А. А. Цыплаков, Н. М. Ибрагимов

**Лекции по курсу**

**Эконометрика: Регрессионный анализ**

под редакцией

В. И. СУСЛОВА и Л. П. ТАЛЫШЕВОЙ

Новосибирск

2021

А. А. Цыплаков, Н. М. Ибрагимов.

Лекция по курсу Эконометрика: регрессионный анализ.

Учебное пособия

под редакцией В. И. Суслова и Л. П. Талышевой

Новосибирск, НГУ, 2021 г.— 1\_\_ с.

Пособие предназначено для студентов и преподавателей экономических факультетов. Подготовлено по материалам лекционного курса, читаемого на экономическом факультете НГУ.

Авторы выражают искреннюю благодарность и признательность А. В. Костину, Е. А. Шильцину и Е. А. Гайворонской за помощь в работе и ценные замечания при подготовке данного пособия.

© Экономический факультет НГУ, 2021

Оглавление

[Лекция: Введение в анализ статистических данных 9](#_Toc94641029)

[Контрольные вопросы 23](#_Toc94641030)

[Экзаменационные вопросы 24](#_Toc94641031)

[Литература 25](#_Toc94641032)

[Лекция: Алгебра линейной регрессии 26](#_Toc94641033)

[Уравнение линейной регрессии 26](#_Toc94641034)

[Метод наименьших квадратов и его свойства 26](#_Toc94641035)

[Показатели точности подбора 28](#_Toc94641036)

[Пример: Аппроксимация для чистой приведенной стоимости 29](#_Toc94641037)

[Центрирование регрессии. Дисперсионное тождество 31](#_Toc94641038)

[Центрирование регрессии 31](#_Toc94641039)

[Дисперсионное тождество 32](#_Toc94641040)

[Коэффициент детерминации как квадрат коэффициента корреляции 33](#_Toc94641041)

[Геометрическая интерпретация МНК 33](#_Toc94641042)

[Случай неединственности решения 34](#_Toc94641043)

[Использование линейной регрессии и метода наименьших квадратов 35](#_Toc94641044)

[Пример: Прогнозирование объема торгов USDRUB\_TOM на Московской бирже 35](#_Toc94641045)

[Пример: Кривая Филлипса по США 37](#_Toc94641046)

[Пример: Зависимость между частотой питателя и скоростью подачи руды 38](#_Toc94641047)

[Пример: Использование регрессии для расчета коэффициента бета 39](#_Toc94641048)

[Контрольные вопросы: 41](#_Toc94641049)

[Экзаменационные вопросы 41](#_Toc94641050)

[Литература 42](#_Toc94641051)

[Лекция: – Классическая модель линейной регрессии 43](#_Toc94641052)

[Классическая модель линейной регрессии 43](#_Toc94641053)

[Оценки и идентификация 45](#_Toc94641054)

[Несмещенность оценок МНК 46](#_Toc94641055)

[Предположения Гаусса–Маркова. Свойство НЛНО 46](#_Toc94641056)

[Контрольные вопросы 50](#_Toc94641057)

[Экзаменационные вопросы 51](#_Toc94641058)

[Литература 51](#_Toc94641059)

[Лекция: – Следствия нормальности. Тестирование гипотез для коэффициентов 52](#_Toc94641060)

[Следствия нормальности 52](#_Toc94641061)

[Тестирование гипотезы для одного коэффициенты 52](#_Toc94641062)

[Тестирование гипотез для нескольких коэффициентов 56](#_Toc94641063)

[Статистическая значимость и практическая значимость 61](#_Toc94641064)

[Представление результатов регрессии в компьютерных программах 65](#_Toc94641065)

[Приложение. Вывод оценок МНК с линейными ограничениями 67](#_Toc94641066)

[Контрольные вопросы 68](#_Toc94641067)

[Экзаменационные вопросы 69](#_Toc94641068)

[Литература 69](#_Toc94641069)

[Лекция: Выбор регрессоров. Мультиколлинеарность 70](#_Toc94641070)

[Выбор регрессоров. Информационные критерии (AIC, BIC) 70](#_Toc94641071)

[Выбор переменных с помощью проверки гипотез 70](#_Toc94641072)

[Выбор регрессоров с помощью числовых показателей 73](#_Toc94641073)

[Проблемы с процедурами выбора регрессоров и используемыми в них статистиками 75](#_Toc94641074)

[Мультиколлинеарность 77](#_Toc94641075)

[Отчеты по моделям регрессии 80](#_Toc94641076)

[Контрольные вопросы 84](#_Toc94641077)

[Экзаменационные вопросы 84](#_Toc94641078)

[Литература 84](#_Toc94641079)

[Лекция: Фиктивные переменные – качественные объясняющие переменные в регрессии 85](#_Toc94641080)

[Фиктивные переменные 85](#_Toc94641081)

[Контрольные вопросы 90](#_Toc94641082)

[Экзаменационные вопросы 91](#_Toc94641083)

[Литература 91](#_Toc94641084)

[Лекция: – Прогнозирование в модели линейной регрессии 92](#_Toc94641085)

[Прогнозирование в классической модели линейной регрессии 92](#_Toc94641086)

[Приложение. Вывод вспомогательной регрессии для получения прогноза 94](#_Toc94641087)

[Контрольные вопросы 96](#_Toc94641088)

[Экзаменационные вопросы 96](#_Toc94641089)

[Литература 96](#_Toc94641090)

[Лекция: – Диагностика регрессии: введение 97](#_Toc94641091)

[Изучение исходных данных 97](#_Toc94641092)

[Простейший графический анализ регрессии 99](#_Toc94641093)

[Принципы проведения диагностических тестов 102](#_Toc94641094)

[Контрольные вопросы 104](#_Toc94641095)

[Экзаменационные вопросы 104](#_Toc94641096)

[Литература 104](#_Toc94641097)

[Лекция: – Функциональная форма модели 105](#_Toc94641098)

[Регрессия как линейная модель 105](#_Toc94641099)

[Что такое «линейная регрессия»? 105](#_Toc94641100)

[Линейные преобразования модели 106](#_Toc94641101)

[Разные функциональные формы и их интерпретация 107](#_Toc94641102)

[Некоторые простые функциональные формы 107](#_Toc94641103)

[Интерпретация коэффициентов в парной регрессии 111](#_Toc94641104)

[Интерпретация коэффициентов множественной регрессии 114](#_Toc94641105)

[Взаимодействия переменных 117](#_Toc94641106)

[Ошибки спецификации функциональной формы 119](#_Toc94641107)

[Последствия неверной спецификации функциональной формы 119](#_Toc94641108)

[Диагностика неверной функциональной формы 120](#_Toc94641109)

[Непостоянство коэффициентов модели. Тесты Чоу 124](#_Toc94641110)

[Непостоянство коэффициентов модели 124](#_Toc94641111)

[Тест Чоу на постоянство коэффициентов 126](#_Toc94641112)

[Тест Чоу на точность прогноза (2-й тест Чоу) 127](#_Toc94641113)

[Приложение. Линейные преобразования в регрессии 128](#_Toc94641114)

[Контрольные вопросы 129](#_Toc94641115)

[Экзаменационные вопросы 129](#_Toc94641116)

[Литература 129](#_Toc94641117)

[Лекция: – Регрессия с ковариационной матрицей ошибок общего вида 131](#_Toc94641118)

[МНК в случае ковариационной матрицы ошибок общего вида 131](#_Toc94641119)

[Последствия нарушения предположений 132](#_Toc94641120)

[Обобщенный МНК 132](#_Toc94641121)

[Доступный обобщенный МНК 134](#_Toc94641122)

[Подходы к моделированию в случае ковариационной матрицы ошибок общего вида 135](#_Toc94641123)

[Контрольные вопросы 135](#_Toc94641124)

[Экзаменационные вопросы 136](#_Toc94641125)

[Литература 136](#_Toc94641126)

[Лекция: – Гетероскедастичность ошибок 137](#_Toc94641127)

[Гетероскедастичность ошибок – определение и последствия 137](#_Toc94641128)

[Диагностика гетероскедастичности 138](#_Toc94641129)

[Взвешенный МНК 142](#_Toc94641130)

[Логарифмирование переменных и гетероскедастичность 146](#_Toc94641131)

[Робастная оценка для ковариационной матрицы коэффициентов МНК 147](#_Toc94641132)

[Контрольные вопросы 148](#_Toc94641133)

[Экзаменационные вопросы 148](#_Toc94641134)

[Литература 148](#_Toc94641135)

[Лекция: – Автокорреляция ошибок 150](#_Toc94641136)

[Некоторые понятия из теории временных рядов 150](#_Toc94641137)

[Автокорреляция ошибок – определение и последствия 152](#_Toc94641138)

[Диагностика серийной корреляции 154](#_Toc94641139)

[Что делать в случае серийной корреляции? 155](#_Toc94641140)

[Другие виды автокорреляции. Пространственная автокорреляция 156](#_Toc94641141)

[Контрольные вопросы 159](#_Toc94641142)

[Экзаменационные вопросы 159](#_Toc94641143)

[Литература 159](#_Toc94641144)

[Лекция: – Отсутствие нормальности ошибок. Выбросы 161](#_Toc94641145)

[Диагностика отсутствия нормальности и наличия выбросов 161](#_Toc94641146)

[Асимптотическая нормальность, проверка гипотез и следствия отсутствия нормальности 164](#_Toc94641147)

[Что делать при сильном отклонении ошибок он нормальности? 164](#_Toc94641148)

[Приложение. Ядерная оценка плотности 167](#_Toc94641149)

[Контрольные вопросы 169](#_Toc94641150)

[Экзаменационные вопросы 169](#_Toc94641151)

[Литература 170](#_Toc94641152)

[Лекция: – Проблема эндогенности 171](#_Toc94641153)

[5.0 Классическая модель линейной регрессии – повторение 171](#_Toc94641154)

[5.1 Случайные регрессоры и экзогенность 171](#_Toc94641155)

[Случайные регрессоры – точная теория для конечного числа наблюдений 172](#_Toc94641156)

[Случайные регрессоры – асимптотическая теория 173](#_Toc94641157)

[Предположение об экзогенности регрессоров 174](#_Toc94641158)

[Последствия отсутствия экзогенности 175](#_Toc94641159)

[Экзогенность и причинность 176](#_Toc94641160)

[5.2 Основные причины эндогенности 178](#_Toc94641161)

[Пропущенные переменные 178](#_Toc94641162)

[Ошибки в переменных 180](#_Toc94641163)

[Одновременность. Системы одновременных уравнений 182](#_Toc94641164)

[5.3 Подходы к проблеме эндогенности 185](#_Toc94641165)

[Сопоставление схожих наблюдений 185](#_Toc94641166)

[Включение контрольных переменных 185](#_Toc94641167)

[Частный случай контрольных переменных – нормировка 186](#_Toc94641168)

[Метод инструментальных переменных с одной эндогенной переменной и одним инструментом 189](#_Toc94641169)

[Метод инструментальных переменных в общем случае 190](#_Toc94641170)

[Использование разрывности в регрессии 191](#_Toc94641171)

[Контрольные вопросы 192](#_Toc94641172)

[Экзаменационные вопросы 193](#_Toc94641173)

[Литература 193](#_Toc94641174)

[Приложения (базовые понятия из курсов теории вероятностей, математической статистике и высшей математики) 194](#_Toc94641175)

[Основные понятия теории вероятностей 194](#_Toc94641176)

[Проверка гипотез 197](#_Toc94641177)

[Свойства оценок: 198](#_Toc94641178)

[Распределения, используемые в эконометрии 199](#_Toc94641179)

[Равномерное распределение 199](#_Toc94641180)

[Нормальное распределение 199](#_Toc94641181)

[Распределение хи-квадрат 199](#_Toc94641182)

[Распределение Стьюдента 200](#_Toc94641183)

[Распределение Фишера 200](#_Toc94641184)

[Многомерное нормальное распределение 201](#_Toc94641185)

[Вспомогательные сведения из высшей математики. 202](#_Toc94641186)

[Матричная алгебра 202](#_Toc94641187)

[Свойства матриц 203](#_Toc94641188)

[Матричное дифференцирование 207](#_Toc94641189)

# Лекция: Введение в анализ статистических данных

В этой лекции излагается материал, с которым, как предполагается, студенты уже знакомы из предыдущих курсов. Поэтому изложение идет в достаточно свободной форме. В частности, сведения из статистики идет вперемешку со сведениями из теории вероятностей. Акцент делается на прикладной анализ статистических данных, а не на теоретические основания.

## Случай одной переменной

Предположим, что имеется набор из наблюдений за некоторой переменной :

(Часто такой набор данных называют выборкой.) Например, в табл. 1 приведены первые несколько наблюдений за некоторым экономическим показателем по регионам России. В данном случае каждое наблюдение – это некоторое действительное число. Мы хотим получить некоторое общее представление о характере имеющихся данных, но довольно большое, так что полный набор чисел сложно охватить одним взглядом. Рассмотрим основные инструменты и понятия, позволяющие исследовать такую переменную .

Таблица 1. Общая площадь жилых помещений, приходящаяся в среднем на одного жителя по регионам России на конец 2019 года, м2 (6 первых наблюдений)

|  |  |
| --- | --- |
| Белгородская область | 32,1 |
| Брянская область | 30,5 |
| Владимирская область | 30,0 |
| Воронежская область | 30,8 |
| Ивановская область | 27,3 |
| Калужская область | 31,0 |
|  |  |
|  |  |

На помощь здесь может прийти графическое представление данных. Например, на рис. 1 имеющиеся данные представлены точками на действительной прямой. По этому графику несложно получить некоторое первоначальное представление о том, с какими данными мы имеем дело. Однако такое описание данных не всегда будет удачным. В частности, точки могут накладываться друг на друга.

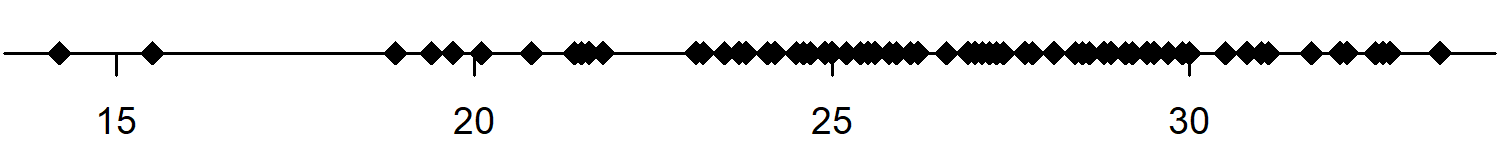


Рисунок 1. Представление данных точками на прямой

Можно также построить точечную диаграмму от номера наблюдения (рис. 2). Такой график будет в некотором смысле произвольным, поскольку порядок наблюдений потенциально может быть любым. В некоторых случаях наблюдения выстроены по какому-то принципу, так что график может выявить интересные особенности. В частности, если это временной ряд, то график по номеру наблюдения будет, фактически, графиком временного ряда по времени. В рассматриваемом случае регионы сгруппированы по федеральным округам и можно увидеть различия в значениях по этим группам.

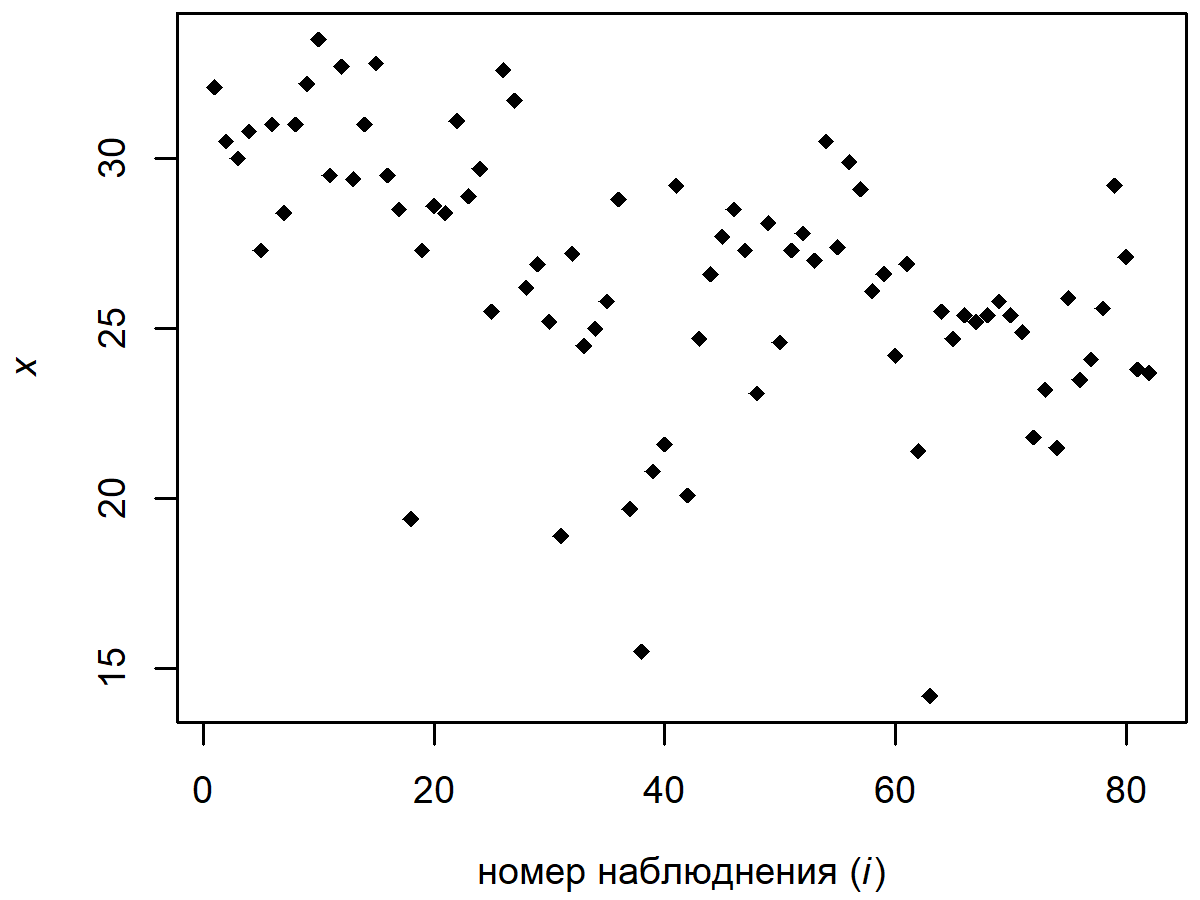


Рисунок 2. Точечная диаграмма по номеру наблюдения

При любом порядке наблюдений графике по дает общее представление о том, какие значения принимает .

Позже мы рассмотрим также более сложное графическое описание данных – гистограмму.

Кроме графиков, можно использовать числовые обобщающие показатели для набора данных. Такие показатели называют **описательными статистиками**.

Простейшие показатели – это **минимальное и максимальное значение** :

По ним мы можем получить информацию о диапазоне значений, принимаемых переменной .

Также довольно простым и понятным показателем является **выборочное среднее**:  
Этот показатель одним числом характеризует местоположение рассматриваемой выборки – это некоторая характеристика ее центра. В векторных обозначениях  
где  
– вектор-столбец наблюдений,  
– вектор-столбец из единиц.

Для рассматриваемого примера среднее по регионам равно . На рис. 3 среднее точек показано вертикальной линией.

Поскольку наблюдения разные, в наших данных имеется разброс наблюдений относительно среднего. Определенное представление о том, насколько широко разбросаны точки, дает интервал и соответствующий показатель размаха выборки , но этот показатель не очень устойчив к случайным выбросам, т. е. резко выделяющимся наблюдениям.

Более удачный и широко используемый показатель разброса – это среднее квадратов расстояний от точек до среднего. Данный показатель ( или ) называется **выборочной дисперсией**:

Мы можем рассмотреть так называемые центрированные наблюдения  
и составить из них вектор :  
В этих обозначениях

Заметим, что при расчете дисперсии можно использовать эквивалентную формулу  
(средний квадрат минус квадрат среднего).

Дисперсия всегда неотрицательна, и может быть равна нулю только когда все совпадают (когда разброс отсутствует). Чем больше дисперсия, тем шире разбросана выборка относительно среднего.

Показатель дисперсии может быть не очень удобен своими единицами измерения. Если измеряется в некоторых единицах ₽, то дисперсия измеряется в ₽2. Поэтому наряду с дисперсией используется также **среднеквадратическое отклонение**. Это среднее квадратическое отклонений точек от :  
Среднеквадратическое отклонение измеряется в тех же единицах, что и сама переменная .

Для рассматриваемого примера дисперсия равна , а среднеквадратическое отклонение . На рис. 3 наряду со средним показаны интервалы плюс-минус 1 и 2 стандартных отклонения по сторонам от . Как видим, для этих данных в интервал плюс-минус 1 стандартное отклонение попадает бо́льшая часть точек, а за пределами интервала плюс-минус 2 стандартных отклонения лежат только редкие точки.

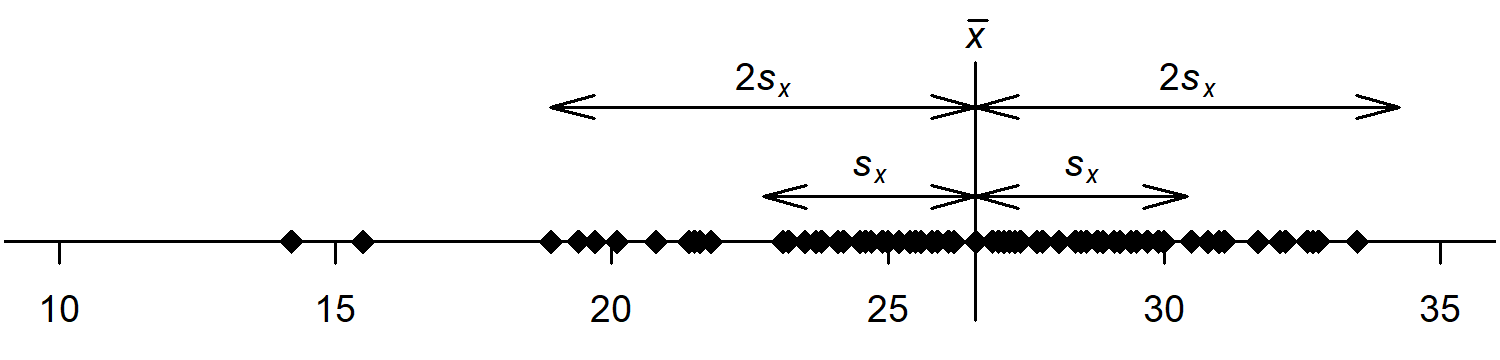


Рисунок 3. Среднее и среднеквадратическое отклонение

Среднее и дисперсия являются частными случаями семейства описательных статистик, называемых **выборочными моментами**. Выборочный момент порядка вычисляется по формуле

При это **центральный момент**, а при – **начальный (нецентральный) момент**.

Очевидно, что среднее – это начальный момент порядка (), а дисперсия – это центральный момент 2-го порядка ().

Заметим, что определенная выше выборочная дисперсия – это так называемая «смещенная дисперсия». Некоторые статистические программы могут рассчитывать «несмещенную дисперсию», в которой сумма квадратов делится не на , а на :

Также иногда используют соответствующий вариант среднеквадратического отклонения . Строго говоря, эти термины – «смещенная» и «несмещенная» – приобретают конкретный смысл только в рамках некоторой статистической модели, описывающей выборку, поэтому их надо рассматривать как неформальные. По крайней мере, следует знать о существовании такого показателя и быть готовым к тому, что в некоторой компьютерной программе могут встретиться и , а не и .

Для дальнейшего обсуждения полезных описательных статистик введем в рассмотрение вероятностную *модель* для изучаемых данных. А именно, нам удобно предположить, что каждое из наблюдений является реализацией некоторой **случайной величины** , имеющей непрерывное распределение на некотором интервале действительной прямой (такой интервал потенциально возможных значений называется **носителем** распределения).

