1 + e^{-(\beta_1 + \beta_2 z_i)}}\right) \end{aligned}$$

Далее для получения оценок параметров достаточно вычислить производные по β_1 и β_2 и приравнять их к нулю. Решение соответствующей системы относительно неизвестных значений параметров и приводит к получению необходимых оценок $\hat{\beta}_1$ и $\hat{\beta}_2$.

Так как данная система является нелинейной, у неё может не быть аналитического решения, поэтому не получится (подобно случаю применения МНК) записать формулы оценок коэффициентов в общем виде. Эконометрические пакеты осуществляют решение системы не аналитически, а численными методами. Функция правдоподобия будет выпуклой вверх, поэтому в найденной точке будет достигаться её максимум.

В силу общих свойств метода максимального правдоподобия, если спецификация модели верна, то полученные ММП-оценки параметров будут состоятельными и асимптотически нормальными. Последнее свойство позволяет тестировать гипотезы по поводу отдельных коэффициентов в логит-модели стандартным образом: вычисляя отношение оценки коэффициента к его стандартной ошибке и используя тот факт, что это отношение имеет стандартное нормальное распределение (см. пример 10.1).

Тестировать гипотезы по поводу выполнения нескольких ограничений (например, для сравнения «короткой» и «длинной» регрессий) можно при помощи теста отношения правдоподобия (likelihood ratio test). Для этого следует оценить параметры модели без ограничений и модели с ограничением и вычислить следующее расчетное значение тестовой статистики:

$$LR = -2(\ln L_R - \ln L_{UR}).$$

Здесь $\ln L_{UR}$ — логарифм максимального значения эмпирической функции правдоподобия в регрессии без ограничений;

$\ln L_R$ — логарифм максимального значения эмпирической функции правдоподобия в регрессии с ограничениями.

Если верна нулевая гипотеза, то эта статистика имеет асимптотическое распределение $\chi^2(q)$, где q — число ограничений.

В этом и предыдущем параграфах мы рассматривали случай анализа парной взаимосвязи. В действительности на вероятность наступления того или иного события могут влиять сразу много факторов.

В этом случае логит-модель оценивается и интерпретируется аналогичным образом с той лишь разницей, что теперь z_i зависит не от единственной переменной, а от произвольного их количества:

$$P(Y_i = 1) = F(z_i) = \frac{1}{1 + e^{-z_i}} = \frac{1}{1 + e^{-(\beta_1 + \beta_2 * z_i^{(2)} + \dots + \beta_k * z_i^{(k)})}},$$

$$z_i = \beta_1 + \beta_2 * x_i^{(2)} + \dots + \beta_k * x_i^{(k)}$$

Так как параметры логит-модели оцениваются не методом наименьших квадратов, то стандартный коэффициент R-квадрат в данном случае неприменим. Зато можно использовать для анализа степени соответствия модели данным некоторые специфические именно для моделей бинарного выбора характеристики:

1. Доля правильно предсказанных исходов. При этом в стандартном случае, если предсказанная вероятность наступления события не меньше 0,5 ($\hat{P}(y_i = 1) \geq 0,5$), то предсказанным исходом считается наступление события. Если же соответствующая вероятность меньше половины, то предсказанным исходом считается ненаступление события. Естественно, более предпочтительной является модель, которая лучше предсказывает исходы.
2. Коэффициент корреляции между исходами и предсказанными вероятностями.
3. Коэффициент псевдо- R^2 , называемый также R-квадрат МакФаддена. Он определяется следующим образом:

$$pseudo-R^2 = 1 - \frac{\ln(L)}{\ln(L_0)}$$

$\ln(L)$ — логарифм функции правдоподобия в модели, для которой мы вычисляем $pseudo-R^2$. $\ln(L_0)$ — логарифм функции правдоподобия в модели, содержащей только константу.

Подобно обычному коэффициенту R-квадрат этот коэффициент лежит в пределах от нуля до единицы (и равен нулю для модели, включающей только константу). Чем лучше модель соответствует данным, тем ближе к единице он окажется.

Пример 10.1. Вероятность приёма на работу

100 кандидатов проходили собеседование о приеме на работу в крупную компанию. Известны следующие данные о кандидатах:

x — стаж работы кандидата (в годах),

gender — бинарная переменная, равная единице для кандидатов-мужчин и равная нулю для кандидатов-женщин,

black — бинарная переменная, равная единице для кандидатов-афро-американцев и нулю для всех остальных кандидатов,

y — бинарная переменная, равная единице для тех кандидатов, которые были приняты на работу.

Результаты оценивания трех моделей на основе доступных данных представлены в таблицах ниже.

(а) Заполните пропуски в таблице 1, вычислив значение коэффициента R-квадрат МакФаддена для всех моделей.

(б) Для модели №3 проверьте значимость уравнения в целом, используя тест отношения правдоподобия.

(в) Сравните модель №2 и модель №3, используя тест отношения правдоподобия.

(г) Для модели №2 интерпретируйте полученный результат, вычислив предельный эффект стажа работы для среднего по выборке работника.

(д) Для модели №2 интерпретируйте полученный результат, вычислив средний предельный эффект стажа работы.

Таблица 10.3. Модели вероятности принятия кандидата на работу

Зависимая переменная: y Метод оценивания: логит-модель			
	Модель 1	Модель 2	Модель 3
x	—	0,49 (0,03)	0,49 (0,15)
Gender	—	—	0,15 (0,43)
Black	—	—	-0,32 (0,43)
Constant	-0,32 (0,20)	-1,02 (0,15)	-0,90 (0,42)
Логарифм функции правдоподобия	-68,0	-62,0	-61,0
R-квадрат МакФаддена			

Таблица 10.3.б. Стаж работы и результаты отбора кандидатов

Стаж работника	Количество кандидатов с таким стажем работы	Количество принятых на работу
0 лет	40	10
1 год	25	10
2 года	10	5
3 года	10	6
4 года	10	7
5 лет	5	4

Решение:

(а) Для каждой из моделей вычислим $pseudo-R^2$:

$$\text{Модель 1: } pseudo-R^2 = 1 - \frac{\ln(L_0)}{\ln(L_1)} = 0$$

$$\text{Модель 2: } pseudo-R^2 = 1 - \frac{62}{68} = 0,09$$

$$\text{Модель 3: } pseudo-R^2 = 1 - \frac{61}{68} = 0,10$$

(б) Для тестирования значимости уравнения в целом нам нужно проверить гипотезу о том, что коэффициенты при всех трех переменных (стаж, пол и раса кандидата) равны нулю. Для этого нам нужно знать значение логарифма функции правдоподобия в модели, в которой нет всех трех этих переменных, то есть в модели, в которую включена только константа.

$$LR = -2(\ln L_R - \ln L_{UR}) = -2(-68 + 61) = 14$$

Критическое значение тестовой статистики Хи-квадрат(3) при уровне значимости 5% равно 7,81. Расчетное значение больше критического, следовательно, гипотеза о равенстве нулю коэффициентов при всех переменных отклоняется. Делаем вывод о том, что уравнение в целом значимо.

$$\text{(в) } LR = -2(\ln L_R - \ln L_{UR}) = -2(-62 + 61) = 2$$

Критическое значение тестовой статистики Хи-квадрат(2) при уровне значимости 5% равно 5,99. Расчетное значение меньше критического, следовательно, гипотеза о равенстве нулю коэффициентов при добавленных переменных **не отвергается**. Делаем вывод о том, что добавление переменных **не** оправдано. Модель №2 является предпочтительной.

(г) Для начала отметим, что коэффициент при анализируемой переменной статистически значим при уровне значимости 5% (0,49/0,03 > 1,96). Средний по выборке стаж работника составляет:

$$\bar{x} = \frac{40 \cdot 0 + 25 \cdot 1 + 10 \cdot 2 + 10 \cdot 3 + 10 \cdot 4 + 5 \cdot 5}{100} = 1,4$$

Предельный эффект в этой точке равен:

$$\frac{d\hat{p}}{dx} = \frac{e^{-(\hat{\beta}_1 + \hat{\beta}_2 x)}}{(1 + e^{-(\hat{\beta}_1 + \hat{\beta}_2 x)})^2} * \hat{\beta}_2 = \frac{e^{-(-1,02 + 0,49 \cdot 1,4)}}{(1 + e^{-(-1,02 + 0,49 \cdot 1,4)})^2} * 0,49 = 0,12$$

Таким образом, для среднего по выборке кандидата один дополнительный год опыта работы увеличивает вероятность оказаться нанятым на работу примерно на 12 процентных пунктов.

(д) Вычисляем предельный эффект для каждого стажа работы, который есть в выборке:

$$x = 0, \frac{d\hat{p}}{dx} = \frac{e^{-(\hat{\beta}_1 + \hat{\beta}_2 x)}}{(1 + e^{-(\hat{\beta}_1 + \hat{\beta}_2 x)})^2} * \hat{\beta}_2 = \frac{e^{-(-1,02 + 0,49 \cdot 0)}}{(1 + e^{-(-1,02 + 0,49 \cdot 0)})^2} * 0,49 = 0,095$$

$$x = 1, \frac{d\hat{p}}{dx} = \frac{e^{-(\hat{\beta}_1 + \hat{\beta}_2 x)}}{(1 + e^{-(\hat{\beta}_1 + \hat{\beta}_2 x)})^2} * \hat{\beta}_2 = \frac{e^{-(-1,02 + 0,49 \cdot 1)}}{(1 + e^{-(-1,02 + 0,49 \cdot 1)})^2} * 0,49 = 0,114$$

$$x = 2, \frac{d\hat{p}}{dx} = 0,122$$

$$x = 3, \frac{d\hat{p}}{dx} = 0,117$$

$$x = 4, \frac{d\hat{p}}{dx} = 0,099$$

$$x = 5, \frac{d\hat{p}}{dx} = 0,076$$

Вычисляем средний предельный эффект:

$$\frac{40 \cdot 0,095 + 25 \cdot 0,114 + 10 \cdot 0,122 + 10 \cdot 0,117 + 10 \cdot 0,099 + 5 \cdot 0,076}{100} = 0,10$$

Таким образом, средний предельный эффект стажа работы составляет примерно 10 процентных пунктов.

10.4. Пробит-модель

Альтернативным способом оценивать параметры модели бинарного выбора является пробит-модель.

Её отличие от логит-модели состоит в том, что вместо логистической функции для описания вероятности наступления события используется функция стандартного нормального распределения.

$$P(Y_i = 1) = \Phi(z_i) = \Phi(\beta_1 + \beta_2 * x_i^{(2)} + \dots + \beta_k * x_i^{(k)}),$$

$$z_i = \beta_1 + \beta_2 * x_i^{(2)} + \dots + \beta_k * x_i^{(k)}$$

Здесь $\Phi(z_i)$ — функция стандартного нормального распределения.

Оценивание, тестирование гипотез и интерпретация результатов в рамках пробит-анализа проводится полностью аналогично случаю логит-анализа с поправкой на использование функции стандартного нормального распределения вместо логистической функции. В частности, предельный эффект изменения переменной $x^{(j)}$ может быть вычислен вот так:

$$\begin{aligned} \frac{\partial P(Y_i=1)}{\partial x^{(j)}} &= \Phi'(\beta_1 + \beta_2 * x_i^{(2)} + \dots + \beta_k * x_i^{(k)}) * \beta_j = \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{(\beta_1 + \beta_2 * x_i^{(2)} + \dots + \beta_k * x_i^{(k)})^2}{2}} * \beta_j. \end{aligned}$$

Здесь $\Phi'(\bullet)$ — производная от функции стандартного нормального распределения, то есть функция плотности вероятности стандартного нормального распределения. Аналогично тому, что мы обсуждали в параграфе 10.2, на практике этот предельный эффект часто вычисляют в точке средних значений регрессоров.

При значениях аргумента, не слишком далеких от нуля, логистическая функция и функция стандартного нормального распределения ведут себя сходным образом, поэтому предельные эффекты, оцененные с использованием логит- и пробит-моделей оказываются примерно одинаковыми. Это делает выбор между двумя указанными подходами непринципиальным. В случае, если результаты их применения все-таки отличаются, можно выбирать модель, характеризующуюся более высоким значением функции правдоподобия и лучшими значениями характеристик качества подгонки модели, которые мы обсудили в параграфе 10.3.

Пример 10.2. Вероятность сдачи зачета

На основе пробит-модели бинарного выбора исследователь анализирует вероятность сдать зачет по некоторому курсу. Исследователь собрал данные о 250 студентах, сдававших зачет: Pass — бинарная переменная, равная единице, если студент сдал зачет; Lectures — количество посещенных студентом лекций по курсу; Male — фиктивная переменная, равная единице для мужчин и нулю для женщин.

В таблице представлены результаты оценивания модели:

Dependent Variable: Pass	
	Probit
Lectures	0,20 (0,03)
Male	- 0,50 (0,02)
Lectures *Male	- 0,05 (0,02)
Constant	- 1,00 (0,12)

(а) Какие коэффициенты в модели являются значимыми?

(б) Иван и Дарья посетили по 10 лекций, но пропустили зачет и будут писать его позже. Используя оцененную модель, вычислите вероятность сдать зачет для каждого из них.

(в) Вычислите предельный эффект посещения лекций для Дарьи.

Решение:

(а) Для всех коэффициентов отношение оценки коэффициента к стандартной ошибке по модулю больше, чем 1,96. Поэтому все коэффициенты статистически значимо отличаются от нуля при уровне значимости 5%.

(б) Для Ивана: $z = 0,2 * 10 - 0,5 * 1 - 0,05 * 10 - 1 = 0$. $\Phi(z) = \Phi(0) = 0,5$. Вероятность сдать зачет равна 50% (здесь $\Phi(z)$ — функция стандартного нормального распределения).

Для Дарьи: $z = 0,2 \cdot 10 - 0,5 \cdot 0 - 0,05 \cdot 0 - 1 = -1$. $\Phi(z) = \Phi(1) = 0,84$. Вероятность сдать зачет равна 84%.

(в) $\frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \cdot 0,2 = \frac{1}{\sqrt{2\pi}} e^{-\frac{1^2}{2}} \cdot 0,2 = 0,24 \cdot 0,2 = 0,048$. Одна дополнительная лекция увеличит вероятность получения зачета примерно на 5 процентных пунктов.

Глава 11. Оценка эффекта воздействия

Часто эконометристу требуется оценить последствия реализации некоторой политики (осуществить policy evaluation). Например, выяснить:

- как изменится занятость в результате введения закона о минимальной заработной плате?
- как скажется на уровне инфляции переход к политике инфляционного таргетирования?
- увеличится ли объем продаж товара благодаря реализации его рекламной компании?
- будет ли уменьшение размера класса способствовать увеличению успеваемости школьников?
- станут ли работники лучше работать, если пройдут курсы повышения квалификации?

Во всех примерах мы сталкиваемся с необходимостью определения эффекта воздействия (treatment effect) некоторой политики на зависимую переменную. Неважно, идет ли речь о монетарной политике, рекламной политике или о каком-либо другом вмешательстве, последствия которого нам интересны.

Если в подобной ситуации есть возможность осуществить контролируемый эксперимент, то получение ответа на любой из указанных вопросов с эконометрической точки зрения становится достаточно простой задачей (мы упоминали это в параграфе 1.3, освежите его в памяти). Поэтому эконометристы любят использовать эксперименты. И поэтому мы обсудим соответствующую методологию вместе со всеми необходимыми определениями в параграфе 11.1.

Подчеркнем, что в параграфе 11.1 вы встретите не новые эконометрические методы, а, скорее новый, способ думать о выявлении причинно-следственных связей.

Так как проведение экспериментов может быть довольно дорогим (а иногда и невозможным) занятием, то на практике для оценки эффектов воздействия часто приходится довольствоваться историческими данными. Для этого разработан ряд специальных методов, которые мы обсудим в параграфах 11.2–11.4. Их общей чертой является попытка преобразовать исторические данные таким образом, чтобы максимально приблизить их к условиям идеального контролируемого эксперимента.

[11.1. Оценка эффекта воздействия в идеальном эксперименте](#)

[11.2. Оценка эффекта воздействия методом разности разностей](#)

[11.3. Локальный средний эффект воздействия \(LATE\)](#)

[11.4. Разрывный регрессионный дизайн](#)

[Задания для самостоятельного решения](#)

11.1. Оценка эффекта воздействия в идеальном эксперименте

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

Начнем с того, что договоримся о терминах. В качестве примера при обсуждении терминологии мы будем использовать гипотетическое исследование, в котором анализируется эффективность лечения. Представим, что мы рассматриваем несколько сотен индивидов, часть из которых подверглась госпитализации и лечению, а часть — нет. И нас интересует причинно-следственная связь между госпитализацией индивида и его уровнем здоровья (будем считать, что мы умеем измерять уровень здоровья численно).

В общем случае группа объектов, которая подверглась воздействию, называется **испытуемой группой** (treatment group). А группа, которая ему не подвергалась — **контрольной группой** (control group).

В нашем примере испытуемая группа — это те, кто был госпитализирован, а контрольная — это все остальные.

Для обозначения принадлежности объекта к той или иной группе будем использовать бинарную переменную D_i .

$D_i = 1$, если i -ый объект вошел в группу, подвергшуюся воздействию (treatment group). В нашем примере $D_i = 1$, если i -ый индивид был госпитализирован. Уровень здоровья i -го индивида в этом случае будем обозначать $Y_i(1)$.

$D_i = 0$, если i -ый объект вошел в контрольную группу (control group). Например, если индивид не был госпитализирован. Уровень здоровья i -го индивида в этом случае будем обозначать $Y_i(0)$.

Тогда изменение здоровья индивида в результате госпитализации можно определить так:

$$Y_i(1) - Y_i(0)$$

Эта величина называется эффект воздействия (treatment effect или causal effect) для i -го индивида. В нашем примере treatment effect для некоторого индивида — это величина, на которую изменится уровень его здоровья, если его подвергнуть лечению, по сравнению со случаем, если его не лечить.

Обратите внимание, что эффект воздействия зависит от индекса i , то есть может быть разным для различных индивидов. Если это так, то эффект воздействия называется гетерогенным. Подобная гетерогенность эффекта — достаточно естественная предпосылка во многих ситуациях. Например, для болеющего человека уровень здоровья сильно зависит от того, подвергнется он лечению или нет, а для человека с крепким здоровьем госпитализация не будет существенно влиять на самочувствие.

Если усреднить эффект воздействия по всем индивидам из генеральной совокупности, то мы получим так называемый **средний эффект воздействия** (average treatment effect, ATE). Примеры ATE:

- На сколько в среднем увеличивается здоровье индивидов в результате их госпитализации?
- На сколько в среднем увеличится успеваемость школьников, если обучать их в маленьком классе вместо класса нормальной величины?
- На сколько в среднем изменится занятость в ресторанах быстрого питания в результате принятия закона о минимальной заработной плате?

Чтобы посчитать эффект воздействия, нужно вычислить разность $Y_i(1) - Y_i(0)$. На практике это невозможно, так как ни для одного объекта мы не наблюдаем одновременно $Y_i(1)$ и $Y_i(0)$. Мы наблюдаем либо одно, либо другое. Действительно, данный конкретный индивид либо госпитализирован, либо нет. Если индивид госпитализирован, то мы наблюдаем $Y_i(1)$. При этом мы не знаем достоверно, что было бы, если индивида не лечили, то есть мы не наблюдаем $Y_i(0)$. Два эти значения называются потенциальными исходами (potential outcomes)

Уровень здоровья, который мы фактически наблюдаем — это и есть доступное нам в данных значение зависимой переменной для этого индивида (в англоязычной литературе его называют observed outcome):

$$Y_i = \begin{cases} Y_i(1), & \text{если } D_i = 1 \\ Y_i(0), & \text{если } D_i = 0 \end{cases}$$

Иногда удобно записывать Y_i следующим образом:

$$Y_i = Y_i(0) + D_i * (Y_i(1) - Y_i(0))$$

Путем непосредственной подстановки нашей бинарной переменной легко убедиться, что обе записи эквивалентны.

Так как мы не можем непосредственно вычислить эффект воздействия для отдельного объекта $Y_i(1) - Y_i(0)$, то мы не можем вычислить и его математическое ожидание $E(Y_i(1) - Y_i(0))$, то есть ATE. Мы не сможем его вычислить, даже если представить, что нам доступны данные по абсолютно всем объектам из генеральной совокупности.

Вместо этого мы можем попытаться оценить этот эффект, используя наблюдаемые данные.

В нашем примере можно попробовать оценить эффект от лечения, сопоставив ожидаемые уровни здоровья тех, кто был госпитализирован, с ожидаемым уровнем здоровья всех остальных. Для этого нужно вычислить такую величину:

$$E(Y_i | D_i = 1) - E(Y_i | D_i = 0)$$

$E(Y_i | D_i = 1)$ — ожидаемое значение зависимой переменной для объектов, которые подверглись воздействию.

$E(Y_i | D_i = 0)$ — ожидаемое значение зависимой переменной для объектов, которые не подвергались воздействию.

Чтобы выяснить, как эта разница математических ожиданий соотносится с интересующей нас величиной АТЕ, осуществим такие преобразования:

$$\begin{aligned} E(Y_i|D_i = 1) - E(Y_i|D_i = 0) &= \\ E(Y_i(1)|D_i = 1) - E(Y_i(0)|D_i = 0) &= \\ E(Y_i(1)|D_i = 1) - E(Y_i(0)|D_i = 1) + \\ + E(Y_i(0)|D_i = 1) - E(Y_i(0)|D_i = 0) &= \\ \underbrace{E(Y_i(1) - Y_i(0)|D_i = 1)}_{ATET} + \underbrace{E(Y_i(0)|D_i = 1) - E(Y_i(0)|D_i = 0)}_{selectionbias} \end{aligned}$$

Последнее выражение состоит из двух слагаемых:

- $E(Y_i(1) - Y_i(0)|D_i = 1)$ — это средний эффект воздействия для индивидов, которые подверглись воздействию (average treatment effect on the treated, ATET)
- $E(Y_i(0)|D_i = 1) - E(Y_i(0)|D_i = 0)$ — это выражение называют **смещением из-за самоотбора** (selection bias). Первое слагаемое здесь — ожидаемый уровень здоровья госпитализированных людей ($D_i = 1$), если бы они не отправились лечиться ($Y_i(0)$). Второе слагаемое — ожидаемый уровень здоровья людей, которые не пошли лечиться.

Таким образом, исходная разность математических ожиданий может быть записана так:

$$E(Y_i|D_i = 1) - E(Y_i|D_i = 0) = ATET + selectionbias$$

Если в нашем примере про больницы предположить, что люди сами решают, идти им лечиться или нет, то естественно ожидать отрицательного смещения из-за самоотбора. Потому что лечиться в этом случае отправятся люди с низким уровнем здоровья (то есть люди с низкими значениями $Y_i(0)$). Следовательно, разность математических ожиданий не будет равна интересующему нас эффекту воздействия.