Как известно, непрерывную[[1]](#footnote-1) случайную величину можно описать с помощью функции **плотности** (рис. 4)

На интервале-носителе плотность (почти всюду) положительна. Удобно считать, что плотность определена на всей действительной прямой, и за пределами носителя принять ее равной нулю. На интуитивном уровне для данной точки плотность показывает, насколько часто случайная величина оказывается рядом с точкой . Если плотность маленькая, то случайная величина относительно редко оказывается рядом с точкой , если большая, то относительно часто. Функция плотности подбирается так, чтобы ее интеграл был равен 1:

плотность

носитель

Рисунок 4. Пример плотности

Например, для **равномерного распределения** на интервале () плотность берется постоянной. Пусть это величина . За пределами интервала плотность 0. Тогда указанный интеграл равен ) и следует взять , чтобы он был равен 1 (рис. 5).

Рисунок 5. Плотность равномерного распределения

Для **нормального распределения** носитель равен , а плотность берется пропорциональной . Это симметричная колоколообразная кривая с вершиной в точке (рис. 6). Чтобы получить интеграл 1, следует разделить указанную функцию на . Таким образом, функция плотности нормального распределения равна

Рисунок 6. Плотность нормального распределения

Распределение случайной величины можно описать с помощью различных характеристик. В чём-то эти характеристики похожи на описательные статистики для выборки. В частности, **математическое ожидание** (теоретическое среднее) является теоретическим аналогом выборочного среднего и характеризует центр распределения. В случае непрерывно распределенной случайной величины оно равно интегралу

Для равномерного распределения имеем Для нормального распределения имеем .

**Дисперсия** случайной величины определяется как математическое ожидание квадрата центрированной случайной величины, где под центрированием понимается вычитание из случайной величины ее теоретического среднего:

Эквивалентно дисперсию можно рассчитать по формуле

Теоретическое **среднеквадратическое отклонение** равно

Дисперсия и среднеквадратическое отклонение характеризуют степень разброса распределения относительно среднего.

Для нормального распределения имеем . Таким образом, два параметра нормального распределения ( и ) – это математическое ожидание и дисперсия.

Также мы можем определить **теоретические моменты**. Момент *q*-го порядка – это

При это начальный (нецентральный) момент, а при – центральный момент.

Математическое ожидание является начальным моментом первого порядка: , а дисперсия – центральным моментом второго порядка: .

Кроме показателей, характеризующих центр и ширину разброса распределения, есть также показатели, которые характеризуют форму распределения. Это, в частности, асимметрия и куртозис.

**Асимметрия** определяется как  
Этот показатель для симметрично распределенной переменной равен нулю. Если он положительный, то имеет место правая асимметрия (тогда, как правило, правый хвост длиннее или толще левого), если отрицательный, то левая асимметрия (наоборот).

**Куртозис** определяется как

У нормального распределения он равен . Если куртозис больше , то распределение, как правило, имеет острую вершину и толстые хвосты, а если меньше , то уплощенную вершину и быстро затухающие хвосты. Показатель **эксцесса** получают вычитанием из куртозиса 3 – это превышение куртозиса над «нормальным» куртозисом.

**Функцией распределения** случайной величины называется функция  
сопоставляющая числу вероятность того, что случайная величина не превышает z. Функция распределения имеет следующие свойства: это неубывающая, непрерывная справа функция, , причем и . Если случайная величина непрерывна, то она имеет плотность , которая связана с функцией распределения соотношениями (рис. 7)

Таким образом, в терминах плотности величина – это площадь под кривой плотности левее точки . Другими словами, это вероятность, соответствующая левому хвосту.

1

Рисунок 7. Функция плотности и соответствующая функция распределения

Функция распределения для данной границы указывает вероятность. Можно рассмотреть и обратную задачу – найти для данной вероятности границу. Такие границы называются квантилями. Квантилью уровня , где , (-квантилью) непрерывной случайной величины называется число , такое что (рис. 7)  
В области, где функция распределения возрастает, мы можем определить обратную к ней (так называемую **квантильную функцию**) и тогда

**Медианой** называется -квантиль. Она делит распределение на две части – справа и слева половина плотности (рис. 8).

1

Рисунок 8. Медиана непрерывного распределения

Если плотность распределения имеет единственный максимум, то можно ввести еще одну характеристику центра распределения, соответствующую точке наибольшей плотности. Такая характеристика называется **модой** (рис. 9):

Рисунок 9. Мода распределения – точка наибольшей плотности

В дальнейшем нам понадобится также определение двухсторонних квантилей. Если распределение непрерывной случайной величины *симметрично* относительно нуля, т. е.  
для всех , то **двусторонней квантилью** уровня называется число , такое что вероятность попадания в интервал равна , т. е.

Дав определения для теоретических характеристик распределения непрерывной случайной величины, мы можем рассмотреть аналогичные характеристики (например, квантили) для выборки.

Опредение **выбочных квантилей** основано на упорядоченной выборке. Пусть – -е по порядку возрастание наблюдение, так что

Тогда выборочная ‍-квантиль – это такая граница, что примерно доля от наблюдений находится левее границы, а примерно доля – правее. Приближенно квантиль уровня следует искать около наблюдения . Мы не будем здесь давать точное определение, заметим только, что по смыслу выборочная квантиль – это не вполне однозначное понятие. Так для наблюдений в качестве медианы можно взять любую точку отрезка .

Для общего описания выборки можно взять несколько разных квантилей. Например, часто рассматривают **квартили**, то есть квантили уровней , и . Минимум и максимум можно в определенном смысле считать квартилями (квантилями уровней и соответственно). **Межквартильный размах** является еще одним показателем разброса выборки. Более детальную информацию о выборке дают **децили**, для которых шаг вероятности равен 0.1.

В табл. 2 приведены несколько основных описательных статистик для выборки. Подобная таблица представляет собой удобное компактное описание данных. В частности, из таблицы видно, что в целом наши данные лежат в диапазоне , причем примерно половина данных лежит в промежутке .

Таблица 2. Описательные статистики для данных

|  |  |
| --- | --- |
| Среднее | 26,58 |
| Дисперсия | 14,76 |
| Среднекв. отклонение | 3,84 |
| Минимум | 14,20 |
| Квантиль 25% | 24,70 |
| Медиана | 26,95 |
| Квантиль 75% | 29,20 |
| Максимум | 33,50 |

Если рассматриваемые наблюдения являются выборкой из какого-то непрерывного распределения, то оценка плотности этого распределения может служить полезным инструментом для визуального анализа. Даже если анализируемые данные не получены на основе с какого-либо непрерывного распределения, мы все равно можем построить оценку плотности, которая дает представление о расположении точек в нашей выборке. Самой распространенной оценкой плотности является гистограмма (рис. 10).

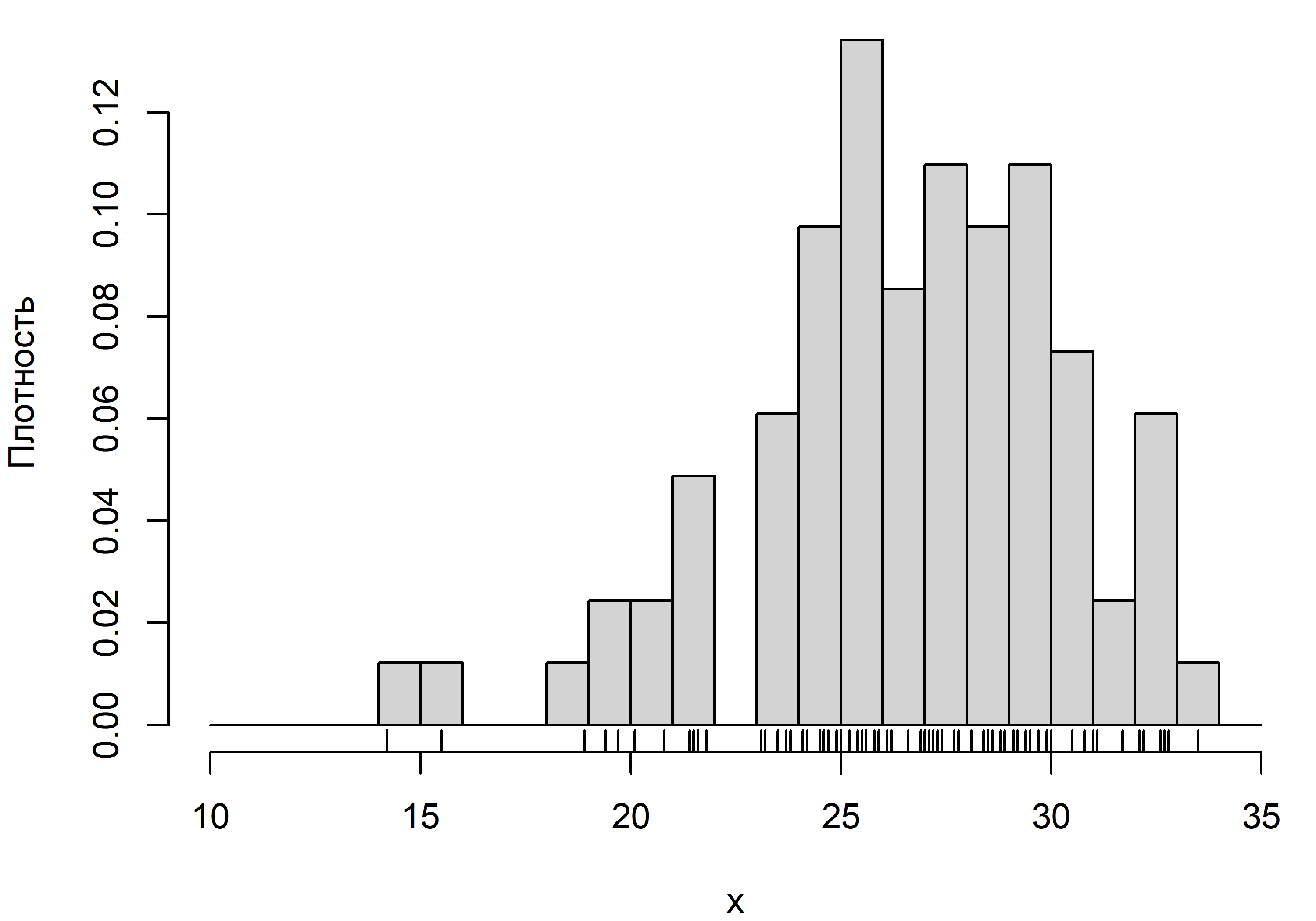


Рисунок 10. Гистограмма данных

Для построения гистограммы выбираются левая и правая границы и , такие что  
Затем отрезок от до разбивается на равных интервалов шириной

Интервалы имеют вид  
Для каждого интервала рассчитывается количество наблюдений , которые в него попадают. Тогда оценка плотности для точек из интервала равна  
За пределами отрезка оценка плотности равна 0.

По гистограмме можно судить о распределении точек выборки в целом. В частности, по ней можно судить о форме распределения. Форму распределения переменной также можно охарактеризовать с помощью выборочных асимметрии и куртозиса. Коэффициент **асимметрии** (скошенности) определяется как

Для симметрично распределенной переменной он около нуля. Коэффициент **куртозиса** определяется как

Соответственно, коэффициент эксцесса меньше на 3. Если распределение переменной близко к нормальному, то выборочный эксцесс примерно равен .

## Случай двух и более переменных

Пусть у нас есть наблюдения за двумя переменными и . Это пар наблюдений

В табл. 3 приведен пример подобных данных в табличном виде.

Таблица 3. Климатическая норма для средней температуры января (°С) и среднемесячная начисленная заработная плата работников организаций (тыс. руб.) по регионам России в 2019 году (6 первых наблюдений)

|  |  |  |
| --- | --- | --- |
| Регион | Температура января | Зарплата |
| Белгородская область | -4,0 | 34,6 |
| Брянская область | -3,8 | 29,9 |
| Владимирская область | -7,3 | 33,1 |
| Воронежская область | -4,9 | 33,7 |
| Ивановская область | -5,9 | 27,6 |
| Калужская область | -5,2 | 41,4 |
|  |  |  |
|  |  |  |

Начать анализ подобных данных можно с анализа каждой из переменных отдельно. Мы можем рассчитать те же самые характеристики, которые обсуждали выше – средние (, ), дисперсии (, ), среднеквадратические отклонения, построить гистограмму, и т. д.

Помимо этого, и могут иметь какие-то совместные характеристики, которые могут отражать взаимозависимость между ними. Чтобы получить представления о взаимосвязях, на начальном этапе анализа можно построить точечную диаграмму (диаграмму рассеяния) в координатах (рис. 11).

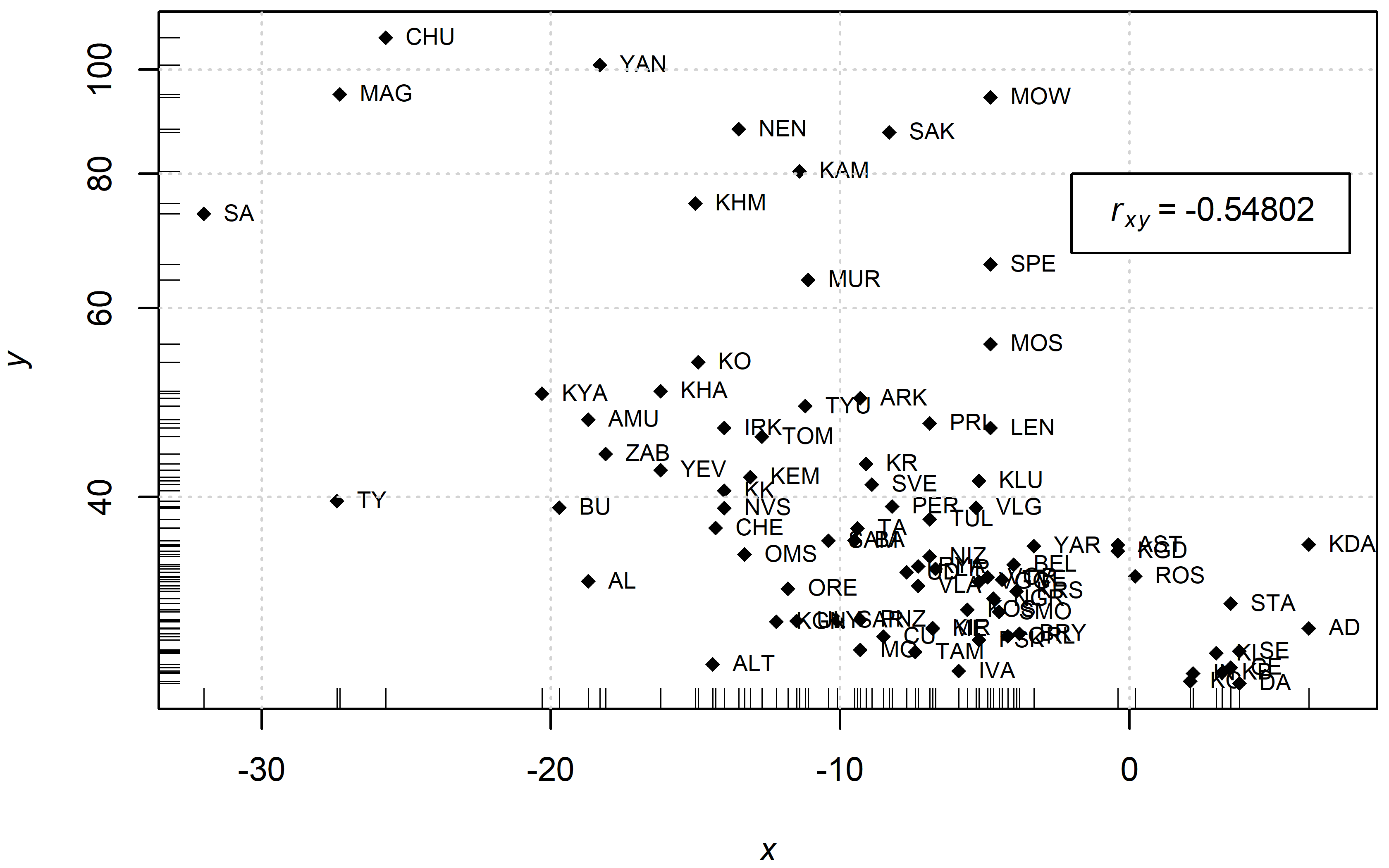


Рисунок 11. Точечная диаграмма с кодами регионов

На рис. 11 в качестве взята температура января, а в качестве – логарифм заработной платы. (Для удобства интерпретации для оси изображена логарифмическая шкала в тысячах рублей.) В частности, на графике можно увидеть, что в более холодных регионах зарплаты в среднем выше, чем в более теплых.

Существуют различные выборочные характеристики, измеряющие тесноту связи между переменными. Простейший показатель такого рода – это выборочная ковариация, показатель тесноты линейной связи между и :  
Здесь , – центрированные значения для двух переменных. В векторной записи  
где , – центрированные векторы и . Можно использовать эквивалентную формулу  
(среднее произведение минус произведение средних). Ковариация симметрична относительно переменных и :

Выборочная ковариация аналогична теоретической ковариации двух случайных величин, которая определяется как  
или, эквивалентно,

Ковариация как измеритель тесноты связи может быть неудобна тем, что она зависит от единиц измерения обеих переменных. Поэтому наряду с ковариацией используют безразмерный показатель или – выборочный коэффициент корреляции и , определяемый как  
(предполагаем, что , ).

Значение коэффициента корреляции всегда принадлежит отрезку . Как и ковариация, он симметричен относительно переменных и ().

Если , то между и существует линейная зависимость:  
причем при и при . График облака наблюдений в координатах – прямая линия с положительным (соответственно, отрицательным) наклоном.

Для рассматриваемых данных ковариация равна , а корреляция . Таким образом, в данных есть довольно существенная отрицательная корреляция между температурой и логарифмом заработной платы.

Рассмотрим теперь случай, когда имеются данные по нескольким переменным.

Пусть есть наблюдения за переменными

Каждую переменную можем представить вектором-столбцом  
а все переменные вместе – матрицей данных размерности :

Как и в случае двух переменных, разумно начать анализ подобных данных с анализа каждой переменной отдельно, рассчитав ее среднее , среднеквадратическое отклонение , и другие статистики, построив различные графики и т. д.

Кроме того, можно рассчитать различные статистики для пар переменных. В частности, можно рассчитать выборочные ковариации для всех возможных пар :  
где – центрированное наблюдение для переменой , а  
– составленный из центрированных значений вектор. Центрирование проводится по-отдельности для каждой переменной.

Составленную из выборочных ковариаций квадратную матрицу   
называют **выборочной ковариационной матрицей** (англ. *variance-covariance matrix*). Ее можно вычислять по матричной формуле  
где – центрированная (по столбцам) матрица :  
Ковариационная матрица является симметричной и положительно полуопределенной (положительно определенной, если переменные линейно независимы). По диагонали матрицы стоят выборочные дисперсии переменных .

Так же точно из корреляций можно составить корреляционную матрицу  
Эта матрица симметричная и положительно полуопределенная, с единицами по диагонали. Такую матрицу в статистических программах выводят в виде таблицы.

В качестве примера рассмотрим темпы прироста цен на молоко в четырех российских регионах: Белгородская область, Алтайский край, Новосибирская область и Приморский край (коды BEL, ALT, NVS, PRI) за несколько лет. На рис. 12 приведены графики четырех рядов.

Таблица 4. Описательные статистики для 4 рядов

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | среднекв. |  |  | таблица корреляций | | | |
|  | среднее | отклонение |  |  | BEL | ALT | NVS | PRI |
| BEL | 0.602 | 1.387 |  | BEL | 1.000 | 0.298 | 0.390 | 0.300 |
| ALT | 0.682 | 2.848 |  | ALT | 0.298 | 1.000 | 0.465 | 0.152 |
| NVS | 0.647 | 1.729 |  | NVS | 0.390 | 0.465 | 1.000 | 0.146 |
| PRI | 0.547 | 0.951 |  | PRI | 0.300 | 0.152 | 0.146 | 1.000 |

В таблице 4 даны основные статистики для четырех переменных. Все переменные по смыслу близки и выражены в одних и тех же единицах, поэтому можно сравнивать их средние и среднеквадратические отклонения между собой. По среднемесячному темпу прироста можно судить о том, в каком регионе цены на молоко росли быстрее, а в каком медленнее. По среднеквадратическому отклонению можно судить о том, в каком регионе наблюдались более резкие колебания цены на молоко (это Алтайский край). Из таблицы корреляций можно сделать вывод, что сильнее всего связаны цены на молоко в Алтайском крае и Новосибирской области.

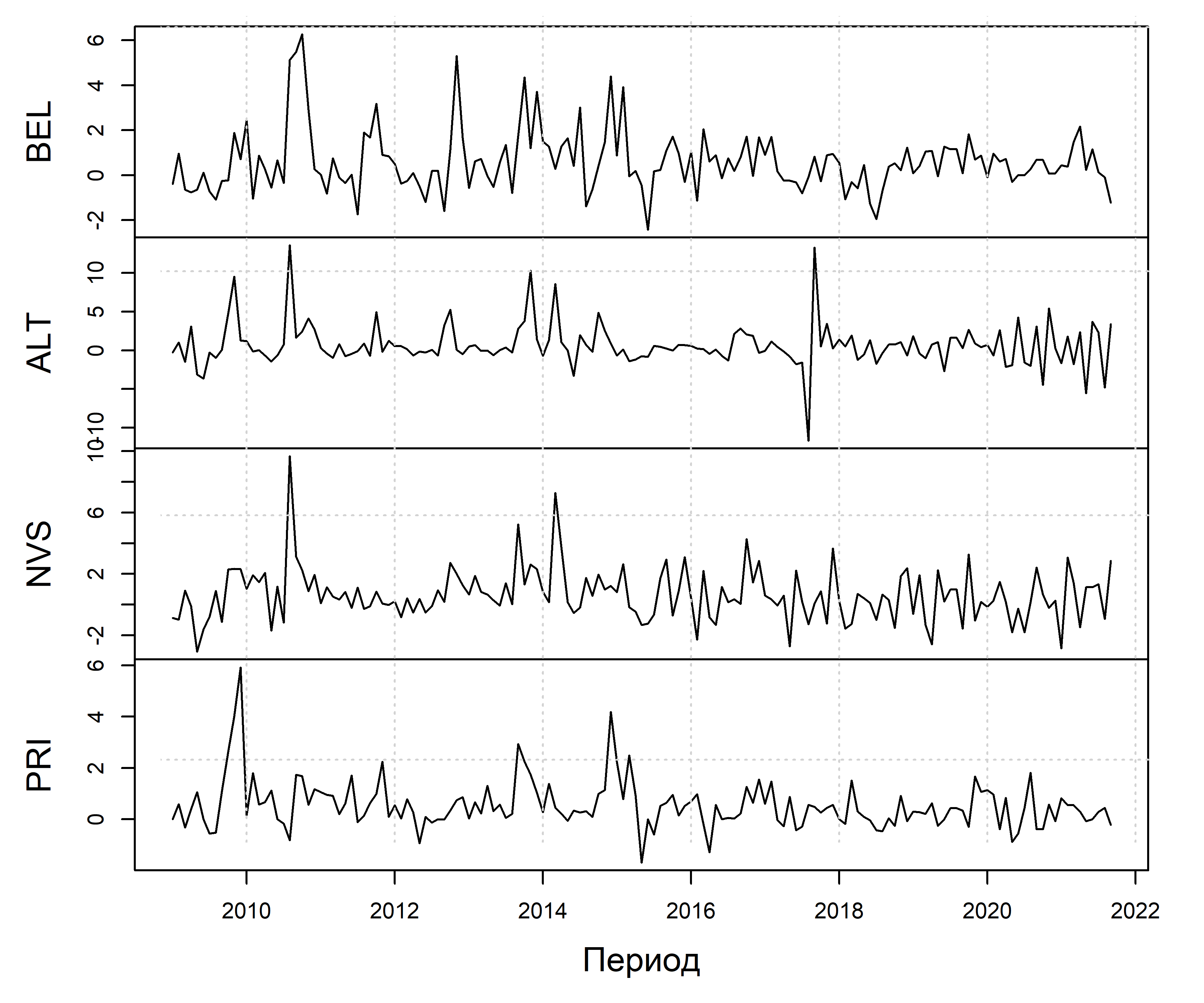


Рисунок 12. Темпы прироста цен молока в 4 регионах

## Контрольные вопросы

1. Для случайных величин и заданы следующие значения:, , , Для случайных величин и вычислите:
   1. , ,,
   2. Можно ли утверждать, что случайные величины и V независимы?
2. Пусть случайные величины независимы и нормально распределены , и
3. Рассмотрим вектор центрированных значений. Представьте вектор ввиде Как выглядит матрица ?
4. Является ли матрица симметричной? Идемпотентной?
5. Представьте скаляр в виде . Как выглядит матрица ?
6. Представьте скаляр в виде . Как выглядит матрица ?
7. Представьте скаляр в виде , где Как выглядит матрица ?
8. Является ли матрица симметричной? Идемпотентной?
9. Покажите, что .
10. Известно, что - вектор-столбец размерности , – матрица , и .
    1. Какую размерность имеет .
    2. Найдите ,.
    3. по вектору
    4. по вектору
11. Известно, что – матрица и , при этом обратима. Рассмотрим , где – единичная матрица
    1. Укажите размер матрицы ,
    2. Найдите,, .
12. Известно, что – матрица . Пусть. Докажите, что матрица имеет полный ранг по столбцам, т. е

## Экзаменационные вопросы

1. Характеристики распределений: функция распределения, плотность, квантили (односторонние, двусторонние), моменты: начальные, центральные моменты, математическое ожидание, дисперсия, асимметрия, куртозис, эксцесс.
2. Характеристики многомерных распределений: совместная функция распределения, плотность, меры связи случайных величин (ковариация, корреляция), матрица ковариаций и корреляций.
3. Распределения, используемые в эконометрии: равномерное, нормальное, хи-квадрат, Стьюдента, Фишера. Где и как в курсе «Эконометрии» используются распределения: нормальное, хи-квадрат, Стьюдента, Фишера.
4. Описательные статистики: выборочные моменты, среднее, дисперсия, асимметрия, куртозис, эксцесс. Примеры
5. Меры связи: ковариация, корреляция. Свойства ковариационной и корреляционной матрицы переменных . Где и как в курсе «Эконометрии» используются эти понятия и свойства.
6. Проверка гипотез: нулевая и альтернативная гипотеза, статистика, критическая область, критическая граница, уровень значимости, уровень доверия, ошибки первого и второго рода. Примеры использования этих понятий в курсе «Эконометрии».
7. Свойства оценок: несмещенность, состоятельность и эффективность оценок. Примеры использования этих понятий в курсе «Эконометрии».
8. При доказательстве каких утверждений и как в курсе «Эконометрии» используются
9. Правиломатричного дифференцирования;
10. Свойства следа матрицы;
11. Свойства положительно полуопределенныхматриц;
12. Свойства идемпотентных матриц;
13. Произведение Кронекера.