Если же предположить, что индивиды случайным образом распределяются между испытуемой и контрольной группой в ходе контролируемого эксперимента, то наш вывод поменяется. Действительно, в условиях случайного распределения по группам (random assignment) попадание объекта в ту или иную группу не будет зависеть от его характеристик. В терминах математических ожиданий это будет означать, что

$$E(Y_i(0)|D_i = 1) = E(Y_i(0)|D_i = 0) = E(Y_i(0))$$

В такой ситуации смещение из-за самоотбора отсутствует:

$$selectionbias = E(Y_i(0)|D_i = 1) - E(Y_i(0)|D_i = 0) = 0.$$

Следовательно, разность условных математических ожиданий равна интересующему нас среднему эффекту воздействия:

$$E(Y_i|D_i = 1) - E(Y_i|D_i = 0) = ATET$$

Поскольку в соответствии с законом больших чисел математические ожидания могут быть состоятельно оценены средними, то состоятельная оценка среднего эффекта воздействия может быть вычислена следующим образом:

$$\bar{Y}_1 - \bar{Y}_0 = \widehat{ATET}.$$

Здесь $\bar{Y}_1$ — среднее по выборке значение зависимой переменной для объектов, попавших в испытуемую группу (в нашем примере это средний уровень здоровья для индивидов, подвергшихся госпитализации). $\bar{Y}_0$ — среднее по выборке значение зависимой переменной для объектов, попавших в контрольную группу.

Ещё раз подчеркнем, что состоятельная оценка эффекта воздействия возможна только в условиях, когда смещение из-за самоотбора отсутствует. Контролируемый эксперимент — это ситуация, когда объекты независимо от своих характеристик случайным образом разделяются на две группы (испытуемую и контрольную). Следовательно, контролируемый эксперимент гарантирует отсутствие смещения из-за самоотбора. Это и обеспечивает фундаментальное теоретическое основание для использования экспериментальных данных с целью выявления причинно-следственных связей.

Оценка эффекта воздействия может быть получена и при помощи обычной парной регрессии. Для этого нужно оценить параметры модели:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 * D_i$$

Вспомнив формулы для обычного МНК, можно доказать (см. соответствующее задание в конце главы), что в этом случае МНК-оценка коэффициента при переменной будет в точности равна оценке интересующего нас среднего эффекта воздействия:

$$\hat{\beta}_2 = \bar{Y}_1 - \bar{Y}_0 \quad (11.1)$$

Если эксперимент построен корректно, то в обычной парной регрессии объясняющая переменная является экзогенной. Это возможно благодаря случайному распределению объектов по группам (благодаря random assignment). Поэтому обычная парная регрессия дает несмещенную и состоятельную оценку среднего эффекта воздействия. Следовательно, в регрессии не обязательно использовать контрольные переменные.

Тем не менее, есть две причины, по которым их использование все-таки может быть полезно:

1. Увеличение точности оценивания: включение контрольных переменных позволяет лучше описать зависимую переменную, снизить стандартную ошибку регрессии и получить более точные оценки коэффициентов.
2. Проверка качества рандомизации: если эксперимент построен правильно, и бинарная переменная D действительно экзогенна, то оценки коэффициента при этой переменной в парной и во множественной регрессии не должны сильно отличаться (так как обе оценки являются состоятельными).

Пример 11.1. Эксперимент STAR

В качестве примера использования экспериментальных данных в эконометрике, рассмотрим проект STAR (Student/Teacher Achievement Ratio), который был реализован в США в 80-х годах. Его авторы поставили перед собой цель выяснить, помогает ли обучение в меньших группах достигать больших академических успехов.

Этот вопрос важен не только с точки зрения образования, но и с экономической точки зрения, так как, чтобы учить школьников в маленьких классах, необходимо больше учителей, и, значит, для их оплаты необходимо больше бюджетных денег.

Следовательно, нужно собирать больше налогов или отвлекать ресурсы от чего-то другого. Поэтому важно понять, есть ли какая-то существенная польза от уменьшения численности стандартного школьного класса.

Для того чтобы это выяснить, исследователям пришлось потратить 4 года и порядка 12 млн. долларов (и это в ценах 80-х годов) для проведения эксперимента. В нём было задействовано несколько тысяч американских школьников. Когда они поступали в начальную школу, их случайным образом распределяли по классам разного типа: некоторые классы имели стандартный размер (22-25 человек), а некоторые — уменьшенный (13-17 человек).

С деталями исследования вы можете познакомиться в статье (Kueger, 1999). Здесь мы сконцентрируемся на результатах расчетов для второклассников. Они представлены в таблице 11.1.

Зависимая переменная — результаты стандартизированного письменного теста, который проводился в конце каждого из четырех лет обучения (Stanford Achievement Test). То есть в конце каждого года обучения школьники во всех школах писали единый тест, что обеспечивало сравнимость их академических успехов.

Переменная интереса — это бинарная переменная, которая равна единице для школьников, попавших в испытываемую группу (treatment group), то есть для школьников, обучавшихся в маленьком классе (в таблице она обозначена как «Маленький класс»).

В таблице 11.1 вы также можете увидеть прочие переменные, которые были включены в те или иные спецификации модели:

- Класс с дополнительной помощью — бинарная переменная, которая равна единице для школьников, обучавшихся в классе, где были доступны дополнительные консультации сверх обычных занятий. В ходе эксперимента такие классы тоже определялись случайным образом.
- Белая/ Азиатская раса — бинарная переменная, равная единице для школьников, относящихся к одной из указанных рас.
- Женщина — бинарная переменная, равная единице для девочек и нулю для мальчиков.
- Право на бесплатный обед — бинарная переменная, равная единице для школьников, которые имели право на бесплатные ланчи. Эта переменная является замещающей переменной для дохода семьи, так как такое право получали школьники из малоимущих семей.
- Белый учитель — бинарная переменная, равная единице для школьников, которых обучал учитель указанной расы.
- Учитель-мужчина — это, как нетрудно догадаться, снова бинарная переменная, которая равна единице для школьников, которых обучал учитель-мужчина.
- Опыт учителя и степень магистра — это переменные, характеризующие стаж учителя (в годах) и наличие у него степени магистра.
- Кроме того, в некоторые спецификации включались бинарные переменные принадлежности к определенной школе (в исследовании приняли участие школьники из нескольких десятков школ).

Судя по результатам оценивания, случайное распределение школьников по классам было осуществлено корректно: попадание в тот или иной класс не коррелировано с прочими контрольными переменными. Такой вывод можно сделать в силу того, что изменения набора контрольных переменных не оказывают существенного влияния на коэффициент при переменной интереса. Во всех спецификациях этот коэффициент приблизительно равен 6.

Можно видеть, что эффект воздействия от попадания в маленький класс устойчив к выбору спецификации: он статистически значим на однопроцентном уровне во всех моделях. Таким образом, можно заключить, что обучение в маленьком классе увеличивает результаты школьника на итоговом тесте примерно на 6 баллов.

Таблица 11.1. Моделирование воздействия размера класса на результаты итогового теста

Объясняющая переменная	(1)	(2)	(3)	(4)
Маленький класс	5,93 (1,97)	6,33 (1,29)	5,83 (1,23)	5,79 (1,28)
Класс с дополнительной помощью	1,97 (2,05)	1,88 (1,10)	1,64 (1,07)	1,58 (1,06)
Белая/ Азиатская раса	—	—	6,35 (1,20)	6,36 (1,19)

Объясняющая переменная	(1)	(2)	(3)	(4)
Женщина	—	—	3,48 (0,60)	3,45 (0,60)
Право на бесплатный обед	—	—	-13,61 (0,72)	-13,61 (0,72)
Белый учитель	—	—	—	0,39 (1,75)
Учитель-мужчина	—	—	—	1,32 (3,96)
Опыт учителя	—	—	—	0,10 0,06
Степень магистра	—	—	—	-1,06 (1,06)
Фиксированные эффекты школ	Нет	Да	Да	Да
R ²	0,01	0,22	0,28	0,28

Примечания: В таблице приведены результаты МНК-оценки модели для второклассников. Зависимая переменная — балл школьника за стандартизированный тест. Во всех моделях кроме перечисленных переменных включена константа, которая не приводится для экономии места. В скобках под оценками коэффициентов указаны робастные стандартные ошибки. Число наблюдений равно 5950. Источник: (Kueger, 1999).

11.2. Оценка эффекта воздействия методом разности разностей

[Внешняя ссылка]: [YouTube](#), [RuTube](#)

Описание метода разность разностей удобно сразу осуществить на примере конкретного исследования. Для этого мы воспользуемся работой Карда и Крюгера (Card, Krueger, 1994).

В 1992 году в штате Нью-Джерси минимальный размер оплаты труда был увеличен с 4,25 до 5,05 долларов. Экономическая теория подсказывает, что подобное решение должно сказаться на занятости работников с низкой квалификацией (ведь именно их труд часто оплачивается по минимальной ставке). Эту гипотезу решили проверить Кард и Крюгер в своей работе. Они собрали данные о занятости работников в ресторанах быстрого питания. Таким образом, отдельный объект в их выборке — это один ресторан быстрого питания. А зависимая переменная — число работников, занятых в этом ресторане полный рабочий день.

Таким образом, средний эффект воздействия (average treatment effect), который интересует авторов работы, — это среднее изменение занятости в ресторане быстрого питания в Нью-Джерси в результате принятия нового закона о минимальной заработной плате.

Как можно было бы посчитать это изменение?

Первый, довольно наивный, подход — взять данные по Нью-Джерси до и после изменения минимальной заработной платы и сравнить их между собой. Этот подход плох, потому что занятость с течением времени может изменяться не только в результате изменения заработной платы, но и по каким-то другим причинам. Есть множество глобальных факторов, которые влияют на всю Америку в целом и могли сказаться на занятости. И, сравнивая такие средние значения до и после, мы не можем понять, влияет ли на происходящее минимальная зарплата или дело в каких-то других факторах. Например, в США могла начаться рецессия, которая способствовала бы снижению занятости во всех штатах, независимо от экономической политики на рынке труда. Или в ресторанном бизнесе могла быть внедрена новая продвинутая технология, которая позволила снизить спрос на труд низкоквалифицированных работников в этой отрасли.

Второй подход состоит в том, чтобы сравнить занятость в среднем ресторане в Нью-Джерси (то есть в испытываемой группе) со средней занятостью в каком-нибудь другом штате, где минимальная зарплата не изменилась (то есть в контрольной группе). Например, в Пенсильвании, в которой в тот же самый период времени минимальная заработная плата осталась на прежнем уровне. Понятно, что такой подход тоже не совершенен. В отличие от совсем «честного» эксперимента, когда мы случайным образом делим рестораны на две группы, здесь эксперимент не совсем чистый, потому что все рестораны первой группы находятся в одном штате, а рестораны второй группы — в другом. И эти штаты, вполне возможно, хотя они и похожи, отличаются не только минимальной зарплатой, но какими-то ещё характеристиками. И поэтому снова невозможно выяснить, объясняются ли различия в занятости в этих двух штатах именно различием в минимальной заработной плате или чем-то ещё.

Вместо каждого из двух указанных подходов можно применить альтернативный метод, который объединяет их преимущества и помогает избежать недостатков. Чтобы понять, как он работает, перечислим основные факторы, которые могут влиять на занятость в типичном ресторане:

- специфические особенности штата, в котором расположен ресторан (эффект штата);
- особенности различных периодов времени, скажем, изменение экономической конъюнктуры (временной эффект);
- эффект изменения минимальной заработной платы (тот самый эффект, который мы пытаемся оценить).

Формально мы можем записать это так:

$$Y_{ist} = \alpha_s + \mu_t + \delta * D_{ist} + \varepsilon_{ist}$$

Здесь индекс i обозначает номер ресторана, индекс s обозначает штат (Нью-Джерси или Пенсильвания), индекс t — момент времени (период **до** изменения заработной платы в Нью-Джерси или период **после** него).

Y_{ist} — число работников, занятых в данном ресторане.

Переменная D равна единице в тех ресторанах, которые находятся в Нью-Джерси в тот период, когда там поменялась заработная плата, и равна нулю во всех остальных случаях.

α_s — эффект штата. Он имеет два значения: $\alpha_{control}$, если наблюдение относится к контрольной группе, то есть к Пенсильвании; $\alpha_{treatment}$ если наблюдение относится к испытываемой группе, то есть к Нью-Джерси.

μ_t — временной эффект. Он равен μ_{before} до изменения заработной платы и μ_{after} после изменения.

δ — эффект воздействия увеличения заработной платы на занятость. Это тот самый эффект, который требуется оценить.

ε_{ist} — случайные ошибки модели

Определим ожидаемое количество занятых в ресторане Нью-Джерси **до** изменения заработной платы:

$$E(Y_{ist} | s = treatment, t = before) = \mu_{before} + \alpha_{treatment}$$

Определим ожидаемое количество занятых в ресторане Нью-Джерси **после** изменения:

$$E(Y_{ist} | s = treatment, t = after) = \mu_{after} + \alpha_{treatment} + \delta$$

Вычитая из второго математического ожидания первое, получим ожидаемое изменение занятости в Нью-Джерси:

$$\Delta_{treatment} = \mu_{after} - \mu_{before} + \delta$$

Теперь определим ожидаемое количество занятых в ресторане Пенсильвании до изменения заработной платы:

$$E(Y_{ist} | s = control, t = before) = \mu_{before} + \alpha_{control}$$

Определим ожидаемое количество занятых в ресторане Пенсильвании после изменения:

$$E(Y_{ist} | s = \text{control}, t = \text{after}) = \mu_{\text{after}} + \alpha_{\text{control}}$$

Аналогично прошлому случаю, вычитая из второго математического ожидания первое, получим ожидаемое изменение занятости в Пенсильвании:

$$\Delta_{\text{control}} = \mu_{\text{after}} - \mu_{\text{before}}$$

Осталось вычесть из первой разности (из $\Delta_{\text{treatment}}$) вторую разность (Δ_{control}), то есть найти ту самую разность разностей, которая дала название анализируемому методу:

$$\Delta_{\text{treatment}} - \Delta_{\text{control}} = \delta$$

Таким образом, мы показали, что интересующий нас эффект воздействия может быть представлен как разность разностей условных математических ожиданий:

$$\begin{aligned} \delta = \Delta_{\text{treatment}} - \Delta_{\text{control}} = \\ [E(Y_{ist} | s = \text{treatment}, t = \text{after}) - E(Y_{ist} | s = \text{treatment}, t = \text{before})] - \\ - [E(Y_{ist} | s = \text{control}, t = \text{after}) - E(Y_{ist} | s = \text{control}, t = \text{before})] \end{aligned}$$

В силу закона больших чисел состоятельной оценкой каждого из этих математических ожиданий является соответствующее среднее значение. Следовательно, состоятельная оценка эффекта воздействия в данном случае может быть рассчитана так:

$$\hat{\delta} = [\bar{Y}_{\text{treatment, after}} - \bar{Y}_{\text{treatment, before}}] - [\bar{Y}_{\text{control, after}} - \bar{Y}_{\text{control, before}}] \quad (11.2)$$

Здесь $\bar{Y}_{\text{control, before}}$ — средний уровень зависимой переменной в контрольной группе до осуществления воздействия (в нашем случае — это средний уровень занятости в ресторанах Пенсильвании до изменения минимальной заработной платы в Нью-Джерси);

$\bar{Y}_{\text{control, after}}$ — средний уровень зависимой переменной в контрольной группе после осуществления воздействия (в нашем случае — это средний уровень занятости в ресторанах Пенсильвании после изменения минимальной заработной платы в Нью-Джерси);

$\bar{Y}_{\text{treatment, before}}$ — средний уровень зависимой переменной в испытуемой группе до осуществления воздействия (в нашем случае — это средний уровень занятости в ресторанах Нью-Джерси до изменения минимальной заработной платы);

$\bar{Y}_{\text{treatment, after}}$ — средний уровень зависимой переменной в испытуемой группе после осуществления воздействия (в нашем случае — это средний уровень занятости в ресторанах Нью-Джерси после изменения минимальной заработной платы);

Таблица 11.2. Оценка воздействия увеличения минимальной заработной платы на занятость методом разности разностей

Переменная		Пенсильвания	Нью-Джерси	
		(1)	(2)	(3)=(2)-(1)
1.	Среднее число занятых в одном ресторане до изменения минимальной зарплат	23,33 (1,35)	20,44 (0,51)	-2,89 (1,44)
2.	Среднее число занятых в одном ресторане после изменения минимальной зарплат	21,17 (0,94)	21,03 (0,52)	-0,14 (1,07)
3.	Изменение среднего числа занятых	-2,16 (1,25)	0,59 (0,54)	2,76 (1,36)

Примечания: в скобках под средними значениями указаны соответствующие стандартные ошибки. Источник: (Card, Krueger, 1994)

Результаты расчетов эффекта воздействия на реальных данных приведены в таблице 11.2. В соответствии с ней эффект воздействия равен:

$$\begin{aligned} \hat{\delta} = [\bar{Y}_{\text{treatment, after}} - \bar{Y}_{\text{treatment, before}}] - [\bar{Y}_{\text{control, after}} - \bar{Y}_{\text{control, before}}] = \\ [21, 03 - 20, 44] - [21, 17 - 23, 33] = 2, 76. \end{aligned}$$

На первый взгляд может показаться, что правильный ответ равен 2,75, а в таблице из статьи и в наших выкладках выписано число 2,76. Это расхождение возникает из-за округлений при промежуточных вычислениях (см. статью).

Можно заключить, что повышение минимальной заработной платы привело к увеличению равновесного уровня занятости в ресторанах быстрого питания Нью-Джерси в среднем на 2,76 человека. Этот результат противоречит выводам стандартных теоретических моделей из микроэкономики (или экономики труда), поэтому вызвал большое обсуждение в литературе. Вы

можете посмотреть для деталей статью Карда и Крюгера, а также цитирующие её более поздние работы, в которых предпринимаются попытки уточнить и объяснить полученные в ней оценки.

Нас же этот пример интересует в качестве иллюстрации применения метода разности разностей. Кроме того, описанная здесь история является примером квазиэксперимента, то есть ситуации, когда распределение объектов на контрольную и испытываемую группу осуществляется не в ходе контролируемого эксперимента, а в силу некоторых внешних обстоятельств, которые при этом позволяют все-таки получить корректную оценку эффекта воздействия.

Геометрическая интерпретация применения метода разности разностей представлена на рисунке 11.1. Пунктирная линия на нём показывает, как менялась бы занятость в Нью-Джерси, если бы этот штат не подвергся воздействию. Подход разности разностей предполагает, что в этом случае динамика занятости была бы аналогична контрольной группе (что в нашем примере означает снижение числа работников).

уровень
занятости

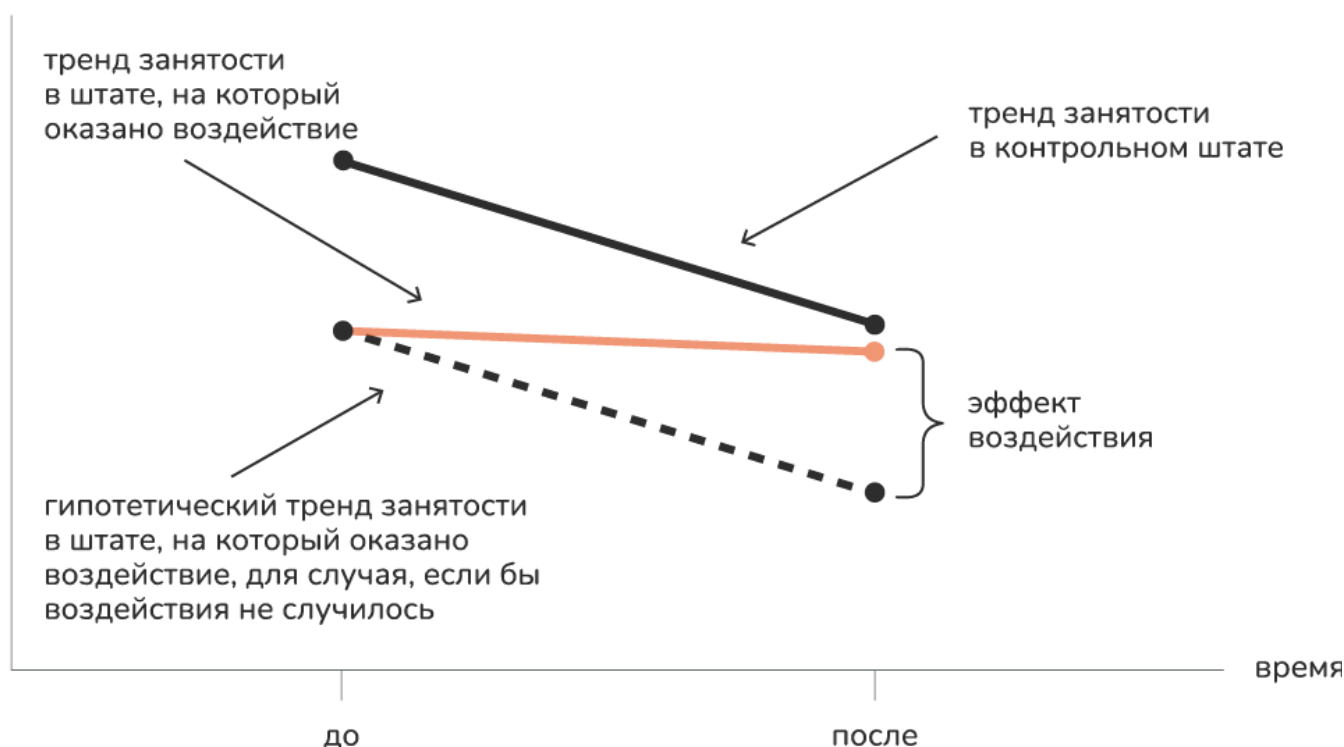


Рисунок 11.1. Геометрическая иллюстрация метода разности разностей

Метод разности разностей непосредственно связан с оценкой моделей при помощи регрессий. В частности, если вы располагаете панельными данными об объектах из испытываемой и контрольной групп за два периода (до и после проведения политики), оценка метода разности разностей может быть получена в результате применения следующей модели:

$$Y_{it} = \beta_0 + \beta_1 * x_i + \beta_2 * z_t + \delta * x_i * z_t + \varepsilon_{it} \quad (11.3)$$

x_i — бинарная переменная, которая равна единице, если i -ый ресторан расположен в Нью-Джерси (то есть относится к испытываемой группе).

z_t — бинарная переменная, которая равна единице для всех наблюдений, относящихся ко второму периоду (периоду после повышения минимальной зарплаты).

В этом случае, применив МНК, получим следующее уравнение:

$$\hat{Y}_{it} = \hat{\beta}_0 + \hat{\beta}_1 * x_i + \hat{\beta}_2 * z_t + \hat{\delta} * x_i * z_t$$

Можно доказать (см. задание в конце главы), что:

$$\begin{aligned} \hat{\beta}_0 &= \bar{Y}_{control, before}, \\ \hat{\beta}_1 &= \bar{Y}_{treatment, before} - \bar{Y}_{control, before}, \\ \hat{\beta}_2 &= \bar{Y}_{control, after} - \bar{Y}_{control, before}, \\ \hat{\delta} &= [\bar{Y}_{treatment, after} - \bar{Y}_{treatment, before}] - [\bar{Y}_{control, after} - \bar{Y}_{control, before}]. \end{aligned}$$

Таким образом, коэффициент при произведении $x_i * z_t$ равен той самой оценке эффекта воздействия, которую мы вывели выше.

Эквивалентный способ оценивания состоит в применении МНК к следующей парной регрессии:

$$\Delta Y_i = \beta_2 + \delta x_i + u_i. (11.4)$$

Здесь x_i — по-прежнему бинарная переменная, которая равна единице, если i -ый объект относится к испытываемой группе, а $\Delta Y_i = Y_{i1} - Y_{i0}$. Если применить к этой модели МНК, то оценка коэффициента при регрессоре снова будет задаваться формулой (11.2).