## Литература

1. Суслов В. И., Ибрагимов Н. М., Талышева Л. П., Цыплаков А. А. *Эконометрия*. – Новосибирск: СО РАН, 2005. [*стр. 691-714*]
2. Вербик М. Путеводитель по современной эконометрике. – М.: Научная книга,. 2008. [стр. 565-592]
3. Магнус Я. Р., Катышев П. К., Пересецкий А. А. Эконометрика. Начальный курс:– Учебник. – 6-е изд., перераб. и доп. – М.: Дело, 2004. [стр. 484-541]

# Лекция: Алгебра линейной регрессии

## Уравнение линейной регрессии

Пусть имеется наблюдений за переменной и за переменными , . Здесь — номер наблюдения.

Рассмотрим задачу аппроксимации некоторой переменной с помощью линейной комбинации переменных . Такая аппроксимация называется **линейной регрессией**:

В этом уравнении , – коэффициенты регрессии. Коэффициент – **константа**, а , – коэффициенты наклона. Переменную называют объясняемой или **зависимой**, а переменные – **объясняющими переменными** или независимыми переменными, а также **регрессорами**. Наконец, – это ошибка данной аппроксимации, называемая **остатком**.

То же уравнение бывает удобно записать в векторной форме:

Здесь

– векторы-столбцы длины , – вектор объясняемой переменной, – вектор, составленный из единиц (), – вектор, составленный из наблюдений за -й объясняющей переменной, – вектор остатков.

Еще более компактно можно записать уравнение регрессии в матричной форме:

Здесь

– матрица объясняющих переменных,

– вектор-столбец коэффициентов регрессии длины .

При данном значении вектора коэффициентов остатки равны

## Метод наименьших квадратов и его свойства

Метод наименьших квадратов (МНК, или OLS, *ordinary least squares*) состоит в том, чтобы подобрать коэффициенты на основе задачи минимизации суммы квадратов остатков ( — *residual sum of squares*) по :

Сумма квадратов остатков в матричном виде:

Используя матричное дифференцирование, можно вычислить вектор первых производных суммы квадратов остатков по коэффициентам регрессии:

Из этих условий первого порядка минимума после несложных преобразований выводим  
или  
Данную систему линейных уравнений обычно называют **нормальными уравнениями**.

Для того, чтобы убедиться, что нормальные уравнения действительно определяют минимум, нужно проверить, что матрица вторых производных (матрица Гессе) положительно полуопределена. Матрица Гессе для равна

Эта матрица не зависит от , всегда одна и та же. Кроме того, она положительно полуопределена. Поскольку матрица Гессе всюду положительно полуопределена, то условия 1-го порядка определяют глобальный минимум суммы квадратов остатков .

Предположение гарантирует единственность минимума, поскольку тогда матрица Гессе положительно определена. Данное условие эквивалентно тому, что матрица имеет полный ранг по столбцам  
т. е. оно выполнено тогда и только тогда, когда регрессоры (столбцы матрицы ) линейно независимы. Далее мы, как правило, будем считать, что это предположение выполнено (невырожденный случай).

Решая нормальные уравнения  
в предположении невырожденности находим вектор коэффициентов МНК:

Остатки МНК равны  
или  
где – **расчетные значения** зависимой переменной :

Из условия первого порядка  
видно, что остатки ортогональны регрессорам:  
или

В частности, так как первый стольбец матрицы является вектором из единиц, то  
Таким образом, сумма остатков равна нулю. Среднее остатков тоже равно нулю:

## Показатели точности подбора

После вычисления коэффицентов и остатков МНК можно посмотреть насколько точно функция регрессии аппроксимирует зависимую переменную . Простейший подобный показатель – это **сумма квадратов остатков** (или остаточная сумма квадратов, англ. *residual sum of squares* или *sum of squared residuals*):

Это значение целевой функции задачи наименьших квадратов в точке минимума.

Похожий показатель – это **остаточная дисперсия**

Это средний квадрат для остатков. С другой стороны, это выборочная дисперсия остатков, (При ее вычислении остатки не требуется центрировать, так как они уже центрированы.)

Можно взять корень из остаточной дисперсии, получив среднеквадратическое отклонение остатков, или как его называют, **стандартную ошибку регрессии**:

Этот показатель рассчитывается в тех же единицах, что и , поэтому по нему может быть более удобно, чем по или , судить, насколько точна наша аппроксимация.

Другой часто используемый показатель точности подбора – это **коэффициент детерминации** . Он показывает на сколько процентов ниже по сравнению с регрессией, в которой есть только константа:

Можно дать определение и через дисперсию – на сколько процентов остаточная дисперсия ниже по сравнению с регрессией, в которой есть только константа:

В знаменателе здесь стоит сумма квадратов остатков (соответственно, остаточная дисперсия) из регрессии, в которой нет объясняющих переменных (), а есть только константа. Действительно, в этом случае имеем

Поэтому

Коэффициент детерминации – это безразмерный показатель, принимающий значения в промежутке (от 0 % до 100 %). При объясняющие переменные не дают ничего для улучшения точности подбора по сравнению с константой. При регрессия без ошибок описывает .

#### Пример: Аппроксимация для чистой приведенной стоимости

В табл. 5 приведены исходные данные – денежные потоки в млрд руб. – по некоторому условному проекту, рассчитанному на 7 лет.

Таблица 5. Денежные потоки по проекту (млрд руб.)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Год () |  |  |  |  |  |  |  |
|  | -45 | 10 | 11 | 12 | 13 | 14 | 15 |

Чистая приведенная стоимость (NPV) по проекту рассчитывается по формуле  
где – процентная ставка (ставка дисконтирования) в процентах. В табл. 6 указаны значения (округленные до 6 знаков после запятой), рассчитанные при разных значениях ставки .

Таблица 6. по проекту при разных процентных ставках

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 |
|  | 10 | 12 | 14 | 16 | 18 | 20 |
|  | 8.236778 | 5.044245 | 2.137734 | -0.514626 | -2.940567 | -5.164288 |

Здесь рассчитывается по нелинейной формуле, но мы можем приближенно связать c с помощью линейной регрессии и получить более простую аппроксимирующую формулу.

Для зависимости от построены две аппроксимирующие модели. В Модели 1 одна объясняющая переменная – . В Модели 2 две объясняющие переменные – и . Результаты подбора регрессии методом наименьших квадратов даны в табл. 7.

Таблица 7. Две аппроксимации для NPV

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Подобранное уравнение |  |  |  |
| Модель 1 |  | 0.546443 | 0.30178 | 99.5654 |
| Модель 2 |  | 0.001403 | 0.01529 | 99.9989 |

Для Модели 1 (парной регрессии с одной объясняющей переменной) можно изобразить результат на графике в координатах (рис. 13).

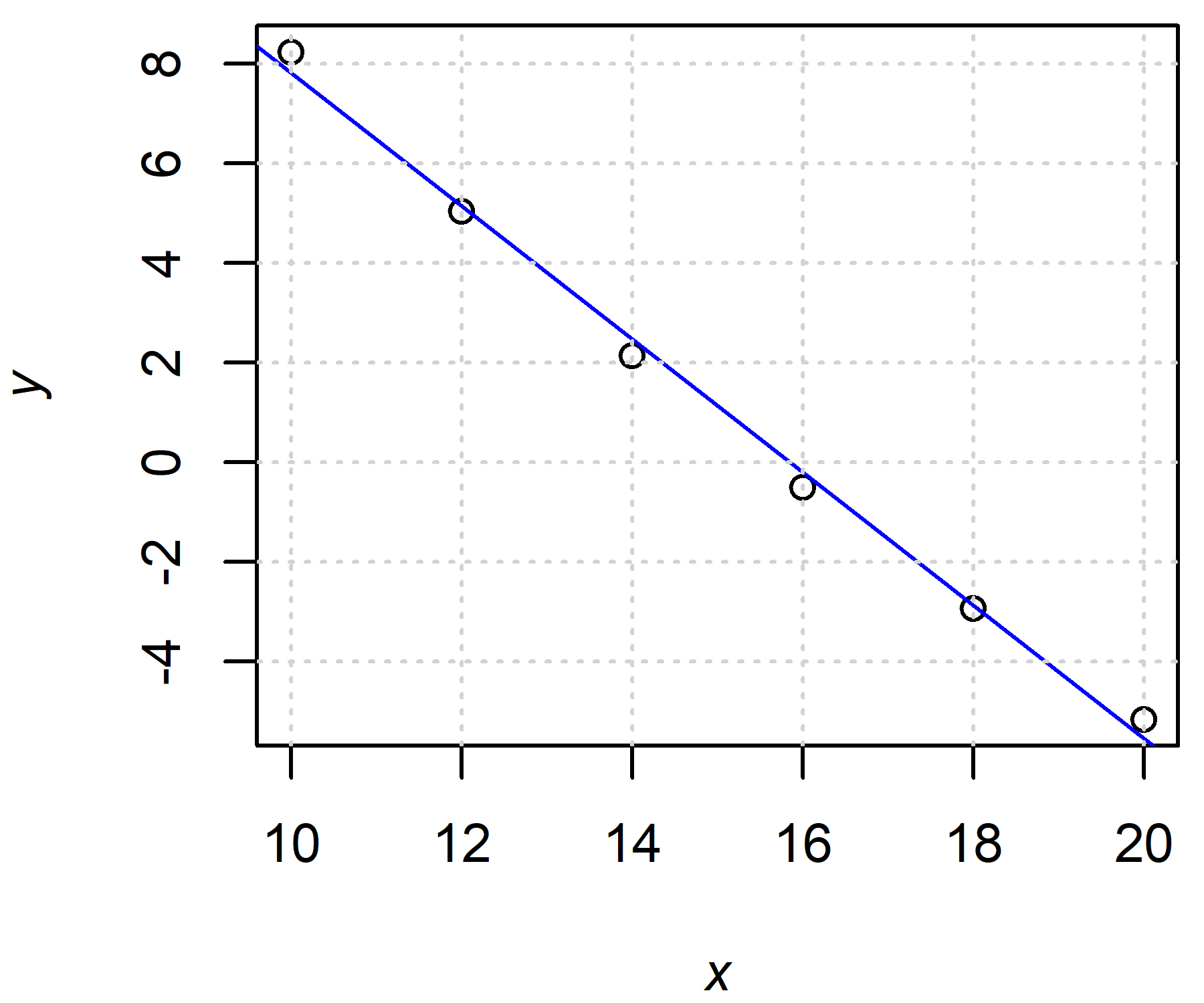


Рисунок 13. Линия регрессии для Модели 1

Для произвольной регрессии – парной или множественной – можно изобразить регрессию в виде точек и подобранной линии в координатах . Линия должна проходить через начало координат и иметь наклон 1.

|  |  |
| --- | --- |
|  |  |

Рисунок 14. Расчетные и фактические значения по Моделям 1 и 2

Чем ближе к наблюдениям удалось подобрать регрессию, тем ближе к линии лежат точки. Модель 2 здесь заметно точнее (рис. 14). Близость точек к линии формально измеряется показателем . Здесь, как мы видели, для Модели 2 более высокий.

## Центрирование регрессии. Дисперсионное тождество

### Центрирование регрессии

Пусть – оценки МНК, а – соответствующие остатки, так что  
Возьмем среднее от обеих частей с учетом свойства :  
где – вектор-строка средних. Значит, гиперплоскость регрессии проходит через точку средних .

Центрируем обе части уравнения регрессии :

Здесь – центрированная зависимая переменна, – центрированный *j*-й регрессор. Мы учли, что при центрировании вектора получаем , а остатки уже центрированы. Фактически, здесь мы из исходного уравнения регрессии вычитаем уравнение для средних, умноженное на вектор .

В матричном виде центрированное уравнение регрессии  
где , . Умножим слева на и учтем ортогональность :  
Получаем формулу оценок МНК для центрированной регрессии:

Видно, что это оценки МНК из регрессии на . Здесь  
выборочная ковариационная матрица объясняющих переменных (регрессоров),  
вектор выборочных ковариаций объясняющих переменных и зависимой переменной.

Константу можем найти из свойства :

### Дисперсионное тождество

Имеем (без доказательства), что расчетные значения в центрированной регрессии  
совпадут с центрированными расчетными значениями в исходной регрессии

Отсюда (с учетом ортогональности и ) получаем тождество для центрированных сумм квадратов:  
или

Здесь – общая сумма квадратов (англ. total sum of squares),  – объясненная сумма квадратов (англ. explained sum of squares).

Аналогично для дисперсий получаем **дисперсионное тождество**, показывающее разложение общей дисперсии объясняемой переменной на две части: объясненную и остаточную.

Здесь – полная дисперсия , – объясненная дисперсия (дисперсия, объясненная регрессией) или дисперсия расчетных значений, – остаточная дисперсия.

Объясненная дисперсия может быть представлена в различных вариантах:

Из дисперсионного тождества получается еще одна формула для коэффициента детерминации:

Это доля объясненной суммы квадратов в общей сумме квадратов или доля объясненной дисперсии в общей дисперсии.

### Коэффициент детерминации как квадрат коэффициента корреляции

Поскольку  
можем переписать R-квадрат как  
или  
Значит, – квадрат выборочного коэффициента корреляции фактических значений и расчетных. Отсюда следует, что .

## Геометрическая интерпретация МНК

Для тех, у кого хорошо развита геометрическая интуиция, полезной будет геометрическая интерпретация регрессии в пространстве наблюдений.

Будем рассматривать все переменные регрессии (, ) как *n*-мерные векторы. При использовании МНК мы подбираем расчетное значение зависимой переменной в виде линейной комбинации регрессоров () с коэффициентами :

Задача МНК состоит в том, чтобы подобрать коэффициенты таким образом, чтобы квадрат евклидова расстояния между вектором и вектором был наименьшим:  
или, что то же самое, чтобы евклидово расстояние между и было наименьшим:

С точки зрения линейной алгебры векторы при всех возможных образуют подпространство -мерного пространства. Вектор соответствует решению задачи МНК, когда он является проекцией вектора на подпространство . При этом вектор остатков будет ортогонален подпространству .

Дадим иллюстрацию этих свойств МНК в случае и (Рис. 15). На рисунке – вектор , — вектор (), – вектор . Векторы и определяют плоскость – это подпространство . Решая задачу МНК, мы ищем на плоскости точку , которая является ближайшей к точке . Такая точка будет проекцией точки на плоскость . Тогда – вектор расчетных значений , – вектор остатков . Вектор ортогонален плоскости .

Рисунок 15. Иллюстрация МНК в случае и

Можно подобную иллюстрацию дать для центрированных векторов и заодно проиллюстрировать коэффициент . В рассматриваемом случае и мы можем рассмотреть проекции всех рассматриваемых точек на плоскость, ортогональную (вектору ) и проходящую через точку . Получим точки , и – проекции точек , и . Проекцией точки будет . (Рис. 16)

Рисунок 16. Иллюстрация МНК в случае и , центрированные векторы

На этом рисунке – вектор , – вектор . Решая задачу МНК, мы ищем на прямой точку ближайшую к точке . Такая точка будет проекцией точки на прямую . Тогда – вектор центрированных расчетных значений , – вектор остатков . Векторы и ортогональны друг другу.

Косинус угла равен коэффициенту корреляции между фактическим и расчетным значением :

Соответственно, коэффициент детерминации – квадрат этого косинуса.

## Случай неединственности решения

Если , то решение задачи МНК не единственное. На самом деле решений в этом случае бесконечно много (континуум). Это означает, что имеющиеся наблюдения не позволяют получить оценки МНК однозначно. В этом случае даже если бы нам были точно (без ошибок ) известны значения величин , мы не могли бы по ним восстановить вектор .

Хотя при линейной зависимости столбцов матрицы оценки МНК не единственны, но условие первого порядка  
все равно выполнено, и расчетные значения и остатки однозначно определяются этим условием. Их можно вычислить следующим образом. Объединим максимальный линейно независимый набор регрессоров в матрицу . Тогда если – коэффициенты МНК из регрессии по , а – соответствующие расчетные значения, то расчетные значения в исходной регрессии тоже равны . Остатки равны , т. е. также вычисляются однозначно.

## Использование линейной регрессии и метода наименьших квадратов

Вообще говоря, метод наименьших квадратов – это только инструмент, некоторый способ получения аппроксимирующей формулы для набора точек. У этого инструмента существует довольно много различных применений.

*Во-первых*, можно использовать МНК *для подбора* *линейной аппроксимации* некоторой более сложной функции. Мы уже рассматривали подобный пример выше (см. пример с NPV). Зависимость может быть точной или содержать ошибки (например, ошибки округления). В рассмотренном примере отклонения точек от подобранной линии регрессии были связаны с тем, что мы приближали нелинейную функцию линейной.

*Во-вторых*, можно использовать линейную регрессию и МНК *для прогнозирования*. В этом случае в левой части регрессии стоит прогнозируемый показатель , а в правой – различные переменные , которые могут помогать в построении прогноза. Сначала оцениваются коэффициенты по имеющимся наблюдениям , а затем полученная оценка используется для построения прогноза при новых , когда значение еще неизвестно.

#### Пример: Прогнозирование объема торгов USDRUB\_TOM на Московской бирже

В роли выступает объем торгов по валютной паре доллар–рубль с поставкой «завтра» (т. е. на следующий рабочий день), измеренный в млрд долл. В роли выступает тот же объем в предыдущий торговый день. По данным за период с 2012-01-04 по 2020-12-30 построена регрессия

Фактические и расчетные значения по этой модели даны на рис. 17.

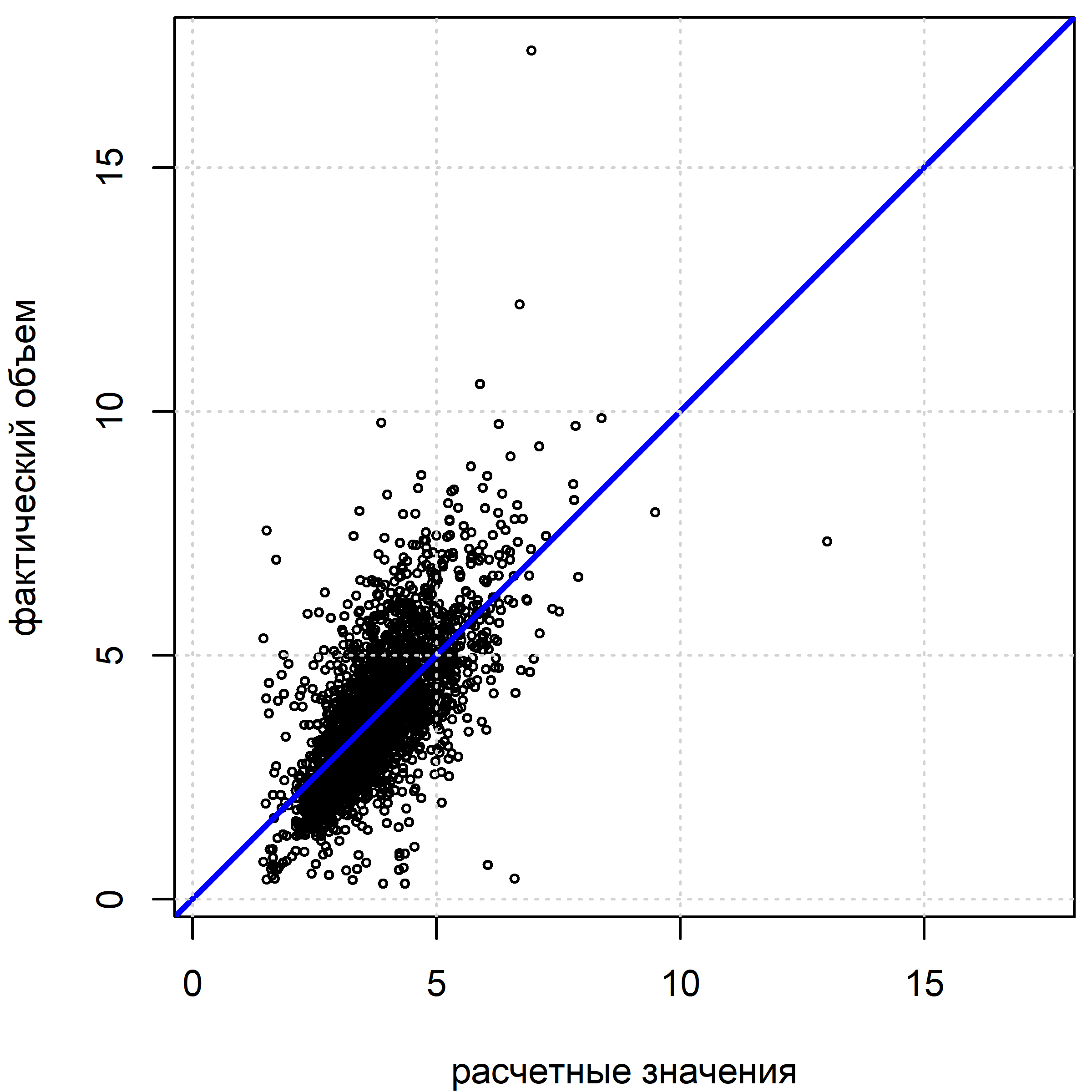


Рисунок 17. Фактические и расчетные значения по модели для объема торгов USDRUB\_TOM

Далее по той же формуле (без пересчета при появлении новых наблюдений) строятся прогнозы объемов на каждый торговый день в период с 2021-01-05 по 2021-09-24 с использованием информации об объемах в предыдущий период. На рис. 7 полученные значения сравниваются с фактически наблюдавшимися объемами.