Для состоятельности этой оценки требуется, чтобы в уравнении (11.3) выполнялись все предпосылки модели со стохастическим регрессором (см. главу 6). В частности, требуется экзогенность объясняющей переменной. Если регрессор оказывается эндогенным из-за пропуска других существенных факторов, влияющих на ΔY , то эта проблема может быть, как обычно, решена включением в уравнение (11.3) контрольных переменных.

Иными словами, метод разности разностей может быть дополнен при помощи учета контрольных переменных, что делает его взаимосвязь с прочими моделями на панельных данных ещё более тесной.

В то же время у метода разности разностей есть важное преимущество по сравнению с ними. Дело в том, что иногда вместо панельных данных вам доступны лишь так называемые повторяющиеся пространственные данные. Это означает, что для разных периодов имеются данные по различным объектам. Скажем, в нашем примере про рестораны быстрого питания это означало бы, что до изменения заработной платы исследователи опросили бы одни рестораны, а после изменения — другие. Тогда непосредственно оценить уравнение (11.3) было бы невозможно, так как нельзя было бы рассчитать разности $\Delta Y_i = Y_{i1} - Y_{i0}$ (потому что по каждому отдельному объекту доступны данные только для одного из двух периодов). Однако интересующий нас эффект воздействия всё ещё мог бы быть рассчитан при помощи метода разности разностей по формуле (11.2).

Метод разности разностей широко используется в современных прикладных исследованиях для анализа последствий применения тех или иных мер экономической политики.

11.3. Локальный средний эффект воздействия (LATE)

В некоторых случаях разделение на группу, подвергшуюся воздействию, и контрольную группу может быть эндогенным. Например, если индивиды сами решают, подвергаться ли им воздействию, то при попытке оценить величину среднего эффекта воздействия мы столкнемся со смещением из-за самоотбора.

Такая ситуация особенно вероятна, если эффект воздействия является гетерогенным, то есть существенно отличается для разных индивидов. В этих условиях корректно оценить ATE затруднительно, зато возможно состоятельно оценить так называемый локальный средний эффект воздействия (local average treatment effect, LATE)

Чтобы пояснить идею работы этого метода, мы будем использовать статью (Angrist, 1990), которая посвящена попытке оценить воздействие службы в армии на будущие доходы.

Простое сравнение средних доходов людей, которые служили и не служили в армии, показывает, что доходы ветеранов устойчиво ниже. Однако эти оценки не вызывают доверия, так как они могут быть смещены из-за эндогенности решения о прохождении службы (возможно, дело в том, что в армию идут менее способные к гражданской работе люди). Подход, который использует Ангрис, позволяет преодолеть эту проблему.

Ангрис обращается к данным о ветеранах войны во Вьетнаме. В те времена в США использовался призыв на военную службу. Приоритетность призыва зависела от так называемого случайного порядкового номера (Random Sequence Number, RSN). Этот номер присваивался каждому мужчине в результате розыгрыша лотереи. Номер изменялся от 1 до 365, так как был привязан к дате рождения. Министерство обороны каждый год определяло некоторый потолок (пороговое значение) случайного порядкового номера. После этого на службу призывались все мужчины с RSN ниже этого порогового значения.

Говоря коротко, на службу призывались мужчины, которым выпало пойти в армию в результате некоторой лотереи (в статье этих победителей называют «draft-eligible»).

Важно отметить, что статус победителя лотереи вовсе не равен статусу ветерана войны (это две разные переменные):

- С одной стороны, вовсе не все ветераны войны были победителями лотереи, ведь кто-то записывался на службу добровольно.
- С другой стороны, не все победители лотереи стали ветеранами, так как кто-то из победителей избежал службы в силу, например, медицинских ограничений.

Такая ситуация называется двусторонним несоблюдением (two-sided noncompliance).

Тем не менее, статус победителя лотереи и статус ветерана положительно коррелированы друг с другом: победители лотереи в среднем оказывались в армии с большей вероятностью, чем остальные мужчины. В то же время статус победителя лотереи не коррелирован с прочими характеристиками индивида, которые могут влиять на его доход (так как присваивался случайным образом). Все это указывает на то, что статус победителя лотереи можно было бы использовать в качестве инструмента для переменной, характеризующей ветеранский статус. Мы, однако, не будем сразу пользоваться уже знакомой методологией двухшагового МНК, а начнем с альтернативного взгляда на описанную ситуацию.

Будем использовать следующие обозначения:

Y_i — значение зависимой переменной (potential outcome). В нашем примере это доход i -го индивида.

D_i — это снова переменная воздействия, которая в отличие от первого параграфа является **эндогенной**. В нашем примере это бинарная переменная, равная единице, если i -ый индивид служил в армии.

Переменная воздействия зависит от некоторого бинарного инструмента Z_i — так называемого внешнего предписания (treatment assignment). В нашем примере это бинарная переменная, равная единице, если i -ый индивид выиграл в лотерею:

$$D_i = D_i(Z_i) = \begin{cases} D_i(1), & \text{если } Z_i = 1 \\ D_i(0), & \text{если } Z_i = 0 \end{cases}$$

В общем случае переменная Y_i может зависеть и от переменной воздействия, и от предписания:

$$Y_i = Y_i(D_i, Z_i) = \begin{cases} Y_i(0, 0), & \text{если } D_i = 0, Z_i = 0 \\ Y_i(0, 1), & \text{если } D_i = 0, Z_i = 1 \\ Y_i(1, 0), & \text{если } D_i = 1, Z_i = 0 \\ Y_i(1, 1), & \text{если } D_i = 1, Z_i = 1 \end{cases}$$

Как обычно, для каждого отдельного индивида мы наблюдаем только одно из четырёх возможных значений (так как один и тот же человек не мог одновременно служить и не служить или одновременно выиграть в лотерею и не выиграть).

Сформулируем предположения, которые потребуются нам для оценки локального среднего эффекта воздействия (LATE).

Предположение 11.1 о независимости (Independence).

Z_i не зависит от $(Y_i(D_i(0), 0), Y_i(D_i(1), 1), D_i(0), D_i(1))$.

В нашем примере это означает, что результат лотереи не зависит от характеристик индивида.

Предположение 11.2 об исключительном ограничении (Exclusion restriction).

Для любого i верно, что $Y_i(1, 0) = Y_i(1, 1)$ и $Y_i(0, 0) = Y_i(0, 1)$.

Эта предположение означает, что инструмент Z_i не оказывает непосредственного влияния на зависимую переменную. Он может влиять на неё только опосредованно: через изменение переменной D_i .

В нашем примере это означает, что выигрыш в лотерее сам по себе не влияет на доход¹. Он влияет на вероятность попадания в армию. А уже служба в армии, в свою очередь, может влиять на доход.

С учетом этой предпосылки зависимую переменную можно записать как функцию от единственного фактора $Y_i(D_i)$:

$$Y_i(1) \equiv Y_i(1, 0) = Y_i(1, 1),$$

$$Y_i(0) \equiv Y_i(0, 0) = Y_i(0, 1).$$

Предпосылка 11.3 о первом шаге (First Stage).

Для любого i верно, что $E(D_i(1) - D_i(0)) \neq 0$.

Это аналог предпосылки о релевантности инструмента. То есть о том, что инструмент влияет на переменную воздействия.

В нашем примере это означает, что выигрыш в лотерею увеличивает для индивида вероятность оказаться в армии.

Предпосылка 11.4 о монотонности (Monotonicity).

Для любого i верно, что $D_i(1) - D_i(0) \geq 0$.

Заметим, что вместо $D_i(1) - D_i(0) > 0$ можно с таким же успехом писать $D_i(1) - D_i(0) = 1$, так как здесь это эквивалентно. Сформулированная предпосылка означает, что разность $D_i(1) - D_i(0)$ может принимать только два значения: 0 или 1.

Чтобы пояснить эту предпосылку, отметим, что, с точки зрения реакции на предписание, каждый индивид может быть отнесен к одному из четырех типов:

- **Complier**: если получил предписание, то идет в армию, а если не получил предписание, то не идет в армию. Для такого индивида верно, что $D_i(1) - D_i(0) = 1 - 0 = 1$
- **Always-taker**: независимо от предписания идет в армию: $D_i(1) - D_i(0) = 1 - 1 = 0$
- **Never-taker**: независимо от предписания не идет в армию: $D_i(1) - D_i(0) = 0 - 0 = 0$
- **Defier**: если получил предписание, то не идет в армию, а если не получил предписание, то идет в армию. Для такого странного индивида верно, что $D_i(1) - D_i(0) = 0 - 1 = -1$

Ясно, что поведение последнего типа индивидов выглядит нереалистично. Предпосылка о монотонности как раз и предполагает, что таких индивидов не существует.

Последняя предпосылка важна для того, чтобы понять, что такое локальный средний эффект воздействия (LATE) — это эффект воздействия для индивидов типа **Complier** (по-русски их можно назвать послушными индивидами, то есть индивидами, которые строго следуют предписанию):

$$LATE = E(Y_i(1) - Y_i(0) | D_i(1) - D_i(0) > 0)$$

Для других групп индивидов оценить АТЕ невозможно:

- Для индивидов типа **Always-taker** у нас нет ни одного наблюдения, показывающего, что с ними будет, если они не подвергнутся воздействию.
- Для индивидов типа **Never-taker** у нас нет ни одного наблюдения, показывающего, что с ними будет, если они подвергнутся воздействию.

Если эксперимент устроен так, что в группу, подвергшуюся воздействию, могут попасть только те, кто получил предписание (а те, кто его не получил, могут оказаться **исключительно** в контрольной группе), то такая ситуация называется односторонним несоблюдением (one-sided noncompliance).

В нашем примере такая ситуация наблюдалась бы, если бы в США законодательно запретили поступать на военную службу тем, кто не выиграл в лотерею (даже в качестве добровольцев).

В случае одностороннего несоблюдения четвертая предпосылка не потребуется нам для оценки эффекта воздействия. А в случае двустороннего несоблюдения, как мы убедимся далее, она необходима.

Теорема о LATE

Если выполнены предпосылки A1-A4, то

$$\frac{E(Y_i | Z_i = 1) - E(Y_i | Z_i = 0)}{E(D_i | Z_i = 1) - E(D_i | Z_i = 0)} = LATE.$$

Для доказательства теоремы рассмотрим сначала два слагаемых в числителе данной дроби:

$$\begin{aligned} E(Y_i | Z_i = 1) &= \{\text{в силу предпосылки A2}\} = \\ E(Y_i(0) + (Y_i(1) - Y_i(0)) * D_i | Z_i = 1) &= \{\text{в силу предпосылки A1}\} = \\ E(Y_i(0) + (Y_i(1) - Y_i(0)) * D_i(1)) \end{aligned}$$

Аналогично:

$$\begin{aligned} E(Y_i | Z_i = 0) &= \{\text{в силу предпосылки A2}\} = \\ E(Y_i(0) + (Y_i(1) - Y_i(0)) * D_i | Z_i = 0) &= \{\text{в силу предпосылки A1}\} = \\ E(Y_i(0) + (Y_i(1) - Y_i(0)) * D_i(0)) \end{aligned}$$

Следовательно, числитель дроби в формуле для LATE можно записать так:

$$\begin{aligned} E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0) = \\ E(Y_i(0) + (Y_i(1) - Y_i(0)) * D_i(1)) - \\ - E(Y_i(0) + (Y_i(1) - Y_i(0)) * D_i(0)) = \\ E((Y_i(1) - Y_i(0)) * (D_i(1) - D_i(0))) \end{aligned}$$

Преобразуем далее полученное выражение. Для этого воспользуемся свойством математического ожидания:

$$\begin{aligned} E((Y_i(1) - Y_i(0)) * (D_i(1) - D_i(0))) = \\ E((Y_i(1) - Y_i(0)) * 1 | D_i(1) - D_i(0) = 1) * P(D_i(1) - D_i(0) = 1) + \\ + E((Y_i(1) - Y_i(0)) * 0 | D_i(1) - D_i(0) = 0) * P(D_i(1) - D_i(0) = 0) = \\ E((Y_i(1) - Y_i(0)) | D_i(1) - D_i(0) = 1) * P(D_i(1) - D_i(0) = 1) + 0 \end{aligned}$$

Здесь мы опираемся на предпосылку 11.4, поэтому не включаем случай $D_i(1) - D_i(0) = -1$, ведь в силу этой предпосылки такой случай невозможен.

Следовательно, дробь, записанная в формулировке теоремы, может быть преобразована так:

$$\begin{aligned} \frac{E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0)}{E(D_i|Z_i = 1) - E(D_i|Z_i = 0)} = \\ \frac{E((Y_i(1) - Y_i(0)) | D_i(1) - D_i(0) = 1) * P(D_i(1) - D_i(0) = 1)}{P(D_i(1) - D_i(0) = 1)} = \\ E((Y_i(1) - Y_i(0)) | D_i(1) - D_i(0) = 1) = LATE, \end{aligned}$$

что и требовалось доказать.

Предпосылка 11.3 необходима, чтобы знаменатель дроби не равнялся нулю.

Если заменить математические ожидания их выборочными аналогами, то получим оценку LATE:

$$\widehat{LATE} = \frac{\bar{Y}_1 - \bar{Y}_0}{\bar{D}_1 - \bar{D}_0}$$

$\bar{Y}_1$ — среднее значение зависимой переменной для индивидов, которые получили предписание. В нашем примере это средний доход тех, кто выиграл в лотерею.

$\bar{Y}_0$ — среднее значение зависимой переменной для индивидов, которые не получили предписание. В нашем примере это средний доход тех, кто проиграл в лотерею.

$\bar{D}_1$ — доля тех, кто подвергся воздействию, среди тех, кто получил предписание. В нашем примере это доля победителей лотереи, которые пошли служить.

$\bar{D}_0$ — доля тех, кто подвергся воздействию, среди тех, кто не получил предписание. В нашем примере это доля проигравших в лотерею, которые при этом всё равно пошли служить.

Мы показали, что

$$E(D_i|Z_i = 1) - E(D_i|Z_i = 0) = P(D_i(1) - D_i(0) = 1)$$

Указанная вероятность — это доля послушных индивидов в генеральной совокупности. В свою очередь, величина $\bar{D}_1 - \bar{D}_0$ — это оценка разности матожиданий

$$E(D_i|Z_i = 1) - E(D_i|Z_i = 0)$$

Следовательно, величина $\bar{D}_1 - \bar{D}_0$ является оценкой доли послушных индивидов в нашей выборке.

Если игнорировать различие между переменной предписания и переменной воздействия, то можно просто оценить влияние предписания на изменение зависимой переменной вот таким образом:

$$ITT = E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0)$$

Эта величина называется эффектом предписания (intention-to-treat effect, ITT effect).

Состоятельная оценка этой величины снова может быть вычислена путем замены матожиданий соответствующими средними:

$$\widehat{ITT} = \bar{Y}_1 - \bar{Y}_0.$$

В силу доказанной нами теоремы

$$LATE = \frac{E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0)}{E(D_i|Z_i = 1) - E(D_i|Z_i = 0)} = \frac{ITT}{P(D_i(1) - D_i(0) = 1)}$$

Так как знаменатель последней дроби лежит между нулем и единицей, то $|LATE| \geq |ITT|$. Причем равенство достигается, только если выборка на 100% состоит из послушных индивидов.

Мы не зря упоминали двухшаговый МНК в начале этого параграфа. Дело в том, что в случае, когда инструментальная переменная и эндогенный регрессор являются бинарными, 2МНК-оценка совпадает с оценкой LATE (доказательству этого факта посвящено одно из упражнений в конце главы).

Таким образом, идеология LATE полезна в том числе тем, что в условиях гетерогенного эффекта воздействия позволяет четко определить, что именно мы оцениваем, используя 2МНК: эффект воздействия политики для индивидов, которые следуют предписанию.

1. Напомним, что здесь речь идет не о лотерее с денежным выигрышем, а о лотерее, победители которой получали повестку от министерства обороны. [↗](#)

11.4. Разрывный регрессионный дизайн

В некоторых квазиэкспериментах вероятность того, что объект подвергнется воздействию, является разрывной функцией от наблюдаемой переменной (или группы переменных). В этом случае для оценки эффекта воздействия может быть применен **разрывный регрессионный дизайн** (regression discontinuity design, RDD).

Он бывает двух видов: четкий (sharp) и нечеткий (fuzzy). Рассмотрим их последовательно.

Четкий разрывный дизайн

Представим, например, что вы хотите оценить влияние обучения индивида в магистратуре университета N на его будущий доход. Скажем, у вас есть результаты вступительных экзаменов в эту магистратуру в 2010 году для всех абитуриентов, а также данные об их доходах в 2020 году. Доход является зависимой переменной. Для поступления в эту магистратуру нужно сдать вступительный экзамен. Пусть проходной балл в 2010 году составил 150 баллов из 200 возможных. В этом случае, если абитуриент набрал 150 и более баллов, то он поступает в магистратуру (то есть оказывается в испытываемой группе). Если же он набрал 149 баллов, то не поступает (то есть оказывается в контрольной группе).

Такая организация данных возникает и в других случаях, когда назначение воздействия определяется какими-то формальными правилами. Например, в двухпартийной системе кандидат от партии побеждает на выборах, если набирает больше половины голосов. Или финансовая помощь может назначаться индивиду, если его доход оказывается ниже некоторого порогового уровня.

Для определенности мы продолжим концентрироваться на примере с поступлением в университет N. В этом случае испытываемая группа — это выпускники университета N, которые в свое время успешно в него поступили, набрав проходной балл. Контрольная группа — это те абитуриенты, которые в свое время не смогли поступить в университет N.

Понятно, что простое сравнение средних уровней доходов в испытываемой и контрольной группе не даст нам состоятельной оценки эффекта от обучения в магистратуре. Ведь, скорее всего, результат вступительного испытания коррелирован со способностями индивида. Тем самым в университет поступили индивиды с более высоким уровнем способностей. А значит, их доход может оказаться выше не из-за того, что они учились в университете N, а из-за того, что они в целом более способные, чем индивиды из контрольной группы.

Мы могли бы устранить это смещение, сравнивая средние уровни доходов для индивидов, которые набрали ровно 150 баллов (ровно проходной балл) и для индивидов, которые набрали 149 баллов (то есть оказались непосредственно под чертой). Судя по результатам вступительного экзамена, уровень способностей для таких индивидов очень близок. Можно считать, что разница в 1 балл из 200 определяется исключительно случайными факторами. А значит, сравнение средних уровней доходов в этом случае даст состоятельную оценку эффекта воздействия обучения в магистратуре университета N на будущий доход.

Проблема последнего подхода состоит в том, что в вашей выборке может оказаться очень мало индивидов, имеющих ровно 149 баллов и ровно 150 баллов. Из-за этого ваша оценка будет иметь низкую точность (высокую дисперсию).

Разрывный дизайн позволяет преодолеть эту проблему. Чтобы показать, как он работает, воспользуемся следующими обозначениями:

$D_i = 1$, если i -ый объект вошел в группу, подвергшуюся воздействию (treatment group). В нашем примере, если i -ый индивид учился в магистратуре университета N. Доход i -го индивида в этом случае будем обозначать $Y_i(1)$.

$D_i = 0$, если i -ый объект вошел в контрольную группу (control group). В нашем примере $D_i = 0$, если i -ый индивид не учился в магистратуре университета N. Доход i -го индивида в этом случае будем обозначать $Y_i(0)$.

Интересующий нас эффект воздействия — это изменение дохода индивида в результате обучения в магистратуре университета N:

$$\rho = Y_i(1) - Y_i(0)$$

Естественно предположить, что ожидаемый доход индивида связан с его баллом за экзамен. Эта связь возникает за счет того, что балл за экзамен является прокси-переменной для способностей индивида. В этом случае доход индивида можно следующим образом записать как функцию от переменных x_i и D_i :

$$Y_i = \alpha + \beta x_i + \rho D_i + \varepsilon_i,$$

$$D_i = \begin{cases} 0, & \text{при } x_i < x^*, \\ 1, & \text{при } x_i \geq x^*. \end{cases}$$

Здесь — пороговый уровень, преодоление которого определяет попадание в испытываемую группу. В нашем примере — это проходной балл, равный 150. Переменную x в разрывном дизайне называют переменной отбора (selection variable¹).

Отличие от стандартного случая использования контрольной переменной, который мы обсуждали в предыдущих главах, здесь состоит в том, что переменная интереса D не просто коррелирована с другим регрессором x , а является неслучайной (детерминированной) функцией от него. Поэтому для состоятельности МНК-оценок достаточно потребовать экзогенности переменной отбора. Тогда переменная D автоматически тоже будет экзогенной.

В условиях этой предпосылки для индивида из контрольной группы условное математическое ожидание зависимой переменной равно:

$$E(Y_i(0)|x_i) = \alpha + \beta x_i.$$

А для индивида из испытываемой группы оно составляет:

$$E(Y_i(1)|x_i) = \alpha + \beta x_i + \rho.$$

Геометрически описанная ситуация проиллюстрирована на рисунке 11.2. Коэффициент равен величине разрыва функции ожидаемого дохода в точке x^* .

Таким образом, чтобы выяснить эффект воздействия, достаточно оценить полученное уравнение регрессии. МНК-оценка и будет состоятельной оценкой эффекта воздействия.

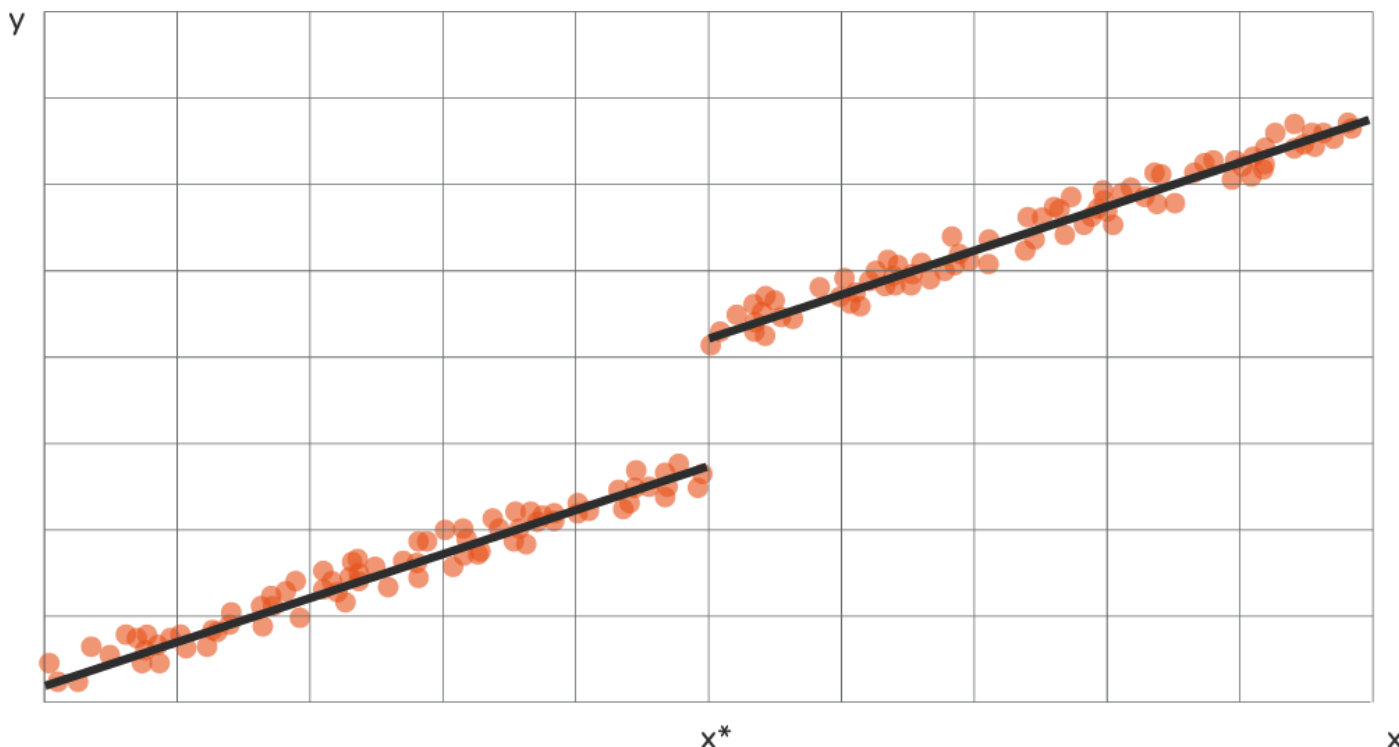


Рисунок 11.2. Пример модели с разрывным дизайном. Линейный случай

Влияние переменной отбора на зависимую переменную может быть нелинейным:

$$Y_i = \alpha + f(x_i) + \rho D_i + \varepsilon_i \quad (11.5),$$

$$D_i = \{0, \text{при } x_i < x^*, 1, \text{при } x_i \geq x^*\}. \quad (11.6).$$