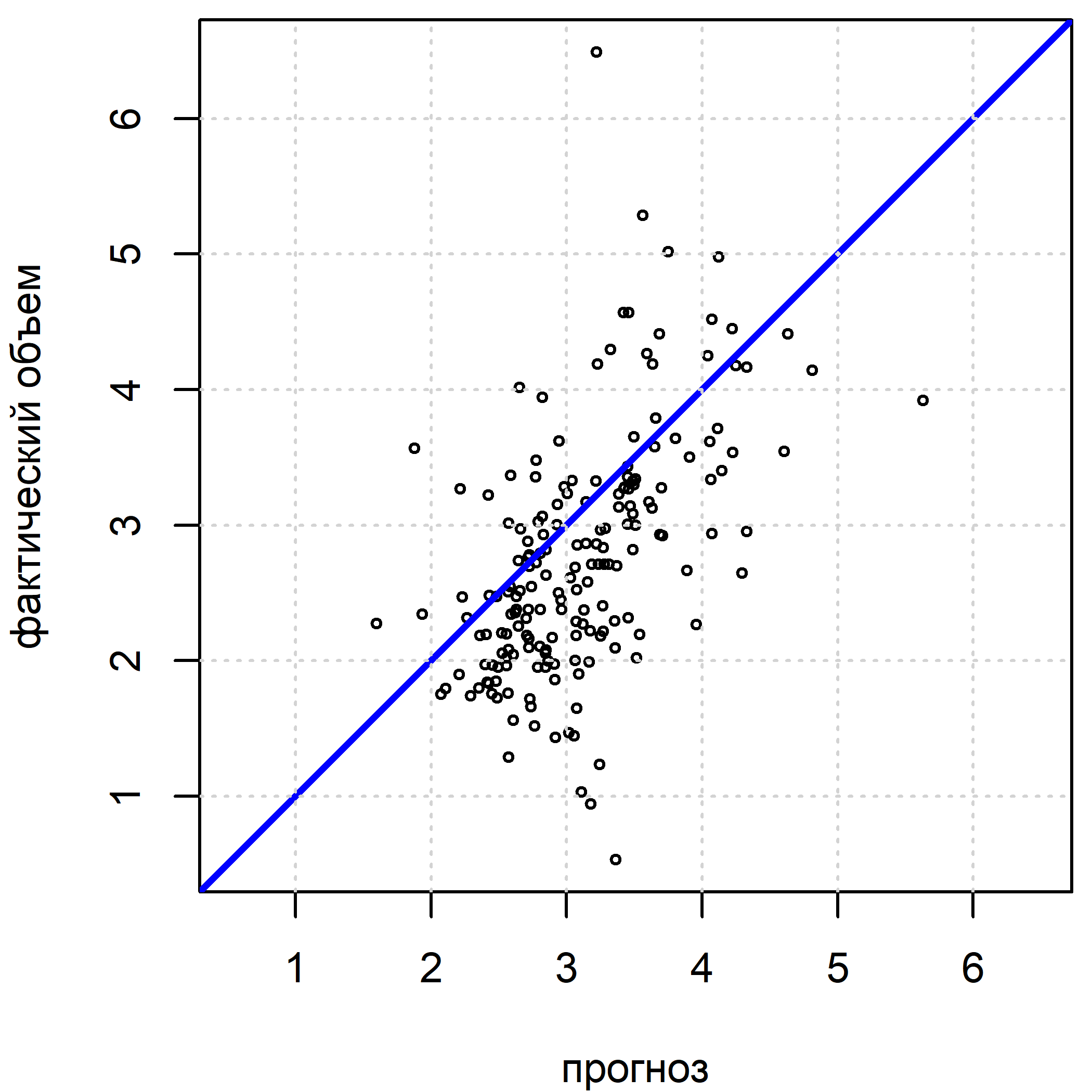


Рисунок 18. Прогнозы по модели для объема торгов USDRUB\_TOM и фактические объемы

*В-третьих*, линейная регрессия может *служить описанием для набора точек*.

#### Пример: Кривая Филлипса по США

Рассмотрим помесячные данные по США за период с марта 1997 г. по март 2020 г. по двум показателям:

wgt – показатель Wage Growth Tracker (% в год),

ue – уровень безработицы (% от общей величины рабочей силы).

Здесь Wage Growth Tracker – это средние темпы прироста заработной платы по участникам обследования населения, которые не меняли работу, вычисляемые Федеральным резервным банком Атланты. На рис. 19 дана диаграмма для двух показателей. Точки соединены линиями, чтобы показать совместную траекторию движения двух временных рядов. Видно, что в среднем, чем выше темпы прироста заработной платы, тем ниже безработица. Наблюдаемое явление представляет собой разновидность так называемой кривой Филлипса.

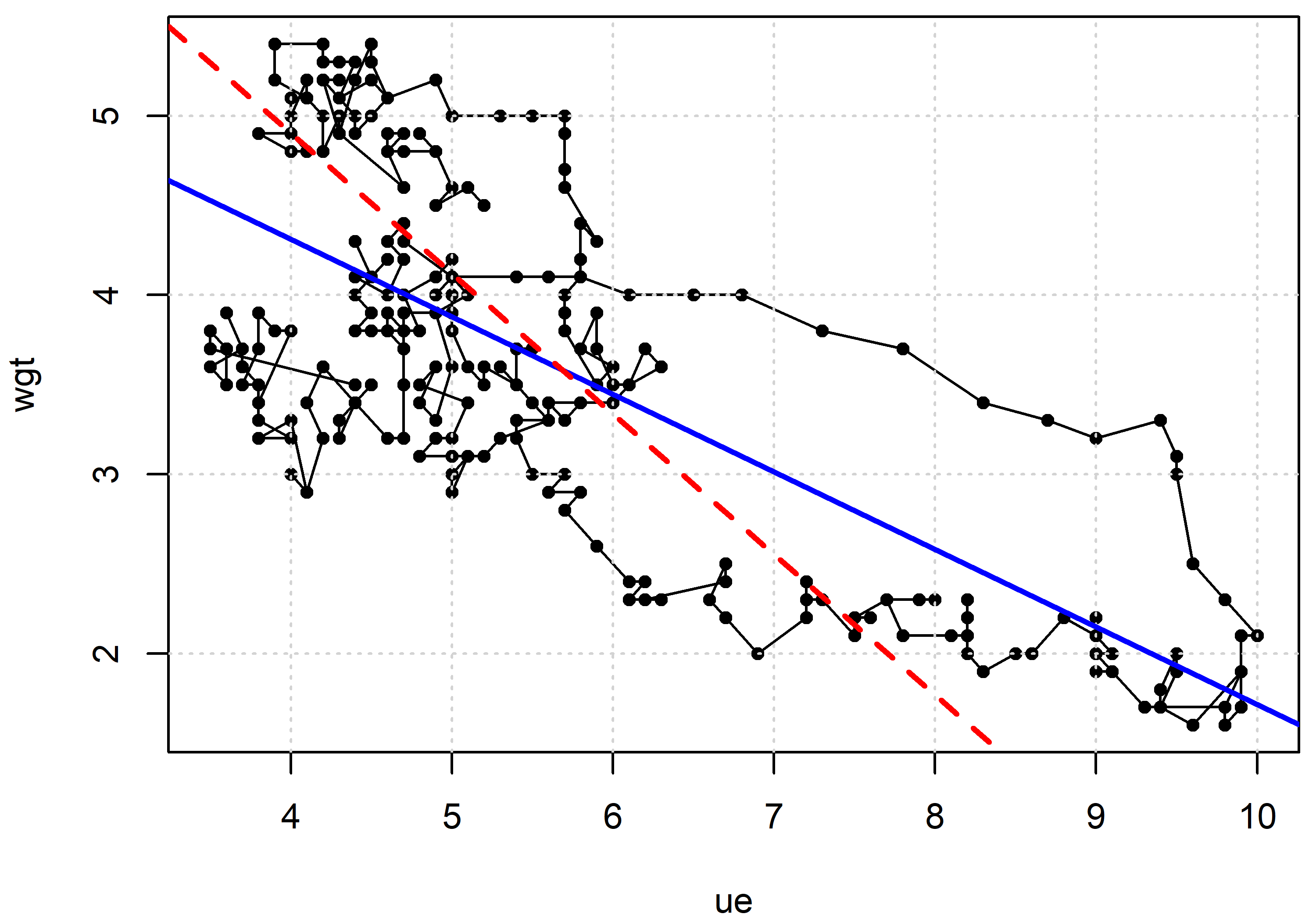


Рисунок 19. Безработица ue и темпы прироста заработной платы wgt и в США

По этим данным мы можем оценить линейную регрессию от :  
(сплошная прямая линия на рис. 19). Правда, в рассматриваемом случае мы не можем однозначно утверждать, что переменная правой части однонаправленно влияет на переменную левой части. Все что мы знаем, это то, что в данный период две переменные были отрицательно коррелированы. Мы могли бы поменять переменные местами и получить следующую обратную регрессию:  
Если перевернуть это уравнение, то получится  
(пунктирная прямая линия на рис. 19).

Согласно одному уравнению, уменьшение уровня безработицы на 1 процентный пункт сопровождалось увеличением темпов прироста заработной платы на процентных пункта. Согласно другому уравнению – на процентных пункта. Таким образом, мы получили два сравнительно грубых приближенных описания для наблюдаемого множества точек. Если бы был выше, то облако точек было бы больше похоже на прямую линию. Тогда линии, полученные по прямой и обратной регрессии, были бы довольно близки и обеспечивали бы более точное описание точек.

*В-четвертых*, можно использовать линейную регрессию *для оценивания причинно-следственных связей* между разными показателями. Это наиболее проблемное применение регрессии, но одновременно и самое интересное с точки зрения научного анализа. Основная проблема состоит в том, что для того, чтобы регрессию можно было интерпретировать в причинном смысле, требуется выполнения достаточно жестких и труднопроверяемых предположений.

Наиболее бесспорным применение регрессии для выявления причинно-следственных связей является ситуация контролируемого эксперимента, когда исследователь сам выбирает значение объясняющих переменных , причем делает это так, чтобы не исказить зависимость. В экономике такая ситуация возникает довольно редко.

#### Пример: Зависимость между частотой питателя и скоростью подачи руды

Здесь мы рассмотрим пример производственного эксперимента, который носит не столько экономический, сколько технологический характер. На некотором производстве скорость подачи руды по конвейерной ленте (, в т/c) зависит от частоты электродвигателя питателя (, в Гц). Были проведены эксперименты по установке частоты на нескольких разных уровнях. При этом измерялась соответствующая скорость подачи. Замеры производились с интервалом 30 с и было получено 349 наблюдений за парами . Методом наименьших квадратов по этим данным получена следующая линейная зависимость:

На рис. 20 показаны соответствующие точки и линия регрессии.



Рисунок 20. Пример регрессии по результатам производственного эксперимента

Данная регрессия позволяет достаточно точно предсказывать по . Пользуясь полученным уравнением можно с помощью изменения добиваться нужного уровня . В том, что причинно влияет на , не возникает сомнений, поскольку мы по своему усмотрению выбираем уровень . Значения выбирались по заранее составленному плану, и на них, в частности, не влияли ранее наблюдавшиеся значения , что позволило избежать смещения при оценке зависимости.

*В-пятых*, алгоритм МНК можно *использовать чисто технически* для расчета величины а также соответствующих расчетных значений , остатков , коэффициента и т. д. Если имеется компьютерная программа для расчета линейной регрессии, то можно взять какой-то вектор , какие-то регрессоры и программа рассчитает для них и другие сопутствующие величины. Такая возможность достаточно часто используется в эконометрике для расчета тестовых статистик и других вычислительных задач.[[2]](#footnote-2)

#### Пример: Использование регрессии для расчета коэффициента бета

Один из финансовых показателей для отдельного актива () – это коэффициент бета:  
где – доходность рассматриваемого актива, а – это так называемая рыночная доходность (доходность рыночного портфеля). Данный коэффициент отражает, как изменения в доходности рассматриваемого актива связаны с изменениями в доходности рынка в целом, и измеряет системный (недиверсифицируемый) риск в данном активе по отношению к риску по всему финансовому рынку. Коэффициента бета – это важный показатель в известной модели CAPM.

На практике один из упрощенных методов оценки коэффициента бета состоит в том, чтобы взять два ряда доходностей и ( за некоторый период времени и использовать вместо теоретических моментов соответствующие выборочные статистики:  
Как несложно понять, данная величина совпадает с оценкой коэффициента наклона в парной регрессии ряда на ряд . Таким образом, коэффициент бета можно вычислять с помощью обычной регрессии.

Рассмотрим пример вычисления коэффициента бета для акций Лукойла (LKOH). В качестве рыночного портфеля возьмем Индекс МосБиржи (IMOEX). Были взяты дневные (т. е. по торговым дням) ряды значений на момент закрытия за период с 2001-10-16 по 2021-10-15. На рис. 21 показаны исходные ряды. Исходные ряды были преобразованы в дневные темпы прироста (в процентах).

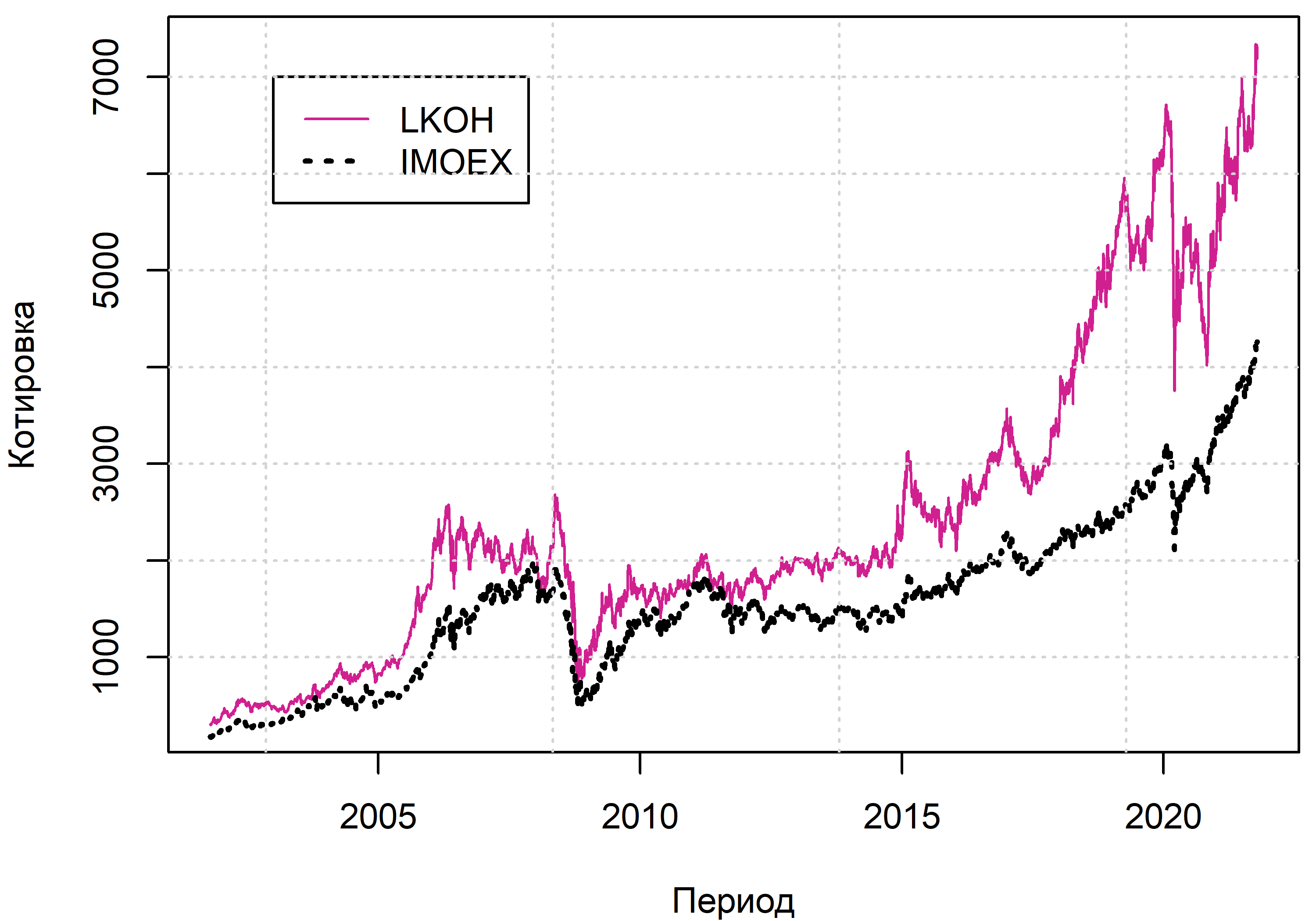


Рисунок 21. Ряды котировок акций Лукойла и Индекса МосБиржи

Для всего периода в целом получены следующие оценки регрессии:

Таким образом, оценка беты равна . Поскольку коэффициент бета со временем может меняться, может быть более правильно использовать более короткие ряды. На рис. 22 показаны оценки беты из регрессий, построенных по данным за отдельные годы.

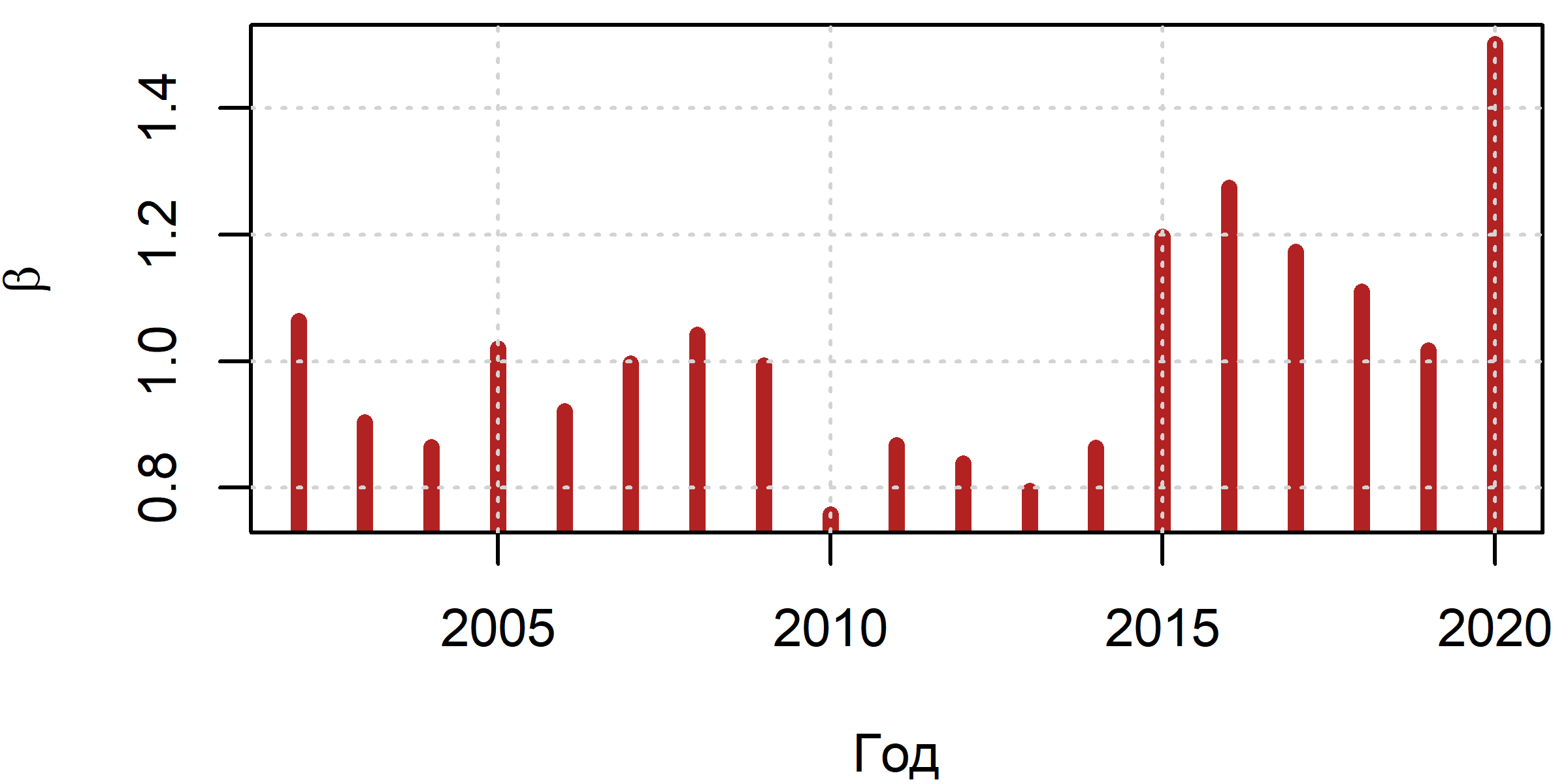


Рисунок 22. Оценки беты акций Лукойла по отдельным годам

## Контрольные вопросы:

1. Что здесь экзогенно, а что эндогенно: ?
2. Доказать, что матрица ковариации факторов является симметрической и положительно полуопределенной. В каком случае она положительно определена?
3. Сколько параметров оценивается в основной модели: ?
4. Что и как позволяет найти метод наименьших квадратов?
5. Какие свойства МНК-оценок обнаруживаются после минимизации остаточной дисперсии по коэффициентам регрессии?
6. Почему ?
7. Средняя МНК-оценок ошибок равна нулю. Что еще известно об этих оценках?
8. Доказать, что .
9. Доказать, что матрица вторых производных (матрица Гессе) для положительно полуопределена?
10. Доказать, что .
11. В чем заключается основное дисперсионное тождество регрессионного анализа?
12. Доказать тождество для нецентрированных сумм квадратов:
13. Доказать, что (расчетные значения из центрированной регрессии совпадают с центрированными расчетными значениями).
14. Как выражается объясненная дисперсия через матрицу ковариации факторов?
15. Доказать, что если , то .
16. Доказать, что если , то и .
17. Вывести МНК оценку коэффициента регрессии с одной константой ().
18. Найти величины ESS, RSS, TSS и длярегрессии
19. Вывести МНК оценку коэффициентов простой парной регрессии ().
20. Найти величины ESS, RSS, TSS и для регрессии

## Экзаменационные вопросы

1. Алгебра множественной регрессии: Оценка МНК параметров регрессии, смысл метода МНК. Условия первого порядка оценок МНК и их следствия.
2. Алгебра множественной регрессии: Расчетные значения, остатки. Объясненная и остаточная дисперсии, дисперсионное тождество. Коэффициент детерминации, в т. ч. его геометрическая интерпретация, связь с парным коэффициентом корреляции.
3. Описать МНК для парной регрессии: выписать целевую функцию, систему нормальных уравнений, оценки коэффициентов. Геометрическая интерпретация.

## Литература

1. Вербик М. Путеводитель по современной эконометрике. – М.: Научная книга,. 2008. [стр. 30-39]
2. Магнус Я. Р., Катышев П. К., Пересецкий А. А. Эконометрика. Начальный курс:– Учебник. – 6-е изд., перераб. и доп. – М.: Дело, 2004. [стр. 69-70, 74-76]
3. Суслов В. И., Ибрагимов Н. М., Талышева Л. П., Цыплаков А. А. *Эконометрия*. – Новосибирск: СО РАН, 2005. [*стр. 199-204*]
4. [Картаев Ф. С.](https://litgid.com/book_author/18780834/) *Введение в эконометрику. Учебник :*– Москва: Проспект, 2019. [*стр. 25-36,69-72]*

# Лекция: – Классическая модель линейной регрессии

## Классическая модель линейной регрессии

Ранее мы просто рассчитывали регрессию методом МНК по имеющимся данным, не задумываясь о том, какую природу они имеют. Чтобы решать более сложные задачи, желательно иметь некоторую модель, которая бы описывала наши данные. В такую модель должны входить случайные числа, то есть модель должна быть вероятностной.

Мы запишем уравнения для модели регрессии в следующем виде

или

где и – наблюдаемые переменные, а — ненаблюдаемая переменная, которая называется **ошибкой** регрессии (или **возмущением** – это слагаемое как бы «возмущает» уравнение, из-за чего взаимосвязь между и не является жесткой линейной взаимосвязью).

В матричном виде

Как и раньше при рассмотрении алгебры МНК здесь и – векторы-столбцы ,  – матрица , в которой первый столбец состоит из единиц,  – вектор-столбец **параметров** (коэффициентов регрессии).

Для чего может быть нужна модель? Вот одна из точек зрения:

«Обычно экономисты хотят больше, чем просто получение наилучшей линейной аппроксимации одной переменной по заданному множеству других переменных Им хочется получить экономические соотношения, в общем являющиеся более адекватными, чем выборка, которую они иногда имеют. Экономисты хотят извлечь выводы о том, что случится, если фактически одна из переменных изменится. То есть: они хотят сказать кое-что о вещах, которые не наблюдаются (еще). В этом случае мы хотим, чтобы соотношение, которое найдено, было бы более чем просто случайное историческое стечение обстоятельств; оно должно отражать фундаментальные отношения. Чтобы прийти к этому, предполагается существование общего соотношения, которое справедливо для всех возможных наблюдений…» [Вербик, с. 39:].

Сам М. Вербик в своем учебнике рассматривает данные как выборку из некоторой генеральной совокупности. Здесь же мы пойдем более простым путем и будем рассматривать наблюдаемые данные , и ненаблюдаемую ошибку как набор случайных чисел, который характеризуется некоторым точно не известным нам многомерным распределением, и постулируем ряд предположений относительно этих случайных чисел.

(**A0**) Функциональная форма. Уравнение регрессии корректно специфицировано и имеет вид

или в матричной записи

Предположение (A0) описывает, как именно связаны между собой те переменные, которые входят в модель. В частности, здесь уравнение регрессии одинаково для всех наблюдений , имеет *линейную форму* (как по переменным , так и по параметрам ), а ошибка входит в уравнение *аддитивно* (то есть прибавляется к основной функции ).

Поскольку ошибку мы не наблюдаем, то, согласно (A0), ошибка просто по определению равна при некоторых неизвестных нам коэффициентах . Без каких-то дополнительных предположений пользы от записанного уравнения регрессии мало.

Вербик, с. 41: «Важно понять, что без дополнительных ограничений статистическая модель не имеет смысла: для любого значения вектора коэффициентов регрессии всегда можно определить множество регрессионных остатков такое, что модель в точности будет справедлива для каждого наблюдения. Таким образом, мы должны принять некоторые предположения, чтобы придать модели смысл».

(**A1**) **Ошибка в среднем равна нулю**.

или в матричной записи

Это предположение означает, что ошибки не носят систематического характера. Ошибки могут быть иногда положительными, иногда отрицательными, но не должно быть систематического смещения в отрицательную или положительную сторону.

(**A2**) **Экзогенность регрессоров**: переменные являются экзогенными для модели. Пока что экзогенность берем в следующем виде:

*Регрессоры являются детерминированными*.

«Детерминированные» в данном случае означает, что мы с регрессорами будем обращаться, как будто они не являются случайными. Мы делаем это для упрощения теоретических конструкций. Позже мы рассмотрим более общую ситуацию и переформулируем предположение об экзогенности.

(**A3**) **Отсутствие гетероскедастичности** (heteroskedasticity) ошибок (**гомо­ске­да­стич­ность** – homoskedasticity). Ошибки всех наблюдений имеют одинаковые дисперсии :

Если ошибка для одних наблюдений гораздо более широко разбросана вокруг нуля, чем для других наблюдений, то сложно работать с такими наблюдениями в рамках одной модели. Здесь мы делаем упрощающее предположение, что дисперсии всех ошибок совпадают. При этом предположении отличие случайной ошибки от нуля, измеряемое ее квадратом, будет иногда больше, иногда меньше, однако не будет такого, что для одних наблюдений это отличие систематически больше, чем для других.

(**A4**) **Отсутствие автокорреляции ошибок**. Ошибки разных наблюдений некоррелированы между собой:

Это условие предполагает отсутствие систематической линейной связи между значениями ошибок в любых двух наблюдениях. Это предположение, в частности, будет выполнено в случае, когда разные наблюдения в совокупности независимы. (Но в принципе независимости наблюдений в этом предположении мы не требуем.)

Предположения (A3), (A4) можно записать одной матричной формулой:

Здесь – ковариационная матрица ошибок , а – единичная матрица .

(**A5**) **Нормальность ошибок**. Вектор ошибок имеет многомерное нормальное распределение. Если учесть предыдущие предположения, то в матричном виде можем записать

Данное предположение можно выразить и по-другому: ошибки регрессии являются независимыми, одинаково нормально распределенными случайными величинами с нулевыми математическими ожиданиями и дисперсиями :

Предположения (A0)–(A5) описывают **классическую модель линейной регрессии** (КМЛР). При условии, что детерминированные, можем КМЛР записать одной формулой:

Правда, в этой записи ошибки «пропали». Их надо дополнительно определить как .

Из предположений классической модели следуют различные теоретические результаты, которые мы в дальнейшем обсудим. При этом не обязательно требовать выполнения сразу всех предположений для того или иного результата. В частности, для многих результатов не нужно предположение (A5).

## Оценки и идентификация

При использовании модели линейной регрессии, как правило, стоит задача получить некоторую оценку для истинных коэффициентов . Сами истинные коэффициенты в модели предполагаются неизвестными, но требуется получить такие , которые бы достаточно хорошо аппроксимировали . Самая широко используемая оценка – это оценка наименьших квадратов (оценка обычного МНК):

Мы здесь говорим об *оценке*, как принято в русскоязычной литературе, хотя более корректно говорить здесь о *функции оценки* или оценивателе (англ. *estimator*). Это некоторая функция , которая по наблюдаемым данным и выдает оценку . Обычно стремятся, чтобы функция оценки обладала «хорошими» свойствами и давала хорошие приближения для .

Для вычисления оценок МНК требуется обратить матрицу . В связи с этим следует ввести дополнительное предположение о невырожденности матрицы .

(**I**) Отсутствие полной мультиколлинеарности регрессоров.

*Cтолбцы матрицы (т. е. регрессоры) являются линейно независимыми.*

Другими словами, мы будем предполагать, что ранг матрицы равен .

Данное предположение можно назвать предположением об **идентифицируемости**. Если вырожденная, то возникает проблема с однозначной идентификацией коэффициентов . Это не только проблема получения оценок МНК, это принципиальная проблема получения однозначных оценок. Если столбцы матрицы линейно зависимы, то кроме найдутся и другие векторы коэффициентов, для которых выполнено аналогичное уравнение:  
В результате мы не можем по имеющимся данным отличить модель от .

(Действительно, если столбцы матрицы линейно зависимы, то для некоторого вектора имеем . Поэтому взяв получим для .)

В дальнейшем, если не оговаривается противное, мы будем всюду предполагать, что условие (I) выполнено.

## Несмещенность оценок МНК

Одно из «хороших» свойств оценок МНК, из-за которых они получили широкое распространение – это несмещенность. Докажем ее. Для этого сначала выразим оценку МНК через ошибки (это представление пригодится нам и в дальнейшем):

Отсюда  
Мы воспользовались здесь предположениями (A0), когда подставили , (A2) когда вынесли функцию от из-под математического ожидания, и (A1).

Таким образом, оценка МНК в среднем равна .

## Предположения Гаусса–Маркова. Свойство НЛНО

Если дополнительно воспользоваться предположениями (A3) и (A4), то можно получить другие важные свойства оценок МНК. Предположения (A0)–(A4) принято называть **предположениями Гаусса–Маркова**:

Мы знаем, что  
Отсюда вычислим ковариационную матрицу оценок МНК:  
Упростив последнюю формулу, получим

Ту же формулу можно вывести немного по-другому:

Здесь при выводе мы использовали свойства ковариационной матрицы. Если – некоторый случайный вектор, – постоянный вектор, а – постоянная матрица подходящих размерностей, то

Ковариационная матрица показывает точность оценок. Чем она меньше, тем точнее оценка. В предельном случае, когда (т. е. ошибки в регрессии нулевые), матрица будет равна нулю и с вероятностью единица, т. е. мы абсолютно точно узнаём .

Можно ли получить более точную оценку чем оценку МНК? Нет, если *ограничиться классом линейных по несмещенных оценок.*

Линейная по оценка будет иметь вид , где матрица размерности получается как некоторая функция от . Оценка МНК является линейной, поскольку она соответствует . Как мы уже видели, оценка МНК является несмещенной.

При справедливости предположений (A0)–(A4) верна **Теорема Гаусса–Маркова**: оценка МНК является **наилучшей линейной несмещенной оценкой** (НЛНО, англ. *BLUE* – *best linear unbiased estimator*).

Действительно, пусть – некоторая другая несмещенная оценка. Выразим эту оценку через ошибки:

Ее математическое ожидание равно

Чтобы оценка была несмещенной при любом векторе истинных коэффициентов , требуется, чтобы . Поэтому

Ковариационная матрица оценки равна

Так как , разность двух оценок равна

Ковариационная матрица этой разности  
Раскрывая скобки и упрощая с учетом требования получим  
или  
Так как это означает, что

Значит, ковариационные матрицы оценок и отличаются на величину положительно полуопределенной матрицы . Таким образом, в матричном смысле не меньше, чем , то есть альтернативная линейная несмещенная оценка не может быть более точной, чем оценка МНК.

Можно переформулировать теорему Гаусса–Маркова без использования сравнения ковариационных матриц. Пусть мы интересуемся некоторой линейной комбинацией коэффициентов , заданной в виде , где  – произвольный -мерный вектор. Тогда, согласно данной теореме,  
то есть не может быть более точной оценкой , чем .

Есть и другие следствия из предположений Гаусса–Маркова.

Обычно в модели регрессии не только коэффициенты , но и дисперсия ошибки неизвестна. Остатки МНК напоминают ненаблюдаемые случайные ошибки . Поэтому естественно в качестве оценки для взять остаточную дисперсию

Найдем математическое ожидание этой оценки. Остатки выразим через ошибки:  
или  
где мы обозначили

Матрица является симметричной идемпотентной, то есть  и . Для остатков выполнено , а их ковариационная матрица равна

Математическое ожидание суммы квадратов остатков равно

Дисперсии стоят по диагонали ковариационной матрицы остатков , а их сумма – это след данной матрицы. Поэтому

(Мы использовали свойства следа матрицы: множитель можно выносить, след суммы – это сумма следов, а матрицы внутри можно циклически переставлять.) Следовательно,

Как видим, эта оценка является смещенной для , но смещение убывает до нуля с ростом количества наблюдений .

Для корректировки смещения можем умножить смещенную остаточную дисперсию на . Получим следующую **несмещенную остаточную дисперсию**:

Величину принято называть **стандартной ошибкой регрессии**.

Отсюда же получим несмещенную **оценку для ковариационной матрицы оценки МНК** , которая пригодится нам в дальнейшем:

Дисперсия оценки отдельного коэффициента равна

где через мы обозначили диагональный элемент матрицы . Соответственно, оценка этой дисперсии равна

Эти дисперсии стоят по диагонали ковариационной матрицы . Корень из дисперсии

называют **стандартной ошибкой оценки** . Она (приближенно) показывает, насколько точной оценкой истинного коэффициента является . В результате оценивания регрессии мы можем получить не только оценки коэффициентов , но и соответствующие стандартные ошибки , которые дают важную информацию о точности как оценки .

Корректность стандартных ошибок зависит от истинности предположений Гаусса–Маркова. Если предположения нарушены, то оценка и стандартные ошибки будут вводить нас в заблуждение относительно точности оценки МНК .

## Контрольные вопросы

1. Чем различаются переменные левой и правой части уравнения регрессии?
2. Какие гипотезы (предположении) обеспечивают несмещенность МНК-оценок?
3. При доказательстве каких фактов «работают» гипотезы**A3** и**A4**?
4. Что доказывается в теореме Гаусса-Маркова?
5. Какие гипотезы обеспечивают взаимную независимость ошибок в наблюдениях?
6. Предположения Гаусса–Маркова.
   1. Что они означают?
   2. Вспомните формулу для МНК-оценок.
   3. Являются ли МНК-оценки линейными по y?
   4. Докажите, что МНК-оценки являются несмещёнными.
   5. Найдите

Пусть – некоторые другие несмещенные оценки.

* 1. Являются ли МНК-оценки линейными по y?
  2. Чему равняется матрица
  3. Найдите и .
  4. На какую ковариационную матрицу похоже выражении, полученное в предыдущем пункте?
  5. Докажите, что
  6. Рассмотрим матрицу . Что находится на её диагонали?
  7. Является ли матрица симметричной?
  8. Докажите, что матрица C является положительно полуопределённой.

1. Покажите, что , где – симметрическая, идемпотентная и положительно полуопределенная матрица.
2. Докажите, что
3. Какие гипотезы обеспечивают, что является несмещенной оценкой для ?
4. Покажите, что является несмещенной оценкой дисперсии ошибок .

## Экзаменационные вопросы

1. Классическая модель линейной регрессии. Несмещенность оценок МНК.
2. Классическая модель линейной регрессии. Оценка ковариационной матрицы оценок МНК.
3. Классическая модель линейной регрессии. Свойство BLUE оценок МНК (Теорема Гаусса–Маркова)
4. Классическая модель линейной регрессии. Оценка остаточной дисперсии (смещенная, несмещенная).

## Литература

1. Вербик М. Путеводитель по современной эконометрике. – М.: Научная книга,. 2008. [стр. 39-49]
2. Магнус Я. Р., Катышев П. К., Пересецкий А. А. Эконометрика. Начальный курс:– Учебник. – 6-е изд., перераб. и доп. – М.: Дело, 2004. [стр. 70-74, 77-78]
3. Суслов В. И., Ибрагимов Н. М., Талышева Л. П., Цыплаков А. А. *Эконометрия*. – Новосибирск: СО РАН, 2005. [*стр. 226-230*]
4. [Картаев Ф. С.](https://litgid.com/book_author/18780834/) *Введение в эконометрику. Учебник :*– Москва: Проспект, 2019. [*стр. 36-48,73-75]*

# Лекция: – Следствия нормальности. Тестирование гипотез для коэффициентов

## Следствия нормальности

Добавим теперь к предыдущим предположениям (предположениям Гаусса–Маркова) предположение (A5) о нормальности остатков:

Поскольку линейно связана с ошибками:  
отсюда следует нормальность оценки МНК:  
То есть имеет многомерное нормальное распределение с математическим ожиданием и ковариационной матрицей .

Оценка отдельного коэффициента имеет одномерное нормальное распределение:  
где через мы обозначили диагональный элемент матрицы .

Также мы видели, что остатки равны , где , причем , а ковариационная матрица остатков равна . Из этого следует, что

Остатки и оценка коэффициентов не коррелируют между собой:  
Здесь мы использовали, что . Составной вектор будет иметь многомерное нормальное распределение, а для такого распределения из некоррелированности следует независимость. Поэтому и независимы между собой.

## Тестирование гипотезы для одного коэффициенты

Указанные свойства можно использовать для статистической проверки гипотез относительно коэффициентов . В частности, рассмотрим следующую величину, имеющую стандартное нормальное распределение:  
Можно было бы с помощью построенной на этом свойстве статистики проверять гипотезы относительно . Но дисперсия нам обычно неизвестна. Вместо нее можно использовать оценку – остаточную дисперсию .

Заметим, что  
где матрица – симметричная идемпотентная. По свойствам многомерного нормального распределения из этого следует, что  
где – это след матрицы .

При этом величины  
независимы между собой как функции от независимых и . Из них можно получить величину, имеющую распределение Стьюдента с степенями свободы:

Таким образом,  
где – стандартная ошибка . Как мы ранее видели, стандартная ошибка получается как корень из диагонального элемента ковариационной матрицы

Предположим, что требуется проверить следующую нулевую гипотезу:  
где – это гипотетическое значение коэффициента . Для этого можно использовать следующую тестовую статистику (**статистику Стьюдента** или ***t*-статистику**):  
При сделанных нами предположениях в условиях истинности нулевой гипотезы эта статистика будет иметь распределение Стьюдента с степенями свободы.

Нулевая гипотеза отклоняется, если оценка сильно отличается от гипотетического значения . В качестве альтернативной гипотезы, как правило, используется  
т. е. не исключается, что может быть как , так и . В таком случае нулевая гипотеза отклоняется, если сильно отличается от как в меньшую, так и в большую сторону. Другими словами, нулевая гипотеза отклоняется, если *t*-статистика сильно отличается от нуля в отрицательную или в положительную сторону. С учетом симметричности распределения Стьюдента поступают следующим образом: нулевую гипотезу отклоняют, если *t*-статистика по модулю больше . При этом, ***t*-тест**, построенный на статистике Стьюдента, осуществляется следующим образом:

|  |
| --- |
| Если , то принимаем и считаем, что . |
| Если , то отклоняем и считаем, что . |

Здесь – это **критическое значение**, получаемое как двусторонняя квантиль распределения Стьюдента со степенями свободы , соответствующая вероятности левого и правого хвостов , где – некоторая маленькая вероятность, которую называют **уровнем значимости** теста.

Другой способ проверки гипотезы состоит том, что рассчитывается так называемое ***p*-значение** (англ. *p-value*). Так же как и выше, рассматривается теоретическое распределение статистики при истинности нулевой гипотезы. *P*-значение – это величина вероятности в соответствующих хвостах распределения, заданных значением статистики. Другими словами, это *граничный уровень значимости*, то есть минимальный уровень значимости, при котором нулевая гипотеза отклоняется. Для *t*-теста это вероятность левого и правого хвоста распределения , то есть вероятность, что случайная величина, распределенная как , по модулю будет больше . Пусть – это *p*-значение статистики . Проверка гипотезы заключается в сравнении с выбранным уровнем значимости :

|  |
| --- |
| Если , то принимаем и считаем, что . |
| Если , то отклоняем и считаем, что . |

Удобство использования *p*-значения состоит в том, что при любом уровне значимости при проверке нулевой гипотезы требуется знать только . Тут не требуется искать квантили, соответствующие разным значениям .

Еще один, третий по счету, способ связан с **интервальным оцениванием** параметров регрессии, когда вместо точечной оценки дается оценка в виде отрезка. Вокруг оцененного параметра строится **доверительный интервал** для истинного значения параметра . Интервал строится таким образом, чтобы вероятность того, что он включает была равна заданной доверительной вероятности. Если – соответствующая двусторонняя квантиль распределения Стьюдента со степенями свободы , то интервал имеет вид

Для проверки нулевой гипотезы нужно просто посмотреть, лежит ли гипотетическое значение в доверительном интервале:

|  |
| --- |
| Если лежит в интервале, то принимаем . |
| Если не лежит в интервале, то отклоняем . |

Чтобы описанный тест обладал уровнем значимости , следует выбрать доверительную вероятность равной , а двустороннюю квантиль взять такую же, как в обычном *t*-тесте – правый и левый хвост соответствуют в сумме вероятности , а середина распределения – доверительной вероятности .

Преимущество использования доверительного интервала состоит в том, что он не привязан к конкретной нулевой гипотезе. Поэтому можно сначала рассчитать интервал, а гипотезу сформулировать и проверить позже. Кроме того, доверительный интервал сам по себе является интервальной оценкой для параметра. Его можно использовать неформально для изучения правдоподобности того, что неизвестный параметр принадлежит или не принадлежит некоторой важной с практической точки зрения области (см. далее обсуждение различия между статистической и практической значимостью). Неудобство доверительного интервала состоит в том, что он привязан к конкретной доверительной вероятности.

Следует обратить внимание на то, что при описанном классическом подходе к проверке гипотез контролируется вероятность так называемой **ошибки первого рода** – т. е. вероятность отклонить нулевую гипотезу при условии, что она верна. Эта вероятность берется достаточно малой – это и есть уровень значимости . Ошибку второго рода, то есть вероятность принять неверную нулевую гипотезу, контролировать сложнее. Дополнительную к вероятности ошибки второго рода величину называют **мощностью теста**. По крайней мере, тесты обычно конструируются так, чтобы при существенной разнице между и в правдоподобных условиях они имели достаточно большую мощность, и, соответственно, вероятность ошибки второго рода была маленькой. Рассмотренный *t*-тест сконструирован именно таким образом.

Если нулевая гипотеза не была отклонена, то исследователь может использовать полученную информацию при моделировании и изменить модель регрессии, наложив соответствующее ограничение. Если подставить ограничение в модель, что получается  
Поскольку коэффициент известен, и его не требуется оценивать, можно перенести соответствующее слагаемое в левую часть:

Получаем новую модель, в которой объясняемой переменной становится и один регрессор из правой части пропадает. В дальнейшем можно работать с этой новой моделью.

Гипотетическое значение коэффициента обычно определяется какими-то теоретическими соображениями. Очень часто берут , что соответствует случаю, когда зависимая переменная не связана с переменной , т. е. коэффициент при ней в регрессии равен нулю и она пропадает из уравнения:  
Статистика Стьюдента для соответствующей гипотезы  
по умолчанию выдается в статистических компьютерных программах в таблице результатов оценивания регрессии. Она равна оценке коэффициента, деленой на ее стандартную ошибку:

Если нулевая гипотеза верна, то, как и в общем случае, эта статистика будет иметь распределение Стьюдента с степенями свободы. Стандартный, *t*-тест, построенный на статистике Стьюдента, осуществляется следующим образом:

|  |
| --- |
| Если , то принимаем и считаем, что  (переменная не нужна в регрессии). |
| Если , то отклоняем и считаем, что  (переменная нужна в регрессии). |

Если гипотеза отклонена, то переменную принято называть **значимой**, а если принята, то **незначимой**.

Если мы хотим проверить гипотезу при ненулевом значении , но есть только компьютерная программа, которая считает *t*-статистику для , то можем положить и использовать замену переменных :

В этой регрессии зависимая переменная , а регрессоры те же, что и в исходной. Проверяя гипотезу, что , то есть что коэффициент при переменной равен нулю, тем самым проверяем гипотезу, что в исходной регрессии.

## Тестирование гипотез для нескольких коэффициентов

Таким же способом можем проверять и другие линейные ограничения на коэффициенты регрессии. Пусть, например, требуется проверить гипотезу  
в модели  
(Производственную функцию Кобба—Дугласа , где – выпуск, – затраты капитала, – затраты труда, можно оценивать с помощью такой линейной регрессии, если обозначить , , , . Ограничение для функции Кобба—Дугласа означает постоянную отдачу от масштаба.)

Можем обозначить и сделать подстановку :  
После преобразования получим

Это тоже линейная регрессия, но с зависимой переменной и регрессорами  
 и . По смыслу она эквивалентна исходной модели. Проверяя гипотезу, что в преобразованной регрессии, мы тем самым проверяем гипотезу, что в исходной регрессии. Если нулевая гипотеза будет принята, то можем далее учесть эту информацию и вместо исходной модели использовать модель  
в которой стало на один неизвестный параметр меньше.

В более общем случае пусть мы хотим в общей модели проверить гипотезу  
где – известный вектор-строка длиной , – известное число. Обозначим  
и выполним подстановку  
где – один из параметров, коэффициент при котором не равен нулю. В результате получим новую линейную регрессию, в которой зависимая переменная и регрессоры некоторым образом поменяются, и одним из регрессоров будет с коэффициентом . В этой новой регрессии мы можем использовать обычную *t*-статистику для переменной для проверки гипотезы, что , и тем самым проверить гипотезу, что в исходной регрессии. В предположении истинности нулевой гипотезы такая *t*-статистика будет иметь обычное распределение Стьюдента с степенями свободы.

Все же вручную делать подстановки для проверки линейного ограничения не всегда удобно. Поэтому в компьютерных программах, оценивающих регрессии, для проверки гипотез есть готовые процедуры.

Пусть требуется проверить то же ограничение  
В качестве основы статистики для проверки этой гипотезы логично взять величину  
Это оценка величины . Также – это **невязка** ограничения , показывающая, «насколько ограничение не выполняется для оценки ».

Мы знаем, что при нормальности  
По свойствам нормального распределения из этого следует, что  
где  
При выполнении нулевой гипотезы будет выполнено  
и  
Как и ранее, используем в статистике величину , для которой  
Случайные величины и независимы между собой и из них можно сформировать случайную величину, имеющую распределение Стьюдента.

Отсюда (выкладки здесь аналогичны предыдущим) получим следующую статистику Стьюдента:  
где – это стандартная ошибка невязки , т. е. квадратный корень из (несмещенной) оценки дисперсии , равной  
Данная статистика имеет стандартный вид так называемого ***t*-отношения**: оценка, деленая на стандартную ошибку этой оценки. При выполнении нулевой гипотезы статистика имеет распределение Стьюдента с степенями свободы.