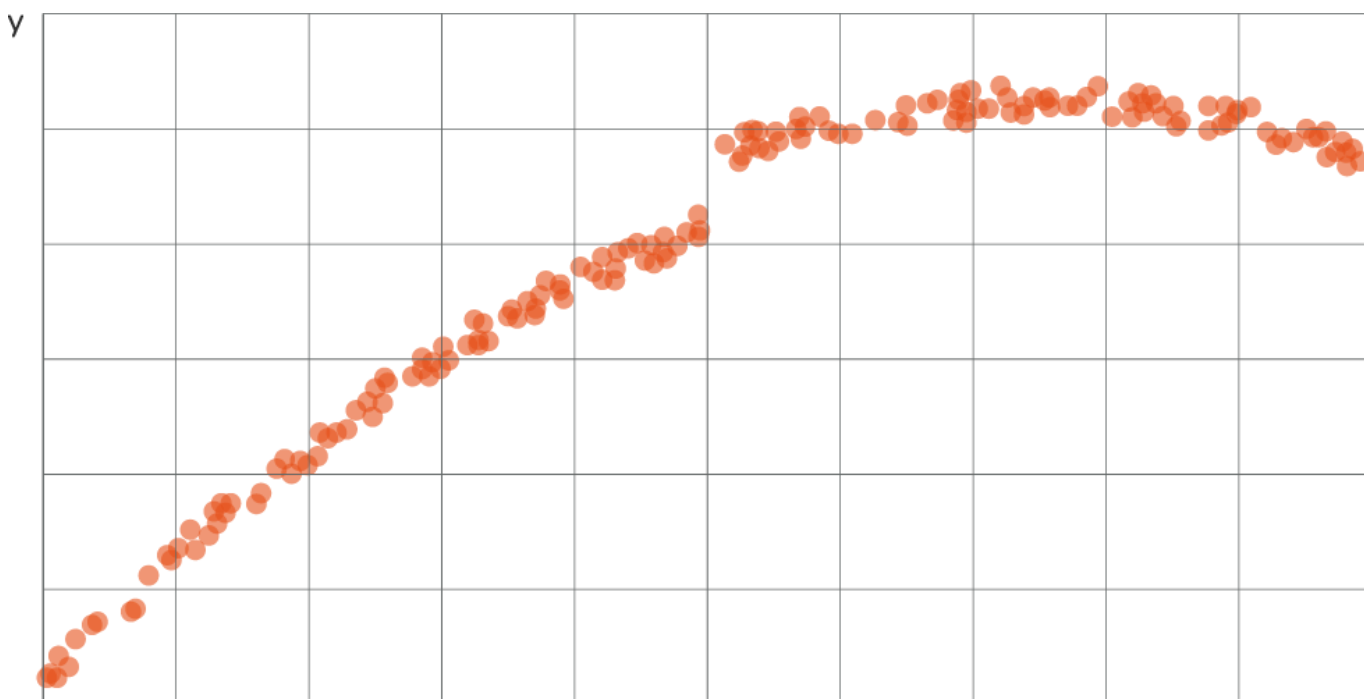
В качестве функции $f(x_i)$ обычно используют полином. Например, для квадратичной функции $f(x_i)$:

$$Y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \rho D_i + \varepsilon_i.$$

Соответствующий пример изображен на рисунке 11.3.

При необходимости спецификация (11.5) может быть дополнена путем включения в уравнение контрольных переменных.

Если эффект воздействия является гетерогенным, то есть различным для разных индивидов, то разрывный дизайн корректно оценивает этот эффект не для любых индивидов, а только для тех, у кого значение переменной отбора близко к пороговому уровню.



X

Рисунок 11.3. Пример модели с разрывным дизайном. Нелинейный случай

Нечеткий разрывный дизайн

При нечетком разрывном дизайне вместо условия (11.6) выполняется следующая предпосылка:

$$P(D_i = 1 | x_i) = \begin{cases} g_0(x_i), & \text{при } x_i < x^*, \\ g_1(x_i), & \text{при } x_i \geq x^*, \end{cases} \text{ где } g_0(x^*) \neq g_1(x^*). \quad (11.7)$$

Таким образом, при нечетком разрывном дизайне преодоление порога влияет на вероятность попадания в испытываемую группу, однако не гарантирует это попадание.

В нашем примере с поступлением такая ситуация могла бы возникнуть, если бы некоторые из абитуриентов, набравшие проходной балл, после этого по каким-то причинам не стали бы учиться в магистратуре университета N. И наоборот, какие-то из абитуриентов, не набравшие проходной балл, все-таки смогли бы пройти обучение (скажем, не на бюджетной основе, а на платной).

В этом случае переменная D в уравнении $Y_i = \alpha + f(x_i) + \rho D_i + \epsilon_i$ больше не является экзогенной, так как зависит не только от преодоления порога, но и от решения индивида. Следовательно, обычный МНК не позволит получить состоятельную оценку коэффициента ρ в уравнении (11.5).

Поэтому в условиях нечеткого разрывного дизайна для оценки эффекта воздействия лучше применить 2МНК, используя в качестве инструмента переменную T :

$$T_i = \begin{cases} 0, & \text{при } x_i < x^*, \\ 1, & \text{при } x_i \geq x^*. \end{cases}$$

В силу того, что вероятность попадания в контрольную группу зависит от того, превышает ли переменная отбора порог или нет, такой инструмент будет релевантным. В то же время он является экзогенным, так как определяется исключительно экзогенной переменной отбора.

Пример 11.2. Партия власти и результаты выборов

В заключение этого параграфа рассмотрим пример применения метода разрывного регрессионного дизайна для модели бинарного выбора (таким образом, этот пример опирается не только на текущую главу, но и на главу 10).

Мы обратимся к статье (Lee, 2008), автор которой задался вопросом, существует ли у партии власти дополнительное преимущество на выборах (Incumbency advantage)?

Точнее говоря, Ли пытается выяснить, имеет ли преимущество кандидат на место в Палате представителей США на текущих выборах, если его партия выиграла предыдущие выборы?

В поиске ответа на этот вопрос есть очевидная трудность: вполне возможно, что лица, занимающие должность, хорошо соответствуют предпочтениям избирателей, имеют большую поддержку и выигрывают благодаря этому, а вовсе не благодаря «административному ресурсу».

Чтобы отделить один эффект от другого, Ли при помощи метода четкого разрывного дизайна анализирует вероятность быть избранным как функцию от соотношения голосов за партии демократов и республиканцев на предыдущих выборах

Он использует тот факт, что победитель на выборах может быть определен следующим образом:

$$D_{i,t} = 1, \text{ если } x_{i,t} \geq 0 \text{ и } D_{i,t} = 0, \text{ если } x_{i,t} < 0,$$

где $x_{i,t}$ — разность между долями голосов, отданных за демократов и за республиканцев (Democratic vote share margin of victory). Если эта разность больше нуля, значит, демократы получили на выборах больше голосов, чем республиканцы, и, следовательно, выиграли их².

$D_{i,t}$ — переменная, равная единице, если на выборах победил кандидат от демократов. Индекс i соответствует избирательному округу, а индекс t — номеру года.

Ли оценивает параметры следующей логит-модели бинарного выбора:

$$P(D_{i,t+1} = 1) = F(\alpha + \beta x_{i,t} + \rho D_{i,t}) \quad (11.8)$$

Здесь $P(D_{i,t+1} = 1)$ — вероятность победы демократической партии на выборах в периоде $t + 1$, $F(\bullet)$ — логистическая функция. (В некоторых спецификациях автор добавляет также контрольные переменные, которые могут влиять на вероятность победы. Например, опыт кандидата.)

Результаты оценивания модели представлены графически на рисунке 11.4. По мере роста доли голосов, которую демократы получили на предыдущих выборах, увеличивается вероятность их победы на следующих выборах. Это увеличение логично и могло бы объясняться увеличением популярности демократов на территории соответствующего избирательного округа, если бы не значительный разрыв в точке $x_{i,t} = 0$ (соответствующей ровно половине голосов, отданных избирателями демократам). Мы видим, что, набрав немного менее 50% голосов на предыдущих выборах, демократы имеют шансы на победу лишь около 15% (см. левую шкалу на графике). Однако если число голосов на прошлых выборах оказывается хоть немного больше 50%, то вероятность победы демократов в текущей избирательной кампании резко растет сразу до 60%.

Величина этого разрыва соответствует оценке коэффициента ρ и характеризует эффект воздействия нахождения партии у власти в момент избирательной кампании на вероятность победить на выборах. Как можно видеть из рисунка, партия власти получает примерно 45-процентный «бонус» к вероятности победы.

Всё это позволяет автору заключить, что партия власти получает существенное преимущество на выборах (независимо от текущих предпочтений избирателей).

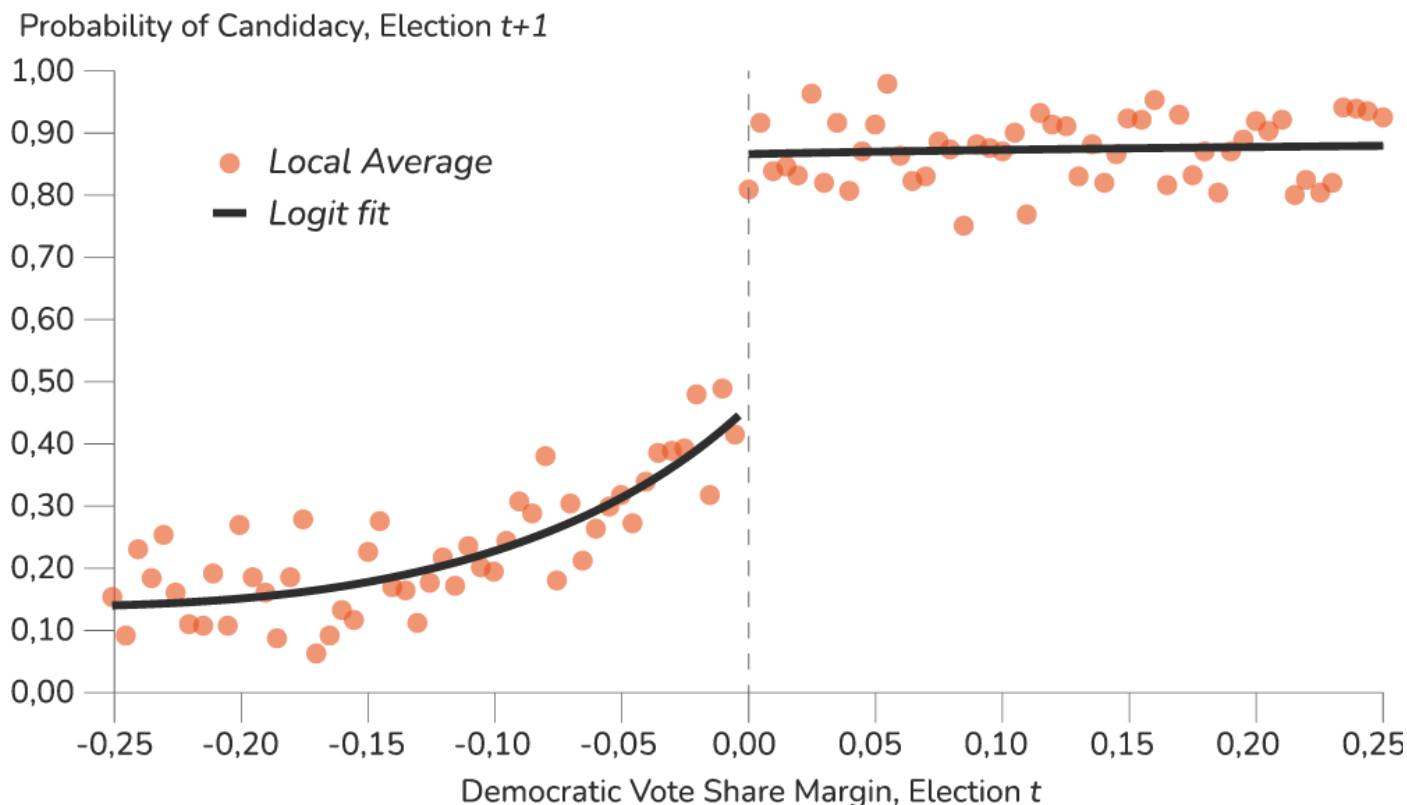


Рисунок 11.4. Зависимость вероятности победы демократов на выборах в периоде $t+1$ в зависимости от доли голосов, набранных в периоде t .

Примечание: Democratic vote share margin of victory, Election t — разность между долями голосов, отданных за демократов и за республиканцев на выборах в периоде t . Следовательно, положительные значения этой переменной соответствуют победе демократов, а отрицательные — победе республиканцев. Источник: (Lee, 2008).

Важным свидетельством корректности выводов автора являются результаты так называемого **теста плацебо** (placebo test). Такое название в литературе об эффектах воздействия носит не какая-то конкретная статистическая процедура, а общий подход к проверке надежности результатов, который устроен так: необходимо проанализировать спецификацию модели, которая похожа на вашу базовую спецификацию, но в которой **точно не должно возникнуть** значимого эффекта воздействия. Если его там действительно не возникает, это говорит в пользу корректности выводов базовой модели.

В случае с работой (Lee, 2008) подобный тест состоит в том, чтобы оценить параметры следующей модели:

$$P(D_{i,t-1} = 1) = F(\beta_0 + \beta_1 D_{i,t} + \beta_2 x_{i,t}) \quad (11.9)$$

В отличие от базовой модели (11.8) в левой части уравнения стоит вероятность победы не на следующих выборах, а на предыдущих ($t - 1$). Ясно, что затруднительно использовать нахождение у власти сегодня для того, чтобы выиграть выборы несколько лет назад. Следовательно, в этом случае коэффициент при переменной $D_{i,t}$ должен быть статистически незначимым, и на соответствующем графике не должно возникать разрыва. Как видно на рисунке 11.5, разрыва в графике вероятности победы на выборах, действительно, нет. Это означает, что модель Ли успешно проходит тест плацебо.

Подчеркнем, что этот тест может применяться не только в рамках разрывного дизайна. Он может быть полезен при проверке корректности выводов модели, оцененной любым из методов.

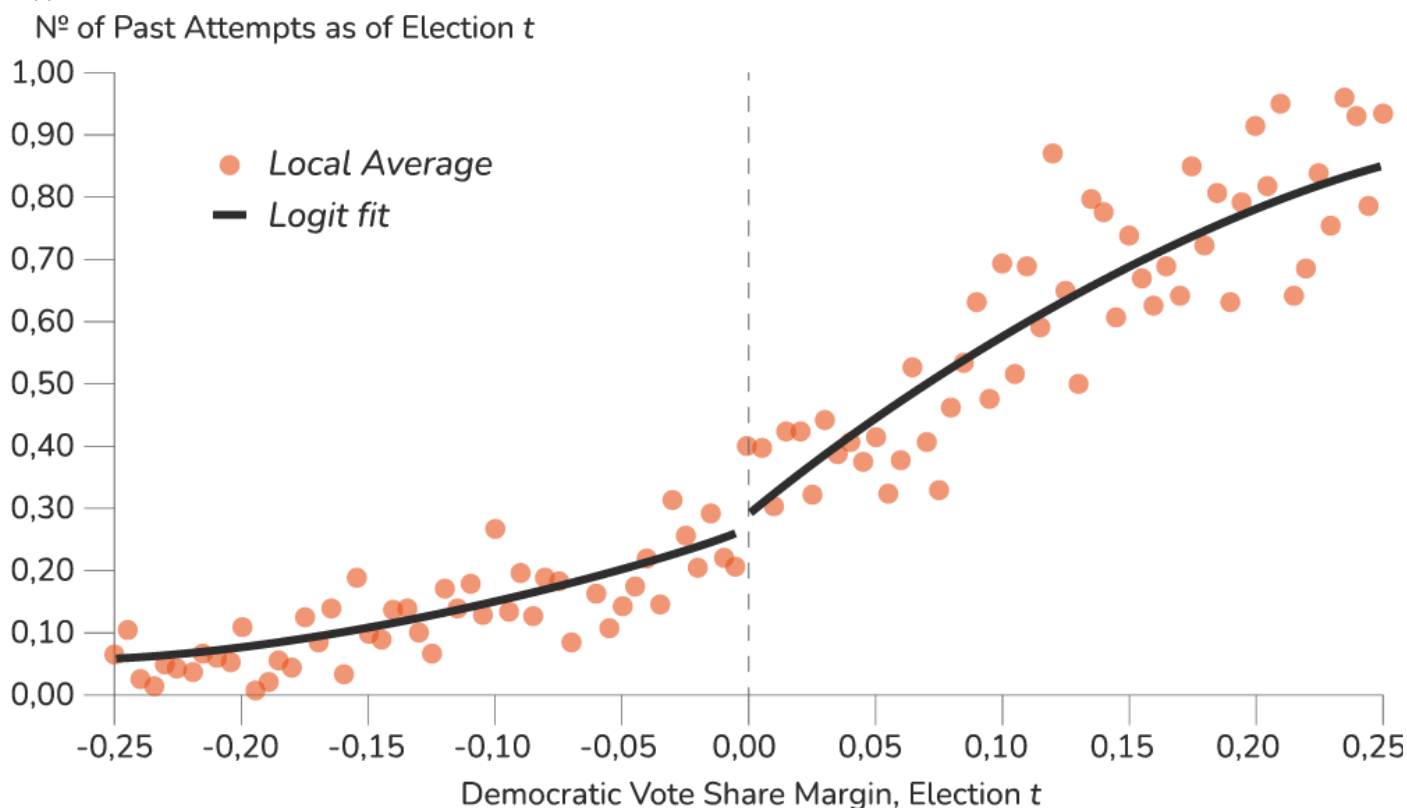


Рисунок 11.5. Результаты теста плацебо. Источник: (Lee, 2008)

1. Иногда в литературе используются также термины «gunning variable» и «forcing variable». [↗](#)
2. Разумеется, такую методологию можно применить только в том случае, когда на победу на выборах претендуют ровно две партии. В случае с США — это партии демократов и республиканцев. [↗](#)

Приложение. Эконометрические кейсы

В этом разделе я выкладываю примеры практических задания (кейсов) на реальных данных, которые могут быть использованы для дополнительной практики в применении эконометрических методов. Часть из кейсов опираются исключительно на темы учебника, а часть требуют знаний по дополнительным разделам эконометрики, которые пока в учебнике не отражены, но будут отражены позже по мере добавления новых глав.

Поскольку тут используются актуальные (и, следовательно, обновляемые) данные, то конкретный числовой ответ на любой из вопросов может меняться со временем. Поэтому для таких кейсов я не публикую единственного решения.

Тем не менее, если вы нуждаетесь в методических указаниях по решению этих заданий (например, чтобы использовать их в качестве методических материалов в своем вузе), то напишите мне на kartaev@gmail.com

[Кейс 1. Население и расходы бюджета](#)

[Кейс 2. Динамика стоимости работ](#)

Кейс 1. Население и расходы бюджета

Кейс подготовлен с использованием данных, предоставленных Департаментом экономической политики и развития города Москвы

В задании используются данные единого хранилища данных ИАС МКР города Москвы (<https://ehd.moscow/>)

Используя сайт единого хранилища выше, загрузите информацию по следующим переменным:

- Численность населения в муниципальных образованиях города Москвы
- Расходы местного бюджета по муниципальным образованиям города Москвы

Для расчетов используйте панельные данные за последние пять лет, (используйте наиболее свежую информацию из доступной в едином хранилище на момент выполнения задания). Обратите внимание, что вам требуются данным по отдельным муниципальным образованиям, а не по округам и не по городу в целом. Удалите лишние строчки из массива.

Если вам нужно освежить в памяти эконометрическую теорию по соответствующей теме, то обратитесь к главе 9 учебника.

В этом задании вам необходимо выяснить, как расходы бюджета зависят от численности населения.

Для этого предпримите следующие шаги:

1. Импортируйте данные в эконометрический пакет.
2. Оцените три модели:
 - Регрессию пула
 - Модель с фиксированными эффектами
 - Модель со случайными эффектами
3. Запишите результаты оценивания в виде единой сводной таблицы. Осуществив необходимые тесты и интерпретировав их результаты, выберите наилучшую модель из трех оцененных вариантов.
4. Для выбранной модели дайте содержательную интерпретацию результатов: поясните, как именно изменение численности населения сказывается на расходах бюджета. Не забудьте проверить, является ли указанное влияние статистически значимым.
5. Оцените модель заново, добавив в нее временные эффекты. Осуществив соответствующий тест, определите, оправдано ли добавление в модель этой группы переменных.
6. Для модели из пункта 5 дайте содержательную интерпретацию результатов: поясните, как именно изменение численности населения сказывается на расходах бюджета (если сказывается). Не забудьте проверить, является ли указанное влияние статистически значимым.
7. Понятно, что расходы бюджета могут зависеть от численности населения муниципального образования нелинейно. Подумайте, как можно учесть этот факт в вашей модели. Оцените соответствующую регрессию и интерпретируйте полученные результаты.

Кейс 2. Динамика стоимости работ

Кейс подготовлен с использованием данных, предоставленных Департаментом экономической политики и развития города Москвы

В задании используются данные единого хранилища данных ИАС МКР города Москвы (<https://ehd.moscow/>)

Используя сайт единого хранилища выше, загрузите информацию по индексам изменения стоимости строительно-монтажных работ. Выберите одну из доступных в базе данных отраслей на ваш вкус.

Ваша цель в этом кейсе — построить прогноз индекса на три месяца вперед.

1. В базе доступны данные о динамике индекса по отношению к предыдущему месяцу и год к году. Какой из вариантов предпочтителен, если вам необходимо минимизировать влияние сезонных колебаний?
2. Используя расширенный тест Дики — Фуллера определите порядок интегрированности временного ряда.
3. Если по итогам предыдущего пункта моделируемый временной ряд оказался нестационарным, то перейдите к стационарным разностям. Далее постройте графики автокорреляционной и частной автокорреляционной функций.
4. Используя результаты предыдущих пунктов, а также информационные критерии Акаике и Шварца, выберите наилучшую спецификацию ARIMA-модели для описания динамики моделируемого индекса изменения стоимости строительно-монтажных работ.
5. Используя выбранную на предыдущем пункте модель, постройте прогноз динамики индекса на один, два и три месяца вперед. Оцените точность прогноза, построив соответствующие доверительные интервалы.