Если требуется проверить сразу несколько ограничений, то уже не получится использовать статистики Стьюдента. Пусть есть линейных ограничений и они записаны в матричном виде:  
где – матрица , – вектор . Как и в случае одного ограничения по свойствам нормального распределения при выполнении имеем  
откуда следует, что

Иcпользуя связь между распределением Фишера и хи-квадрат, получим, что следующая *F*-статистика  
при истинности нулевой гипотезы будет иметь распределение Фишера со степенями свободы и .

Как и для *t*-статистики, можем здесь ввести величину невязки ограничения  
Это вектор длины . Тогда *F*-статистика равна

где – (несмещенная) оценка ковариационной матрицы . Она вычисляется через оценку ковариационной матрицы коэффициентов :  
Таким образом, имея приближенную ковариационную матрицу коэффициентов , мы можем проверять с помощью нее разные гипотезы о коэффициентах .

***F*-тест**, построенный на статистике Фишера, осуществляется следующим образом:

|  |
| --- |
| Если , то принимаем и считаем, что . |
| Если , то отклоняем и считаем, что . |

Здесь – это критическая граница, получаемая как односторонняя квантиль распределения Фишера со степенями свободы и , соответствующая вероятности правого хвоста , где – уровень значимости теста.

Другой способ проверки основан на *p*-значении , которое вычисляется в данном случае как вероятность правого хвоста распределения Фишера, соответствующая . Процедура здесь такая же, как и в случае распределения Стьюдента:

|  |
| --- |
| Если , то принимаем и считаем, что . |
| Если , то отклоняем и считаем, что . |

Особый случай *F*-теста – это так называемый **тест на удаление переменных**. Проверяется нулевая гипотеза, что коэффициенты при нескольких () регрессорах одновременно равны нулю. Если нулевая гипотеза принимается, то проверяемые регрессоры называют в совокупности **незначимыми** и их можно удалить из регрессии.

Очень важно понимать, что набор обычных *t*-статистик для группы регрессоров не может здесь заменить *F*-статистику. Дело в том, что *t*-тесты предназначены для проверки значимости *отдельных* переменных (гипотез ), а *F*-тест на удаление переменных тестирует *все переменные в совокупности*. Таким образом, если сразу несколько переменных в регрессии незначимы по *t*-статистикам, то не следует спешить с выводом, что все они незначимы в совокупности и поэтому их можно все одновременно удалить из регрессии. Если пользоваться *t*-статистиками, то следует удалять незначимые переменные только по одной. После удаления одной из незначимых переменных остальные могут стать значимыми по *t*-статистикам. Можно порекомендовать удалять из регрессии те незначимые переменные, которые имеют самые маленькие по модулю *t*-статистики (и самые большие *p*-значения).

Особый случай теста на удаление переменных – это ***F*-тест для модели** в целом. Нулевая гипотеза в таком тесте состоит в том, что коэффициенты *при всех регрессорах, кроме константы*, одновременно равны нулю:

Здесь ограничений. (Вопрос: какой вид имеют матрица и вектор в данном случае?) Если указанная гипотеза принимается, то *регрессия в целом незначима*. Тогда у нас нет достаточных оснований по имеющимся данным говорить о том, что объясняющие переменные помогают объяснить переменную .

Альтернативный способ расчета статистики Фишера основан на тождестве:  
где – сумма квадратов остатков в исходной регрессии («длинная» регрессия), а (*restricted RSS*) – сумма квадратов остатков в модели с наложенным ограничением («короткая» регрессия). Тождество доказано в Приложении к главе (и там же выведена одна из возможных формул для оценки МНК с ограничением ). Используя это тождество, запишем альтернативную формулу для *F*-статистики:  
Пользоваться данной формулой удобно, когда несложно оценить как одну модель, так и другую. Например, она удобна при проведении теста на удаление переменных.

В случае *F*-теста для модели в целом ограниченная модель – это модель с одной константой, для которой, как мы знаем, и . Поэтому для него  
(Мы учли, что количество ограничений .) Здесь видна связь с коэффициентом детерминации, который вычисляется на основе тех же сумм квадратов:  
А именно,  
Статистика при нулевой гипотезе о «ненужности» всех регрессоров, кроме константы, имеет распределение Фишера со степенями свободы и .

Учесть в регрессии ограничение общего вида несколько сложнее, чем ограничение . Если требуется учесть одно линейное ограничение , то можем сделать подстановку  
перенести слагаемые с известными коэффициентами в левую часть и переобозначить переменные по той же схеме, как это, например, было сделано выше для случая ограничения . Похожим способом можно учесть сразу несколько линейных ограничений на параметры регрессии. После подстановки линейных ограничений получим некоторую модель линейной регрессии, в которой вместо коэффициента будет на коэффициентов меньше, т. е. . Оценив модель, получим в ней сумму квадратов остатков , которую затем можно использовать в *F*-тесте для проверки наложенных ограничений.

Относительно описанных в этом разделе статистик следует сделать три важных замечания.

Во-первых, при выводе распределения статистик мы использовали предположение о нормальности ошибок регрессии. В реальных данных очень часто наблюдаются существенные отклонения от нормальности, но это не означает, что описанные статистики нельзя использовать. Просто распределение соответствующих *t*- и *F*-статистик только приближенно будет *t* и *F*. Для обоснования их использования приходится прибегнуть к асимптотической теории, что выходит за рамки данного пособия. К этому вопросу мы еще вернемся в соответствующей главе.

Во-вторых, в реальных данных часто наблюдаются гетероскедастичность и автокорреляция. При этом стандартная оценка для ковариационной матрицы  
перестает быть корректной и соответствующие тесты дают вводящие в заблуждение результаты. В то же время, мы по-прежнему можем использовать *F*-статистики, рассчитанные по формуле  
и *t*-статистики

где и , если заменим здесь оценку ковариационной матрицы на вариант, который состоятелен при гетероскедастичности и автокорреляции (так называемую робастную ковариационную матрицу, о ней пойдет речь в другой главе). Более того, можно порекомендовать использовать такую поправку на регулярной основе при проведении тестов, чтобы не прийти к ложным выводам.

Обоснование использования робастных ковариационных матриц для расчета *t*- и *F*-статистик требует асимптотической теории. Распределение соответствующих статистик только приближенно будет *t* и *F*.

В-третьих, значения статистик, рассчитанных по суммам квадратов остатков:

и по совпадают с теми, которые основаны на стандартной оценке ковариационной матрицы, поэтому их *не следует использовать, когда есть заметные гетероскедастичность или автокорреляция, которые искажают стандартную оценку ковариационной матрицы*.

## Статистическая значимость и практическая значимость

Мы рассмотрели, каким образом можно проверять гипотезы относительно коэффициентов регрессии. Если гипотеза о равенстве нулю коэффициентов отклоняется (*p*-значение маленькое), то говорят, что соответствующие регрессоры значимы. Здесь следует помнить, что речь идет о значимости только с точки зрения статистической проверки гипотез, но не о практической значимости. Поэтому во избежание путаницы лучше уточнять: «переменные являются статистически значимыми (на таком-то уровне)».

Довольно широко распространена ошибка, когда исследователи смотрят только на статистическую значимость или незначимость переменной (на данном уровне значимости) и не обращают внимания на величину самого коэффициента. Если переменная оказывается значимой (*p*-значение маленькое), то и измеряемый эффект считают достаточно сильным и практически важным, а если незначимой (*p*-значение большое), то влияние считают отсутствующим или слабым, а соответствующий эффект считают не важным с практической точки зрения.

На самом деле, если переменные значимы, то это не обязательно означает, что соответствующий эффект силен и обладает практической значимостью. И наоборот, если переменные не значимы, то это не означает, что соответствующий эффект не может быть сильным и обладать практической значимостью.

Поясним данную мысль на примере простой модели, в которой есть только константа:

Предположим, что параметр измеряет некоторый эффект, который нас интересует. Предположим, кроме того, что с практической точки зрения эффект силен и важен, если его величина по модулю достаточно велика. Более конкретно, будем считать, что эффект важен, если , и не важен в противном случае, где – некоторое граничное значение. Заметим, что здесь практическая значимость и сила эффекта связана с истинной величиной , которая по предположению является ненаблюдаемой.

Оценивая по наблюдениям параметр методом МНК мы получаем оценку

При стандартных предположениях математическое ожидание этой оценки равно измеряемому эффекту , а теоретическая дисперсия этой оценки равна

где . Вместе с оценкой мы вычисляем ее стандартную ошибку, которая в данном случае равна

где

О статистической значимости эффекта мы, следуя стандартной процедуре, будем судить по *t*-статистике для гипотезы :  
При нормальности ошибок и эта статистика будет подчиняться обычному (центральному) *t*-распределению, а при произвольном – так называемому нецентральному *t*-распределению.

При большом количестве наблюдений () будет стремиться к а распределение *t*-статистики будет похоже на нормальное . Даже если эффект слабый ( заметно меньше ), но не нулевой, при достаточно большом количестве наблюдений математическое ожидание статистики будет большим, так что с очень большой вероятностью будет происходить отклонение нулевой гипотезы и делаться вывод о том, что эффект статистически значим. И это будет вполне корректный статистический вывод для гипотезы . С ростом вероятность отклонить нулевую гипотезу при заданном уровне значимости растет и в пределе равна 100%.

С другой стороны, может быть противоположная ситуация, когда эффект сильный ( заметно больше ), но дисперсия ошибки большая, а количество наблюдений маленькое. Тогда мы очень часто будем принимать нулевую гипотезу и делать вывод о том, что эффект статистически незначим.

Рассмотрим условный пример, когда правительство проводит эксперимент по влиянию введения частичного безусловного базового дохода на желание работать. Из населения выбирается случайно группа людей трудоспособного возраста. Для каждого из этих людей выясняется продолжительность рабочей недели в среднем за последний год. Далее им в течение года выплачивается некоторый фиксированный доход (скажем, 20000 рублей в месяц) и для каждого измеряется продолжительность рабочей недели в среднем за этот год. Пусть для *i*-го человека – это изменение в продолжительности рабочей недели по сравнению с предыдущим годом (в часах). Предположим, что правительство считает практически значимым эффектом среднюю по населению величину , выходящую за интервал 1 ч/нед.

В ситуации, когда , и эксперимент проводится с группой из человек, нулевая гипотеза при уровне значимости 5% принимается с вероятностью около 60%. Здесь эффект является практически значимым, но чаще всего статистически незначим. В этой ситуации собранные данные дают слишком мало информации, чтобы делать уверенные выводы о практической значимости.

В ситуации, когда , и эксперимент проводится с группой из человек, нулевая гипотеза при уровне значимости 1% отклоняется с вероятностью около 92%, а при уровне значимости 5% – с вероятностью около 98%. Здесь эффект не является практически важным, но с вероятностью близкой к 100% будет статистически значимым.

В ситуации, когда , и эксперимент проводится с группой из 0 человек, нулевая гипотеза при уровне значимости 1% или 5% отклоняется с вероятностью около 100%. Здесь эффект является практически важным и почти всегда статистически значим.

На этом примере мы видим, что практическая значимость и статистическая значимость – это разные понятия и их не следует смешивать.

Как же можно делать выводы по поводу практической важности некоторого измеряемого эффекта?

Вообще говоря, *что такое практическая важность и как она связана с коэффициентами регрессии, далеко не всегда понятно*. Мы рассмотрим только сравнительно простую ситуацию, когда нас интересует некоторый количественно выражаемый эффект , который является линейной функцией от коэффициентов регрессии :  
Будем считать, что этот эффект практически важен, если , где – некоторое граничное значение.

Прежде всего, следует понять, что проверка точечной нулевой гипотезы в рассматриваемой ситуации – это очень неудачная идея! Нам требуется разобраться, насколько совместима с наблюдаемыми данными принадлежность величины областям и . Поэтому в данном случае стоит смотреть не на статистики, соответствующие гипотезе , а на интервальную оценку для .

Мы можем вычислить из регрессии оценку на основе МНК-оценки :  
и для этой оценки стандартную ошибку

Отсюда получим доверительный интервал для :

где – двусторонняя граница *t*-распределения для выбранной доверительной вероятности.

Если полученный доверительный интервал отстоит достаточно далеко вправо (влево) от интервала практически неважных эффектов , то можем с большой степенью уверенности сказать, что эффект практически важен и является положительным (соответственно, отрицательным). Если доверительный интервал лежит глубоко внутри интервала практически неважных эффектов , то можем с большой степенью уверенности сказать, что эффект не важен с практической точки зрения. Если же взаимное расположение доверительного интервала и интервала не дает основания для столь уверенных выводов, особенно если доверительный интервал содержит точки как внутри интервала , так и за его пределами, то желательно получить дополнительную информацию, прежде чем делать выводы.

В частном случае сам по себе коэффициент при одной из переменных является измерителем некоторого эффекта (). В этой ситуации желательно при анализе результатов регрессии смотреть не столько на *t*-статистику и *p*-значение, сколько на доверительный интервал для этого коэффициента, который указывает на диапазон правдоподобных значений для интересующего нас эффекта.

## Представление результатов регрессии в компьютерных программах

Существует довольно много различного программного обеспечения для проведения статистических расчетов. Линейная регрессия – это стандартная статистическая процедура, и поэтому возможность ее оценить есть практически в любой статистической программе. Представление результатов расчетов по линейной регрессии может отличаться в разных программах, но сформировался некоторый стандартный «джентельменский набор» показателей, который выдается на экран по умолчанию.

**A** зависимая переменная ;

**B** количество наблюдений ;

**C** количество степеней свободы ;

**D** список регрессоров (включая константу, если она есть) ;

**E** коэффициенты регрессии ;

**F** стандартные ошибки коэффициентов ;

**G** *t*-статистики для регрессоров ;

**H** *p*-значения для *t*-статистик () ;

**I** доверительные интервалы для коэффициентов ;

**J** сумма квадратов остатков ;

**K** (несмещенная) остаточная дисперсия ;

**L** стандартная ошибка регрессии ;

**M** полная сумма квадратов ;

**N** объясненная сумма квадратов ;

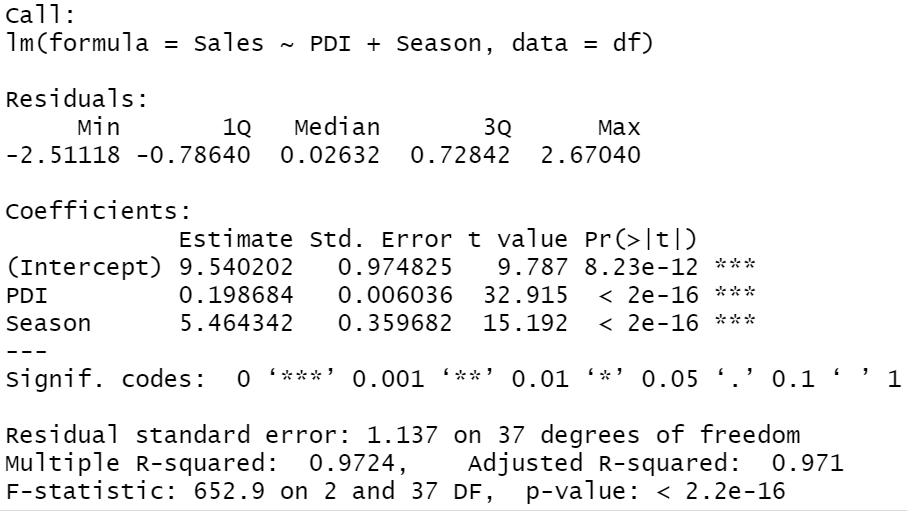
**O** коэффициент детерминации (R-квадрат) ;

**P** скорректированный коэффициент детерминации ;

**Q** *F*-статистика для регрессии в целом ;

**R** *p*-значение для *F*-статистики ().

На рис. 23, 24, 25 представлена одна и та же регрессия, рассчитанная в трех разных программах. Поверх таблиц расставлены метки, указывающие на расположение той или иной информации. Основная часть – таблица с коэффициентами, стандартными ошибками, *t*-статистиками и *p*-значениями – очень похожа. Stata и Excel представляют суммы квадратов в виде так называемой таблицы дисперсионного анализа (англ. ANOVA). В этой таблице также указываются степени свободы и несмещенные дисперсии. R сначала указывает описательную статистику для зависимой переменной.



Q

R

O

P

L

A

C

E

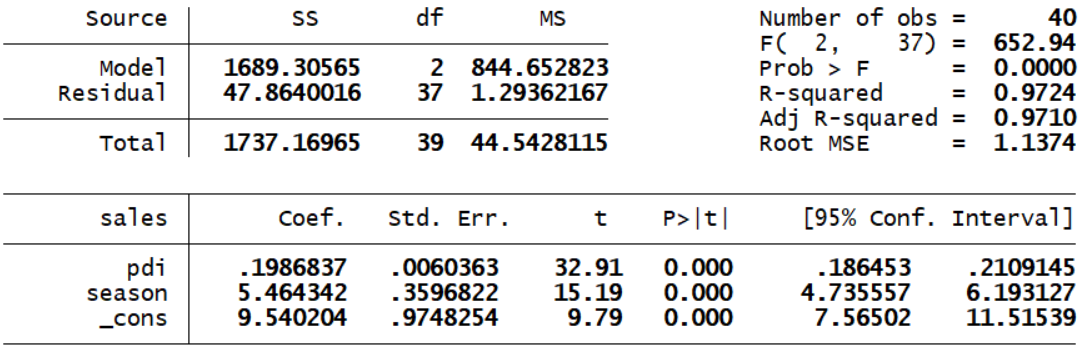
D

F

G

H

Рисунок 23. Отчет программы R



B

Q

R

O

P

L

N

J

C

M

K

E

D

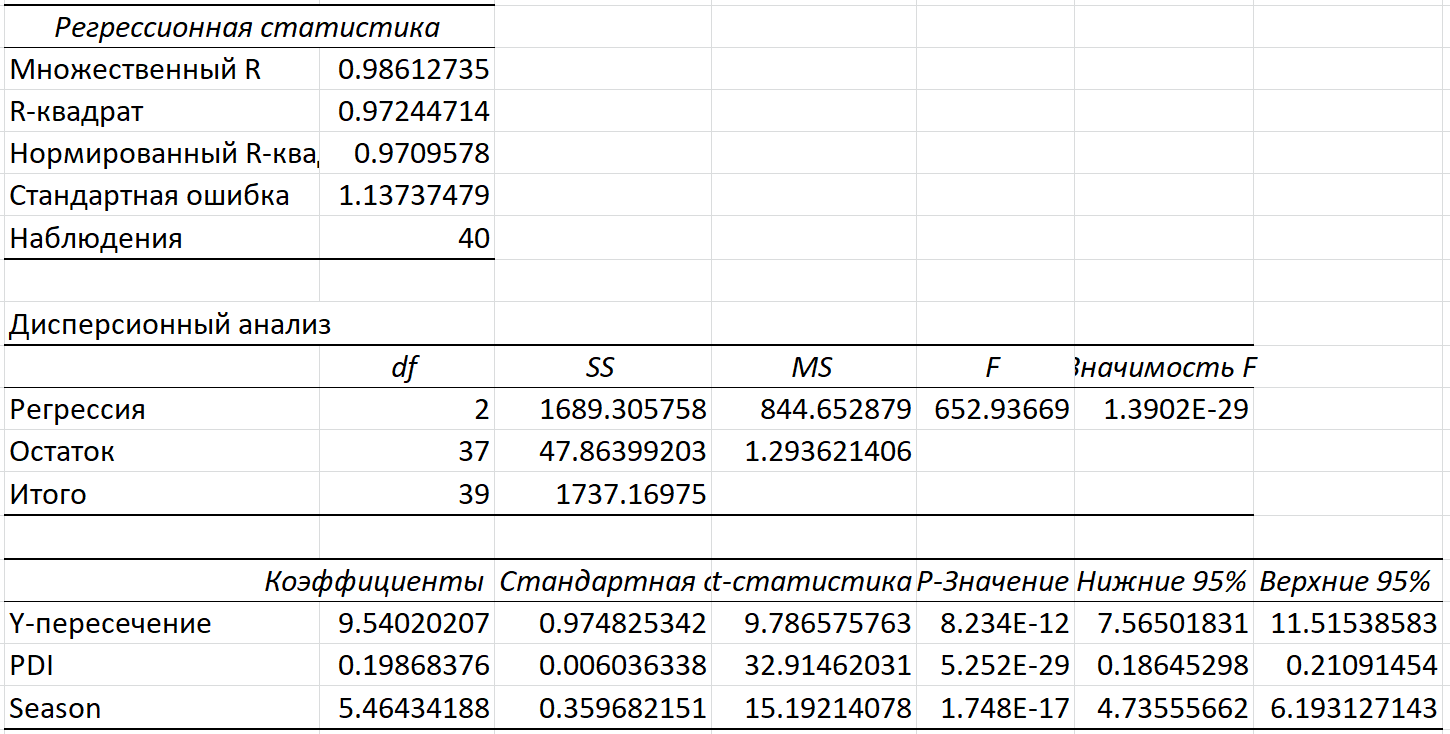
F

G

H

I

Рисунок 24. Отчет программы Stata



B

Q

R

O

P

L

N

J

C

M

K

E

D

F

G

H

I

Рисунок 25. Отчет программы Excel

## Приложение. Вывод оценок МНК с линейными ограничениями

Сумма квадратов остатков как функция от равна  
Как мы знаем, минимум этой функции достигается в точке и равен  
Функцию суммы квадратов остатков несложно переписать в виде

Докажем алгебраически, не прибегая к взятию производных, что минимум той же функции при линейном ограничении достигается в точке  
где

Мы можем вместо минимизировать величину

Подставим сюда  
и получим  
или, с учетом определения матрицы ,

Подставим сюда вместо произвольный вектор , удовлетворяющий ограничению . При этом учтем, что для тоже выполнено , и поэтому . Получим  
Второе слагаемое здесь не зависит от , а первое неотрицательно, поскольку это сумма квадратов элементов вектора . Первое слагаемое достигает минимума, равного нулю, при . Следовательно, действительно достигает минимума в точке . Значит,  
как и требовалось доказать.

## Контрольные вопросы

1. Докажите, что остатки и оценки коэффициентов не коррелируют между собой.
2. Пусть случайные величины независимы и нормально распределены  
   где , и
3. Рассмотрим вектор центрированных . Представьте вектор ввиде Как выглядит матрица ?
4. Является ли матрица симметричной? Идемпотентной?
5. Представьте скаляр в виде . Как выглядит матрица ?
6. Представьте скаляр в виде . Как выглядит матрица ?
7. Представьте скаляр в виде , где Как выглядит матрица ?
8. Является ли матрица симметричной? Идемпотентной?
9. Покажите, что
10. Запишите матрицы (R и r) ограничений на параметры регрессии в случае проверки значимости j-го коэффициента регрессии.
11. Почему в случае незначимости влияния j-го фактора статистика имеет распределение Стьюдента с степенями свободы.
12. Оценивается функция Кобба-Дугласа (в логарифмическом виде) с ограничением однородности первой степени. Запишите матрицы (R и r) ограничений на параметры регрессии.
13. Опишите F-тест для гипотезы о нескольких линейных ограничениях: сформулируйте нулевую и альтернативную гипотезы, способ получения тестовой статистики, её распределение при верной нулевой гипотезе, вид критической области.
14. Запишите матрицы (R и r) ограничений на параметры регрессии в случае проверки гипотезы о незначимости регрессии в целом.
15. Покажите, что
16. Почему в случае незначимости влияния всех факторов статистика , имеет распределение Фишера со степенями свободы и ?
17. Покажите схему проверки незначимости влияния j-го фактора с помощью статистики Фишера.
18. Эквиваленты ли гипотеза: и гипотеза ?

## Экзаменационные вопросы

1. Гипотеза о нормальности ошибок и ее следствия: Распределение оценок МНК. Тестирование одного линейного ограничения коэффициентах регрессии. Статистика Стьюдента. P-значение. Доверительные интервалы для коэффициентов.
2. Гипотеза о нормальности ошибок и ее следствия: Распределение оценок МНК. Проверка гипотез о постоянной отдаче от масштаба для функции Кобба—Дугласа.
3. Гипотеза о нормальности ошибок и ее следствия: Распределение оценок МНК. Проверка совместно несколько линейных ограничений на коэффициенты регрессии. Статистика Фишера. P-значение.

## Литература

1. Вербик М. Путеводитель по современной эконометрике. – М.: Научная книга,. 2008. [стр. 54-69 ]
2. Магнус Я. Р., Катышев П. К., Пересецкий А. А. Эконометрика. Начальный курс:– Учебник. – 6-е изд., перераб. и доп. – М.: Дело, 2004. [стр. 46-51,54-55,78-88]
3. Суслов В. И., Ибрагимов Н. М., Талышева Л. П., Цыплаков А. А. *Эконометрия*. – Новосибирск: СО РАН, 2005. [*стр.230-233* ]
4. [Картаев Ф. С.](https://litgid.com/book_author/18780834/) *Введение в эконометрику. Учебник :*– Москва: Проспект, 2019. [*стр. 48-52, 81-85]*

# Лекция: Выбор регрессоров. Мультиколлинеарность

## Выбор регрессоров. Информационные критерии (AIC, BIC)

**Спецификация модели** состоит в выборе различных аспектов модели, таких как функция, описывающая зависимость между переменными, форма, в которой входят переменные (например, переменная в исходном виде или ее логарифм), набор переменных-регрессоров, которые могут входить в модель, и т. д. Мы рассмотрим здесь только один аспект спецификации модели – выбор регрессоров.

Бывают ситуации, когда нас интересует только одна объясняющая переменная – мы хотим узнать, как она влияет на зависимую переменную. Даже при такой простой постановке исследовательской задачи может быть существенно важным включить в регрессию какие-то другие переменные. Дело в том, что мы, как правило, хотим узнать как одна переменная влияет на другую *при прочих равных условиях*, но равные условия не всегда удается обеспечить. Поэтому важно бывает включить переменные, отвечающие за, так сказать, окружающую обстановку. Такие переменные называют **контрольными переменными**.

Допустим, мы исследуем влияние посещения дополнительных занятий на оценки студентов по какой-то дисциплине. Но студенты не находятся в одинаковых условиях. Например, у студентов со слабой подготовкой может быть более сильный стимул посещать занятия. Или, наоборот, сильные студенты могут быть более дисциплинированными в посещении занятий. Таким образом, при анализе влияния дополнительных занятий на успехи в учебе надо по возможности контролировать предварительную подготовку студентов, их дисциплинированность и другие факторы. Простой регрессией с одной объясняющей переменной здесь не обойтись.

Рассматривается классическая модель линейной регрессии с зависимой переменной и заданным набором потенциальных регрессоров, от которых может зависеть :  
Предполагаем, что коэффициенты при некоторых из регрессоров могут быть нулевыми. Такие регрессоры бывает желательно исключить из модели, чтобы не вносить в нее лишний «шум» и в целом сделать модель более компактной.

Даже если истинный коэффициент при переменной не равен в точности нулю, но довольно мал, ее может быть желательно исключить из модели по тем же соображением. В целом можно поставить задачу выбрать из данного набора регрессоров некоторый набор из регрессоров, где , с тем, чтобы оставить в модели только эти регрессоры. Все методы выбора регрессоров *имеют свои недостатки и смещают оценки*, но некоторые из таких методов неплохо работают на практике и помогают в прикладном моделировании.

### Выбор переменных с помощью проверки гипотез

Основной способ выбора регрессоров, который используют на практике – это тесты на удаление переменных на основе *t*- и *F*-статистик, которые мы рассмотрели ранее. При этом *t*-тест на удаление переменных предназначен для проверки значимости *одного* регрессора, а *F*-тест на удаление переменных для группы из *нескольких* регрессоров (в случае одного регрессора результаты *F*-теста совпадут с результатами двустороннего *t*-теста). Незначимые переменные удаляются из регрессии.

Если в регрессии нет некоторых особых групп переменных (например, групп фиктивных переменных, относящихся к одному качественному признаку), то удобнее всего удалять переменные по одной на основе *t*-статистик, т. е. использовать следующую **шаговую процедуру**:

* Оценивается регрессия по текущему набору переменных.
* Если все переменные значимы при заданном уровне значимости, то процедура останавливается – подходящий набор регрессоров найден.
* В противном случае выбирается «самая незначимая» переменная, т. е. такая, которая имеет наименьшую по модулю *t*-статистику (или, что то же самое, самое большое *p*-значение) и удаляется из набора регрессоров. Процедура при этом продолжается.

При проведении шаговой процедуры удаления переменных следует помнить о следующем:

* Константу лучше всегда оставлять в регрессии, даже если она не значима.
* Если среди регрессоров есть переменные, которые являются особо интересными с точки зрения проводимого исследования, то их тоже лучше оставлять в регрессии, даже если они не значимы.
* Если в регрессии имеется группа связанных между собой по смыслу переменных (например, сезонные фиктивные переменные для разных сезонов), то не следует проверять их значимость по-отдельности. Удаляется вся группа в целом, если она незначима по *F*-статистике.
* Уровень значимости, при котором процедура подбора регрессоров останавливается, выбирается исследователем по своему усмотрению, однако лучше не выбирать слишком низкий уровень значимости, такой как часто используемый уровень 5 %. (Можно взять, например, уровень 20 % или хотя бы 10 %.)

Данные рекомендации объясняются, в числе прочего, следующими соображениями. Наличие в регрессии незначимых переменных не является ошибкой спецификации. С теоретической точки зрения МНК и другие методы оценивания могут оценить коэффициенты, даже если они нулевые. В то же время *удаление переменных на основе тестов всегда приводит к смещению оценок*. Тут используется уже не просто МНК, который, как мы видели, при соответствующих предположениях дает несмещенные оценки, а **МНК в комбинации с предварительным тестированием**.

Если в окончательно выбранной регрессии остались малозначимые, но интересные с точки зрения проводимого исследования переменные, то это дает возможность читателям отчета об исследовании своими глазами увидеть, что данные переменные были учтены, но оказались незначимыми. Кроме того, читатели могут захотеть по своим соображениям ориентироваться на какой-то другой уровень значимости, чем автор исследования.

Вообще, как уже обсуждалось, статистическую значимость не следует смешивать с практической значимостью. В последние годы все больше уважаемых известных исследователей говорят о том, что в статистических исследованиях *надо очень осторожно относиться к тестам значимости и p-значениям* и не абсолютизировать их роль. Это серьезная проблема и о ней написано много работ.

Выше рассмотрен подход **«сверху вниз»** или «от общего к частному», когда переменные исключаются из регрессии.

Существует и противоположный подход – **«снизу вверх»** ли «от частного к общему». При таком подходе тестируются потенциальные регрессоры, которых нет в текущем наборе – не следует ли их добавить в регрессию. Здесь тоже можно использовать *t*-статистики (*F*-статистики в случае нескольких переменных). Это **тесты на добавление переменных**.

К подходу «снизу вверх» следует относиться с еще большей осторожностью, чем к подходу «сверху вниз». Основная причина состоит в следующем: если в текущей регрессии пропущено несколько нужных переменных с ненулевыми коэффициентами, то регрессия некорректна. Тесты на добавление отдельных переменных в такой регрессии будут проводиться на некорректной модели и их результаты могут быть ложными. Например, могут существовать две нужные переменные, которые при использовании *t*-статистик на добавление переменных незначимы, но если бы использовалась *F*-статистика на добавление обеих переменных, то она могла бы показать их большую значимость.

Более того, в подходе «снизу вверх» сложно контролировать ошибку первого рода (отклонение верной нулевой гипотезы). Пусть для *t*-теста на добавление переменной выбран уровень значимости . Вероятность ошибиться, включив ненужную переменную при одиночном тестировании, будет равна . Однако если тестируется сразу несколько переменных, то вероятность включить ненужную переменную возрастает. Если, тестируется независимых между собой ненужных переменных, то вероятность того, что хотя бы одна из них значима по *t*-статистике на уровне , будет равна . Например, при и данная вероятность будет равна , т. е. примерно 40%.

Таким образом, с одной стороны, при использовании подхода «снизу вверх» *важные взаимосвязи могут оказаться необнаруженными* из-за того, что они связаны с комбинированным эффектом двух и более переменных. С другой стороны, этот подход при большом числе потенциальных регрессоров *может привести к включению в модель большого количества «мусора» и появлению ложных «научных открытий»* о взаимосвязях между переменными. Особенно велика такая опасность при применении механического перебора. В связи с указанными проблемами рекомендуется выбирать достаточно полный набор регрессоров исходя из содержательного экономического анализа, а затем использовать подход «сверху вниз».

Однако следует понимать, что подход **«**снизу вверх**»** тоже широко используется и полезен. В частности, исследователь вполне может по ошибке не включить какую-то важную переменную в модель. Тогда модель может оказаться некорректной. Поэтому желательно уже построенные модели подвергать различным диагностическим проверкам на предмет выявления каких-то упущенных важных эффектов. **Диагностика модели** – это очень важная часть эконометрического моделирования.

### Выбор регрессоров с помощью числовых показателей

Пусть есть несколько разных вариантов регрессий. Можно ли предложить некоторый числовой показатель, чтобы его можно было посчитать для каждой из регрессий и выбрать ту регрессию, для которой значение показателя наибольшее (или, наоборот, наименьшее)? Да, такие показатели существуют и некоторые из них неплохо зарекомендовали себя на практике, хотя достаточно строгого теоретического обоснования для них нет.

Регрессия основана на минимизации суммы квадратов остатков. Нельзя ли показатель суммы квадратов остатков регрессии () использовать для того, чтобы выбирать регрессоры? В исходном виде использовать для выбора регрессоров не получится. Также не получится использовать и связанные с ним показатели смещенной остаточной дисперсии , соответствующей стандартной ошибки регрессии и коэффициента детерминации . Дело в том, что при добавлении дополнительных регрессоров , и всегда падают, а всегда растет. (Если быть более точным, в особых случаях эти показатели могут не измениться.) В предельном случае, когда количество регрессоров (включая константу) равно количеству наблюдений (), остатки в регрессии становятся равными нулю и будет выполняться , и .

Все указанные статистики – это показатели **точности подбора** регрессии к наблюдаемым данным, которые не учитывают, что включение большого количества ненужных переменных может быть нежелательным. Поэтому кроме точности подбора следует как-то учитывать *количество переменных* в модели и штрафовать за большое количество переменных, мало увеличивающих точность.

Одним из первых предложенных показателей такого рода был так называемый **скорректированный коэффициент детерминации**. Он строится по образцу обычного коэффициента детерминации, но происходит корректировка на количество степеней свободы:  
Здесь  
это несмещенная остаточная дисперсия в рассматриваемой регрессии с степенями свободы, а  
это несмещенная остаточная дисперсия в регрессии с одной константой (там степень свободы). Чем меньше (т. е. больше точность подгонки), тем больше. Чем меньше количество объясняющих переменных , тем больше. Выбирается модель с наибольшим значением . (Заметим, что скорректированный *R*-квадрат в отличие от обычного при малой точности подгонки может быть отрицательным.)

Вместо можно использовать просто несмещенную остаточную дисперсию . При этом выбирается модель с наименьшей диспер­сией.

Скорректированный коэффициент детерминации обладает следующим свойством: при добавлении дополнительного регрессора возрастает тогда и только тогда, когда соответствующая *t*-стати­стика для добавления переменной по модулю больше 1. При большом количестве степеней свободы такое критическое значение соответствует уровню значимости примерно 32 %. Таким образом, при использовании для выбора регрессоров в модель будут включены не очень значимые переменные. Таким образом, хотя включает штраф за количество переменных, но этот штраф не очень сильный.

Другие известные числовые показатели для выбора регрессоров – это так называемые информационные критерии. Самые популярные – это информационный критерий Акаике и байесовский информационный критерий.

Информационный критерий Акаи́ке:

Байесовский информационный критерий (информационный критерий Шварца):  
где – общее количество параметров в модели, – значение функции правдоподобия в максимуме. В рассматриваемом нами случае регрессии с константой и неизвестной дисперсией количество параметров равно

Максимум функции правдоподобия для классической модели регрессии с нормальными ошибками равен

Таким образом,

Определения данных критериев в разных источниках несколько отличаются. Например, иногда при подсчете количества параметров не учитывают дисперсию, т. е. берут . Иногда приведенные выше величины делят дополнительно на количество наблюдений .

Из нескольких сравниваемых регрессий выбирается такая, которая соответствует наименьшему значению AIC (или BIC). Оба критерия при прочих равных условиях отдают предпочтение моделям с малой величиной и малым числом переменных . При не очень малом числе наблюдений () BIC является более жестким критерием, чем AIC так как коэффициент при штрафном слагаемом у него больше ().

Информационным критериям, так же как и показателю , можно сопоставить уровень значимости для тестирования добавления одной переменной, но здесь уровень значимости сильнее зависит от количества наблюдений. Асимптотически при большом для AIC уровень будет примерно 16 %. Для BIC такого асимптотического значения нет, этот критерий с ростом становится все жестче и жестче. В правдоподобном случае, когда и , BIC соответствует уровень значимости примерно 4 %. Таким образом, часто используемый в статистике уровень значимости 5 % может при использовании шаговой процедуры «сверху вниз» давать результаты, похожие на BIC. Если же и , то BIC соответствует уровень значимости меньше 1 %.

### Проблемы с процедурами выбора регрессоров и используемыми в них статистиками

Рассмотренные процедуры выбора регрессоров становятся ненадежными, когда количество регрессоров сопоставимо с количеством наблюдений. Например, можно проделать эксперимент, в котором все регрессоры (кроме константы) генерируются с помощью датчика случайных чисел и никак не связаны с зависимой переменной. При большом количестве регрессоров разброс значений показателя будет очень большим и с довольно большой вероятностью будет получаться «статистически значимая» по *F*-статистике регрессия. В пределе, когда количество степеней свободы очень маленькое, заметная доля сгенерированных регрессий будет иметь очень высокий .

Боле того, рассмотренные процедуры выбора регрессоров правильно работают только тогда, когда для исходного набора регрессоров модель правильно специфицирована, то есть выполнены предположения классической модели линейной регрессии. В противном случае здесь можно получить некорректный результат.

Следует понимать, что хотя и информационные критерии в определенном смысле используются для выбора «наиболее удачной» модели, но они ни в коем случае *не отражают качество модели*. Если большой, то это не значит, что модель «хорошая», а если маленький, то «плохая». Коэффициент детерминации показывает, насколько близко мы аппроксимировали зависимую переменную с помощью регрессоров, но он ничего не говорит о корректности самой модели, из которой мы исходим. Так же точно, если в регрессии переменные значимы, то это не значит, что модель «хорошая», а если незначимы, то «плохая».

Если взять произвольную переменную в качестве зависимой и произвольный набор других переменных той же длины в качестве регрессоров, то всегда можно применить к этим данным МНК, получить оценки коэффициентов и разные статистики. Далее можно к любой регрессии применить процедуры выбора регрессоров. Однако далеко не всякая подобная регрессия будет правильно специфицирована и иметь хоть какой-то содержательный смысл.

В принципе совсем не сложно построить регрессию, в которой будет высокий , все переменные будут значимыми, но сама регрессия будет полностью некорректной и бессмысленной.

Выше мы уже обсуждали эксперимент, в котором регрессоры генерируются с помощью датчика случайных чисел и никак не связаны с зависимой переменной, причем количество регрессоров сопоставимо с количеством наблюдений (например, , ). Применим к каждой такой регрессии шаговую процедуру и оставим только значимые переменные (например, на уровне 10 %). С очень большой вероятностью мы получим модель с несколькими «статистически значимыми» переменными, которая сильно «статистически значима» по *F*-статистике. А если взять исходное количество переменных близким к , то получаемые бессмысленные регрессии еще и будут достаточно часто иметь высокий .

Кроме того, следует учесть, что если, добавив переменную в регрессию, мы существенно повысили коэффициент детерминации, то мы получили более близкую аппроксимацию для зависимой переменой, но не более того. Для полученной модели могут перестать выполняться нужные предположения. Если модель неверно специфицирована, то оценки коэффициентов могут оказаться сильно смещенными, а *t*- и *F*-статистики для них – полностью некорректными.

Особенно «коварным» бывает нарушение предположения об экзогенности регрессоров. Например, можно справа поставить зависимую переменную и получить . Конечно, это бессмысленная модель и вряд ли кто-то ее станет оценивать. Но бывают и более тонкие случаи нарушения экзогенности, когда наличие проблем с эндогенностью неочевидно даже для человека, поднаторевшего в эконометрике.

Коэффициент детерминации – это доля, т. е. безразмерная величина. Поэтому может показаться, что можно сравнивать с помощью него любые регрессии. Однако это совсем не так. Если в регрессиях разные зависимые переменные, то их нельзя сравнивать по коэффициенту детерминации. Например, если в одной регрессии зависимая переменная (где ), а в другой – , то коэффициенты детерминации в них совершенно несопоставимы.

Коэффициент детерминации мы определили для линейной регрессии. В то же время для многих широко используемых эконометрических моделей нет аналога и . Для ряда моделей аналоги коэффициента детерминации предложены, но смысл их не очень понятен.

В целом желательно не преувеличивать значение коэффициента детерминации. Вербик: «величина , как правило, не самая важная характеристика наших результатов оценивания».

## Мультиколлинеарность

Как мы видели, при наличии линейной зависимости между регрессорами (полной мультиколлинеарности регрессоров) возникает проблема однозначной идентификации коэффициентов регрессии. Проблема проявляется в неоднозначности как теоретической модели, так и оценок МНК, а также в невозможности применить обычную формулу для оценок МНК

Линейная зависимость между регрессорами наблюдается, например, когда количество наблюдений меньше количества коэффициентов . Ранг матрицы регрессоров не может быть больше числа строк , поэтому в рассматриваемом случае он будет меньше . Ясно, что если , то у нас не хватает информации, чтобы однозначно оценить коэффициенты. Например, если и , то мы имеем единственную точку, чтобы подобрать по ней линию регрессии; эта задача не имеет однозначного решения – через данную точку можно провести бесконечно много прямых с разными наклонами.

Дело здесь не в самом по себе количестве наблюдений, а в целом — в отсутствии адекватной информации для выявления изучаемой зависимости. Например, если исследователь хочет изучить влияние курения на производительность работников, но в его выборку не попал ни один курильщик, то зависимость будет невозможно оценить.

Проблема полной мультиколлинеарности может возникнуть также, если исследователь по ошибке включает в регрессию однозначно взаимосвязанные переменные. Например, если строится регрессия по регионам и в число объясняющих переменных входят доля работоспособного населения и доля неработоспособного населения, то возникает линейная зависимость: в сумме эти две переменные дают вектор из единиц. Здесь имеет место избыточность информации в данных: если мы знаем одну переменную, то можем однозначно рассчитать другую по линейной формуле. Достаточно оставить в регрессии одну из двух переменных. Очень похожая проблема возникает при использовании в регрессии полного набора фиктивных переменных; об этом речь пойдет в соответствующем разделе.

Кроме полной мультиколлинеарности бывают ситуации, когда набор регрессоров в некотором смысле близок к линейной зависимости. Обычно такую ситуацию и называют **мультиколлинеарностью**.

Самый распространенный случай мультиколлинеарности – это когда два регрессора (скажем и ) сильно коррелируют между собой, т. е. (или ). При этом получается, что система из этих двух регрессоров и вектора из единиц, соответствующего константе регрессии, близка к линейной зависимости.

Например, если рассматривается зависимость цены квартиры от жилой площади и от общей площади, то для некоторых классов квартир два показателя площади будут сильно коррелировать.

Наличие сильных корреляций между объясняющими переменными можно заметить, построив таблицу их попарных корреляций. Такая возможность есть в большинстве статистических компьютерных программ.

Если мультиколлинеарность затрагивает более двух объясняющих переменных, то парные корреляции могут не продемонстрировать ее наличия. Можно тогда рассмотреть серию регрессий, в каждой из которых зависимая переменная – это один из регрессоров (), а регрессоры – остальные исходные регрессоры () и константа. Если коэффициент детерминации в такой регрессии (обозначим его ) очень близок к единице, то это свидетельствует о наличии мультиколлинеарности.

Следствием мультиколлинеарности является **слабая идентификация** (неинформативность имеющихся наблюдений с точки зрения интересующей нас зависимости): мы не можем точно оценить коэффициенты регрессии. Мы можем понять, что цена квартиры, вообще говоря, сильно зависит от ее площади, но каждый из отдельных коэффициентов при двух регрессорах – общей и жилой площади – может иметь достаточно большую стандартную ошибку и даже быть незначимым по *t*-статистике.

С мультиколлинеарностью, среди прочего, связана необходимость использовать *F*-ста­тис­тику для теста на значимость группы переменных, а также необходимость удалять незначимые по *t*-статистикам переменные только по одной, а не все сразу. Достаточно часто бывает, что переменная, которая была незначимой, становится значимой при удалении другой незначимой переменной.

В примере с общей и жилой площадью мы можем получить, что каждая из этих переменных по *t*-статистике незначима, в то время как по *F*-ста­тис­тике они в совокупности значимы. При удалении одной из этих переменных другая становится значима по *t*-статистике. Мы совершили бы серьезную ошибку, если бы удалили из регрессии сразу обе переменные на том основании, что они не значимы по *t*-статистикам.

О том, насколько сильно зависимость между регрессорами влияет на точность оценивания коэффициентов можно судить по показателю коэффициента увеличения дисперсии (англ. *variance inflation factor*, *VIF*). Коэффициент показывает, насколько дисперсия оценки коэффициента при переменной больше, чем дисперсия такой же оценки в гипотетической ситуации, когда переменная никак не связана с остальными объясняющими переменными, т. е. в описанной выше регрессии. Коэффициент увеличения дисперсии можно рассчитать по формуле  
Если очень большой, то дисперсия оценки во много раз больше, чем могла бы быть в гипотетической ситуации отсутствия мультиколлинеарности.

Мы дали неформальное определение мультиколлинеарности, описали последствия и рассмотрели ряд показателей, по которым ее можно обнаружить. Но здесь важно сделать несколько важных оговорок.

* Не существует каких-то определенных границ для рассмотренных показателей. Нельзя сказать, что если по модулю больше, чем, скажем, , то это уже мультиколлинеарность. То же самое с и .
* Мультиколлинеарность – это не нарушение предположений модели линейной регрессии. Можно рассмотреть все наши гипотезы (A0)–(A5). Ни в одной из них не говорится о мультиколлинеарности.
* Хотя из-за мультиколлинеарности некоторые элементы ковариационной матрицы оценок МНК могут быть большими, и стандартные ошибки оценок некоторых коэффициентов могут быть большими, но мультиколлинеарность не приводит к некорректности оценок этих показателей. Да стандартная ошибка большая, но мы это видим по выдаваемым нашей компьютерной программой результатам.
* Проблема мультиколлинеарности во многом неправомерно раздута авторами некоторых учебников и руководств по регрессионному анализу. В то же время, многие уважаемые ученые-статистики призывают не преувеличивать значение этого явления.
* Мультиколлинеарность – это проблема недостаточности имеющейся информации.

С практической точки зрения все это означает, что мы можем заниматься моделированием, не обращая особого внимания на мультиколлинеарность и не высчитывать каждый раз перечисленные выше показатели. Даже если мы видим, что коэффициент корреляции между и равен 0,9999, это вовсе не означает, что мы обязаны удалить одну из этих переменных из регрессии и вообще должны как-то волноваться по этому поводу.

Например, в примере с ценой квартир мы можем получать очень точный прогноз цены даже когда отдельные коэффициенты незначимы по *t*-статистикам. Мы можем выбрать два пути. Один путь – оставить в регрессии как общую, так и жилую площадь. Это не помешает прогнозированию. Другой путь – удалить самую незначимую из двух переменных на том основании, что информация фактически дублируется и удаляемая переменная не помогает сколько-нибудь заметно улучшить прогноз цены.

Если регрессия применяется не для прогнозирования, а для структурного анализа зависимостей, то бывает важно достаточно точно знать значения отдельных коэффициентов регрессии. Мультиколлинеарность может этому мешать. Но надо понимать, что поскольку мультиколлинеарность – это проблема недостаточности имеющейся информации, то *без получения дополнительной информации с этой проблемой невозможно справиться*. Если другой информации нет, то надо просто привести полученные оцени и их стандартные ошибки.

Например, если и тесно коррелированы, то можно наложить ограничение, что коэффициент при одной из переменных равен нулю. Правда, тогда возникает вопрос: если мы заранее знали, что коэффициент при переменной равен нулю, то зачем мы ее вообще рассматривали как потенциальный регрессор? Такое произвольное внесение информации постфактум не очень обосновано с методологической точки зрения.

Более оправданы усилия по поиску дополнительных наблюдений, которые могут содержать более полную информацию. Так в примере с курением можно попробовать найти дополнительные данные, которые бы включали работников-курильщиков.

Наиболее благоприятная ситуация с точки зрения получения дополнительной информации – это когда исследователь использует экспериментальные данные и сам выбирает значения объясняющих переменных в эксперименте. Можно так спланировать эксперимент, чтобы разные объясняющие переменные были между собой почти некоррелированны, а их значения колебались в достаточно широких пределах, чтобы экспериментальные данные были как можно более информативными с точки зрения оценки коэффициентов. К сожалению, такая ситуация довольно редка в экономических исследованиях.

## Отчеты по моделям регрессии

Мы уже видели, как представляют результаты расчетов по отдельной регрессии различные компьютерные программы. Часто при оформлении расчетных заданий или квалификационных работ студенты вставляют в текст такие выданные программой таблицы. Такой «ленивый» подход приводит к тому, что, во-первых, работа выглядит неэстетично, а, во-вторых, ее сложно читать и воспринимать результаты. В целом это свидетельствует о непрофессионализме составителя отчета.

Составление отчета по результатам эконометрического моделирования требует отдельных усилий и тщательного обдумывания. Надо, прежде всего, разобраться, какие сведения являются важными для потенциального читателя отчета, а какие являются неинформативными техническими деталями, которые только «захламляют» отчет. Ключевые сведения (например, использованную зависимую переменную) надо обязательно включить в отчет, а ненужную информацию (например, лишние знаки после запятой) исключить из него.

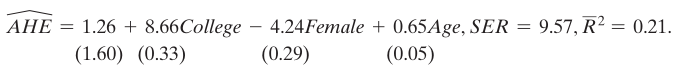
Отчет по результатам моделирования должен включать не только некое финальное уравнение регрессии, но и основные вехи самого процесса моделирования – какие данные использовались для оценивания, как данные анализировались и обрабатывались, на каком основании выбрано именно это уравнение и т. п. Часто имеет смысл привести несколько основных альтернативных моделей, вместе со статистиками, которые обосновывают выбор между ними, чтобы были достаточно ясно видны последствия выбора той или иной спецификации. Отчет можно дополнить основными сведениями о проведенных диагностических процедурах, информативными графиками. (О диагностике регрессии речь пойдет в дальнейшем.)

Можно дать следующие общие рекомендации.

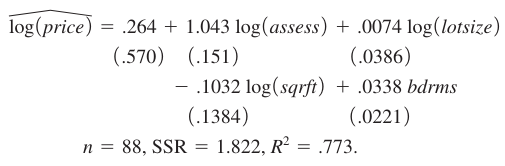
* Не использовать выдачу компьютерной программы, а самому оформить аккуратный ясный отчет.
* Явно указать зависимую переменную регрессии.
* Указать, по каким наблюдениям строится регрессия, и сколько наблюдений было использовано. Для временных рядов указать период.
* Цифры округлить до небольшого числа значащих цифр. Например, коэффициенты до 3–5 цифр, стандартные ошибки до 2–4 цифр, и т. д.
* Для коэффициентов привести статистики, показывающие точность их оценивания. Лучше всего привести стандартные ошибки, но можно использовать и другие варианты (интервалы, *t*-статистики, *p*-значения). Не следует давать все показатели сразу, чтобы не загромождать отчет.
* Можно звездочками указать статистическую значимость оценок коэффициентов. Например, \* – значим на уровне 5%, \*\* – значим на уровне 1%, и т. п. (При этом следует помнить, что большое количество звездочек не служит признаком качества модели, а статистическая значимость – это не то же, что практическая значимость.)
* Привести наиболее информативные статистики по регрессии в целом (Например, , стандартную ошибку, AIC и т. п.).
* Можно использовать как запись в виде уравнения, так и таблицу. Таблица удобнее, если сравниваются несколько разных моделей. В такой таблице исключенные переменные указываются прочерками. В уравнении стандартные ошибки (или другие показатели) приводятся под коэффициентами в скобках. Часто так же поступают с таблицами.

Чтобы пояснить сказанное, приведем несколько примеров оформления оцененных уравнений регрессии из известных учебников по эконометрике.

Сток, Уотсон:



Wooldridge:



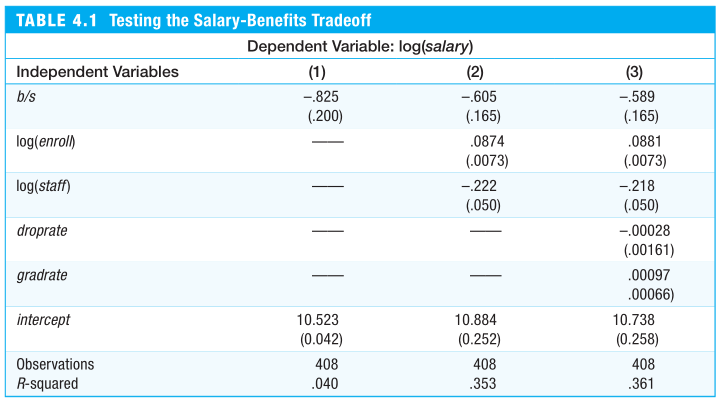
Вербик (Таблица 3.1):



Сток, Уотсон:



Wooldridge:



## Контрольные вопросы

1. Как получена формула для коэффициента детерминации, скорректированного на число степеней свободы?
2. Покажите, что для линейной модели регрессии при k > 1 cкорректированный коэффициент детерминации не больше, чем коэффициент детерминации.
3. Показать, что добавление новых факторов в регрессию не меняет оценки параметров при ранее введенных факторах (“старые” оценки параметров), если
   1. коэффициент детерминации в регрессии по ранее введенным факторам уже равен единице;
   2. вектор остатков в регрессии по ранее введенным факторам ортогонален новым факторам;
   3. новые факторы ортогональны ранее введенным факторам.
4. Что такое мультиколлинеарность, и ее основные последствия?
5. Как можно проверить наличия мультиколлинеарности?
6. Перечислите наиболее характерные признаки мультиколлинеарности.
7. Пусть оценивается регрессия с 2-мя факторами. Как дисперсия МНК-оценки связана с коэффициентом корреляции между факторами?

## Экзаменационные вопросы

1. Выбор регрессоров с помощью проверки гипотез. Диагностика мультиколлинеарности (корреляции факторов, коэффициенты детерминации для факторов).
2. Выбор регрессоров с помощью числовых показателей. Диагностика мультиколлинеарности (корреляции факторов, коэффициенты детерминации для факторов).

## Литература

1. Вербик М. Путеводитель по современной эконометрике. – М.: Научная книга,. 2008. [стр. 51-54, 81-86]
2. Магнус Я. Р., Катышев П. К., Пересецкий А. А. Эконометрика. Начальный курс: – Учебник. – 6-е изд., перераб. и доп. – М.: Дело, 2004. [стр. 74-78, 109-112]
3. Суслов В. И., Ибрагимов Н. М., Талышева Л. П., Цыплаков А. А. Эконометрия.– Новосибирск: СО РАН, 2005. [стр. 234-244]
4. Картаев Ф. С. Введение в эконометрику. Учебник : – Москва: Проспект, 2019. [стр. 78-80,106-110]

# Лекция: Фиктивные переменные – качественные объясняющие переменные в регрессии

## Фиктивные переменные

Переменные, которые используются в регрессиях, принимают числовые значения на действительной прямой. Это так называемые количественные переменные. Примеры из экономики: объем производства, цена, заработная плата, издержки и т. п. Можно ли учесть в регрессии качественную информацию? Да, можно.

Качественные информацию, как правило, учитывают путем разбиения наблюдений на определенные **группы** (**категории**). Как правило, в исходном виде соответствующие **качественные (категориальные) переменные** определенным образом кодируются. Например, у нас может быть переменная, в которой каждое наблюдение имеет метку «A», «B» и «С». Для мужчин и женщин могут использоваться метки «м», «ж». Также могут использоваться числовые коды (2, 3, 4, 5 для оценок по курсу). На основе качественной переменной можно далее построить набор **фиктивных переменных** (англ. *dummy variable* или просто *dummy*), по одной для каждой группы. Это переменные, принимающие значения 1 и 0: 1, если наблюдение принадлежит данной группе, и 0 в противном случае. (Такие переменные также называют бинарными, индикаторным и т. п.). Эти переменные уже можно непосредственно ввести в регрессию для отражения качественной информации.

### Случай одной качественной переменной

Пусть, например, требуется изучить влияние пола на заработную плату. Тогда в качестве зависимой переменной можно взять заработную плату в рублях, а в качестве объясняющей – следующую фиктивную переменную:

Кроме фиктивной переменной имеет смысл ввести количественные объясняющие переменные в роли контрольных переменных, поскольку, скорее всего, мы захотим изучить влияние пола *при прочих равных условиях*. Пусть, например, в регрессию будет входить – стаж работы в годах. Другие объясняющие переменные для простоты не будем включать. Тогда модель будет иметь следующий вид:

По виду это обычная линейная регрессия (выполнено предположение A0). Если она удовлетворяет и другим предположениям классической модели регрессии, то можем применить МНК для ее оценки. Оценив модель, мы, в частности, получим оценку параметра . По смыслу будет показывать, на сколько рублей зарплата женщины при прочих равных условиях отличается от зарплаты мужчины. Если , то зарплата женщин при прочих равных условиях ниже, чем зарплата мужчин. Поскольку это обычная модель регрессии, то можем использовать обычную проверку гипотез. Так можно проверить гипотезу и тем самым выяснить, является ли различие в зарплатах мужчин и женщин статистически значимым.

Другой характерный пример – **сезонные фиктивные переменные**. Пусть мы используем для оценки регрессии квартальные данные за несколько лет, и пусть – объем продаж продукции фирмы, – цена продукции. Если спрос на продукцию носит сезонный характер (например, это верно для пива), то продажи могут зависеть от квартала. Тогда можем ввести в модель четыре квартальные сезонные переменные, :

В результате модель будет иметь вид

При использовании фиктивных переменных в регрессии надо учитывать некоторые особенности. Мы рассмотрим их на примере, но особенности имеют общий характер. Пусть зависимая переменная , количественная объясняющая переменная , а наблюдения разбиты на три группы: , и . Для каждой из групп имеется соответствующая фиктивная переменная. Например, для группы

Уравнение регрессии имеет вид  
Коэффициенты , и в данной модели интерпретируются как **эффекты групп**, и . Они показывают, как влияет на зависимую переменную принадлежность к той или иной группе.

В этой регрессии имеет место *линейная зависимость между регрессорами* (полная мультиколлинеарность). Сумма трех фиктивных переменных равна единице:  
Это связано с тем, что каждое наблюдение принадлежит одной и только одной из категорий , и . Если мы знаем значение двух из фиктивных переменных, то можем вычислить и значение третьей. Матрица регрессоров имеет вид  
В ней столбцы **,**, и линейно зависимы между собой. Оценки МНК нельзя рассчитать по обычной формуле, так как матрица не имеет обратной.

Проблема в том, что коэффициенты , , и нельзя идентифицировать однозначно. Нужно наложить какое-нибудь подходящее ограничение на коэффициенты, чтобы обеспечить идентификацию.

Есть несколько удобных для практического применения ограничений. Самый простой и распространенный подход состоит в том, что одна из групп берется в качестве базовой и коэффициент при ней приравнивается к нулю. Например, примем группу за базовую и наложим ограничение . Это означает, что мы исключаем фиктивную переменную из модели:  
В полученной модели переменные **,** и линейно независимы, поэтому проблема исчезает.

Коэффициенты , – это эффекты групп и , и они являются *относительными*. То есть показывает, насколько выше в группе по сравнению с группой , а показывает, насколько выше в группе по сравнению с группой при прочих равных условиях (рис. 26).

A

Рисунок 26. Иллюстрация регрессии, в которой группа является базовой

Можно взять и другую группу в качестве **базовой**. При этом модель останется по сути такой же, но интерпретация эффектов групп поменяется. Положим, например, . Запишем соответствующее уравнение, пометив коэффициенты в нем штрихами:

Мы можем однозначно пересчитать коэффициент одной модели в коэффициенты другой, что и доказывает их эквивалентность.

Действительно, заметим, что и подставим в уравнение первой модели:  
что можно переписать в виде  
Приравнивая коэффициенты при соответствующих переменных, получим систему, взаимно-однозначно связывающую два набора коэффициентов:

При этом пересчитывается только константа и коэффициенты при фиктивных переменных, а коэффициент при количественной переменной одинаковый в обеих моделях.

Другой подход состоит в том, чтобы наложить ограничение и удалить константу из регрессии:

В такой модели получается как бы несколько разных констант – по одной константе для каждой из групп (рис. 27).

Рисунок 27. Иллюстрация регрессии без общей константы

Еще один подход состоит в том, чтобы сделать средний эффект равным нулю. Тогда можно интерпретировать как общую для всех групп константу, а , , – как отклонения от общего уровня в ту или иную сторону в зависимости от принадлежности наблюдения к определенной группе. Пусть , – количество наблюдений в каждой из групп. Накладывается ограничение, что среднее взвешенное эффектов групп равно нулю:

Для оценивания можно выразить один эффект через остальные и подставить в исходное уравнение. Например,

Получим модель  
Здесь исключается из регрессии, а в и вносятся поправки. После оценивания модели мы будем иметь оценки для и , а оценку для найдем из приведенной выше формулы .

Если в группах примерно равное число наблюдений (), то удобнее использовать ограничение, что среднее арифметическое эффектов равно нулю, т. е.

В каждом из рассмотренных вариантов на коэффициенты , , и накладывается одно линейное ограничение, которое позволяет обеспечить идентификации. *Все эти варианты эквивалентны между собой и различаются только интерпретацией коэффициентов*.

Здесь важно следить, чтобы в каждой из групп было достаточно много наблюдений для оценивания модели. По крайней мере, количества наблюдений в каждой группе должны быть не равны нулю (, ).

Вернемся к рассмотренным ранее примерам.

В примере с заработной платой мужчин и женщин мы ввели одну фиктивную переменную – для женщин. К ней существует парная , которая принимает значение 1 для мужчин. Поскольку не вводилась в регрессию, то в ней не было линейной зависимости. Фактически, мы просто взяли мужчин за базовую группу.

В примере с квартальными данными и сезонностью в записанном уравнении есть линейная зависимость. Чтобы от нее избавиться, мы можем, например, взять один из кварталов за базовый, или принять среднюю сезонность за ноль, т. е. наложить ограничение

Второй вариант может быть более удобным с точки зрения интерпретации, но его несколько сложнее оценивать.

Мы рассмотрели случай с одной качественной переменной. Можно учесть в регрессии сразу несколько качественных переменных. Например, для регрессии с заработной платой можно рассмотреть одновременно пол человека и уровень образования (начальное, среднее, высшее).

### Случай двух качественных переменных

Рассмотрим более подробно случай двух качественных переменных: , , (3 категории) и , (2 категории). Исходная модель будет иметь вид

В данной модели две линейные зависимости между регрессорами:  
и

Заметим сразу, что не получится избавится от линейных зависимостей, наложив ограничение . Требуется наложить *два* ограничения на коэффициенты , , , , и . Самый простой подход – исключить по одной фиктивной переменной из каждого набора. Например, можно наложить ограничения

Модель примет вид

При проверке на значимость качественной переменной имеет смысл проверять коэффициенты при всех фиктивных переменных из соответствующего набора одновременно. В рассматриваемой модели, исключена переменная , поэтому следует проверить на равенство нулю коэффициенты при оставшихся переменных и . Поскольку коэффициентов больше одного, для проверки следует использовать *F*-статистику. Для качественной переменной , можно проверять по *t*-статистике.

Можно усложнить модель, включив **эффекты взаимодействия** между качественными переменными (эффекты второго порядка), т. е. совместного влияния качественных факторов. Это может быть в случае, если сочетания признаков позволяют улучшить объяснение зависимой переменной. Например, в примере с зарплатой женщина с высшим образованием может иметь дополнительный бонус к зарплате, который не сводится просто к сумме эффектов пола и образования. Исходная модель с эффектами взаимодействия будет иметь следующий вид:

Здесь и т. д. Включаются эффекты как первого порядка (, , и , ), так и эффекты второго порядка ( и т. д.).

В подобной модели будет сразу несколько линейных зависимостей. Самый простой способ избавиться от линейных зависимостей – исключить по одной категории из каждой классификации и удалить из модели соответствующие фиктивные переменные. Например, можем исключить и :

При использовании эффектов взаимодействия следует следить, чтобы наблюдений для каждой группы, соответствующей сочетанию признаков, ( и т. д.), было не очень мало. Если наблюдений для одного из сочетаний не найдется, то модель не получится оценить.

В рассматриваемой модели можем проверить гипотезу об отсутствии эффектов взаимодействий:

Также можно проверить гипотезу об отсутствии влияния одной из качественных переменных. Отсутствие влияния , :

Отсутствие влияния , , :

## Контрольные вопросы

1. Для чего нужны в регрессии фиктивные переменные?
2. Какие из перечисленных факторов учитываются в регрессии с помощью фиктивных переменных: а) профессия; б) курс доллара; в) численность населения;  
   г) размер среднемесячных потребительских расходов?
3. Запишите исходную форму регрессии с фиктивными переменными, в случае поквартальных наблюдений.
4. Почему невозможна оценка параметров МНК непосредственно по исходной форме регрессии с фиктивными переменными?
5. Основные практические способы включения фиктивных переменных в регрессию в случае поквартальных наблюдений. Как проверяется значимость фактора? Интерпретация оценки коэффициентов в зависимости от способа включения этих факторов.
6. Запишите переменную совместного влияния (взаимодействия) факторов пола (два уровня: мужской и женский) и образования (четыре уровня: нет образования, образование начальное, среднее, высшее).
7. В уравнение регрессии для доходов населения вводятся два качественных фактора: пола (два уровня: мужской и женский) и образования (четыре уровня: нет образования, образование начальное, среднее, высшее). Сколько фиктивных переменных (с учетом взаимодействия факторов) в исходной и преобразованной (после устранения линейных зависимостей) форме уравнения?

## Экзаменационные вопросы

1. Качественные факторы в регрессии. Фиктивные переменные. Способы включения фиктивных переменных в регрессию. Интерпретация коэффициентов при фиктивных переменных.
2. Способы включения фиктивных переменных в регрессию. Взаимосвязь оценок коэффициентов для разных способов.
3. Качественные факторы в регрессии. Фиктивные переменные. Способы включения совместного влияния качественных факторов (эффекты второго порядка) в регрессию. Пример.

## Литература

1. Магнус Я. Р., Катышев П. К., Пересецкий А. А. Эконометрика. Начальный курс:– Учебник. – 6-е изд., перераб. и доп. – М.: Дело, 2004. [стр. 112-118]
2. Суслов В. И., Ибрагимов Н. М., Талышева Л. П., Цыплаков А. А. *Эконометрия*. – Новосибирск: СО РАН, 2005. [*стр. 289-295]*
3. [Картаев Ф. С.](https://litgid.com/book_author/18780834/)*Введение в эконометрику. Учебник :*– Москва: Проспект, 2019. [*стр. 110-116]*

# Лекция: – Прогнозирование в модели линейной регрессии

## Прогнозирование в классической модели линейной регрессии

Одно из направлений использования регрессионного анализа – это прогнозирование. Пусть мы получили оценку МНК , используя имеющиеся наблюдения

Если известны значения регрессоров  
для некоторого нового наблюдения с номером , где , то можно использовать полученную оценку для построения **прогноза** величины . Для этого требуется просто подставить новые данные в оцененное уравнение регрессии:

Какими теоретическими свойствами будет обладать подобный прогноз? Свойства эти связаны с теоретическими характеристиками **ошибки прогноза** , равной разности фактической величины и прогнозной:

Для анализа свойств прогноза мы должны сделать ряд предположений. Фактически, это те предположения (A0)–(A4), которые мы делали ранее, но соответствующим образом модифицированные для расширенного набора наблюдений Как мы видели ранее, если для наблюдений выполнены предположения (A0)–(A4), то для оценок имеем и . Относительно нового наблюдения будем предполагать следующее.

(A0) Новое значение получается по той же формуле регрессии, что и предыдущие , :

(A1) Ошибка нового наблюдения в среднем равна нулю:

(A2) Значения регрессоров в новой точке являются детерминированными.

(A3) Дисперсия ошибки нового наблюдения такая же, как и у предыдущих наблюдений:

(A4) Ошибка нового наблюдения не коррелирует с ошибками предыдущих наблюдений:

Можем записать ошибку прогноза в виде

Данная ошибка состоит из двух частей. Это ошибка регрессии для нового наблюдения и добавка , которая связана с тем, что мы не знаем истинные коэффициенты и берем вместо них оценку .

Первое простое свойство состоит в том, что *прогноз является несмещенным*, т. е. ошибка прогноза в среднем равна нулю:  
Здесь мы использовали предположения (A0), (A1) и (A2).

Далее, дисперсия ошибки прогноза равна

При выводе этой формулы мы использовали предположения (A0)–(A4). В частности, два слагаемых ошибки прогноза некоррелируют по предположению (A4), поэтому дисперсия равна сумме их дисперсий.

**Точность прогноза** можно измерить **средним квадратом ошибки прогноза** (англ. *mean squared error*, MSE). Этот показатель из-за несмещенности совпадет с дисперсией ошибки прогноза:  
Чем меньше этот показатель, тем более точным является прогноз. Как видим, он, как и ошибка прогноза, состоит из двух слагаемых: дисперсии ошибки нового наблюдения и дополнительного слагаемого, связанного с точностью оценки коэффициентов .

Без доказательства отметим, что в случае парной регрессии () можно рассчитать по следующей упрощенной формуле:  
где – выборочное среднее, а – выборочная дисперсия объясняющей переменной по наблюдениям . Формула показывает, что чем дальше новая точка от среднего , тем менее точен прогноз.

Дисперсию ошибки прогноза мы точно не знаем, поскольку не знаем , но можем получить оценку на основе остаточной дисперсии:

Точечный прогноз и его стандартную ошибку  
можно получить в любой компьютерной программе, которая считает регрессии, использовав вспомогательную регрессию со следующими переменными:

Здесь зависимая переменная дополнена нулем, матрица регрессоров дополнена строкой , а также добавлен регрессор, равный для нового наблюдения и 0 в остальных случаях. Коэффициент при добавленной переменной будет равен прогнозу , а стандартная ошибка этого коэффициента – стандартной ошибке прогноза . (Доказательство в приложении.)

Мы рассмотрели прогноз в виде одного числа, то есть **точечный прогноз**. Поскольку любой прогноз неточен, более полную информацию дают интервальные прогнозы, а еще более полную информацию – плотностные прогнозы, то есть прогнозы в виде функции распределения. Для построения таких прогнозов нам следует дополнительно принять предположение о нормальности ошибки регрессии.

(A5) Ошибка нового наблюдения имеет нормальное распределение:  
а в целом вектор всех ошибок регрессии имеет многомерное нормальное распределение:

В условиях предположений (A0)–(A5) ошибка прогноза имеет нормальное распределение:

Поскольку истинное значение неизвестно, мы не можем использовать этот факт для построения плотностного прогноза. По аналогии с выводом распределения *t*-статистики мы можем здесь вывести, что величина имеет распределение Стьюдента:  
где – стандартная ошибка прогноза. Это свойство позволяет предложить следующее прогнозное распределение (**плотностной прогноз**):  
Данную запись следует понимать следующим образом: для дается прогноз в виде распределения Стьюдента со степенями свободы , масштабированного на величину стандартной ошибки прогноза и сдвинутого на величину точечного прогноза .

Из плотностного прогноза можно получить другие виды прогнозов. В частности, **интервальный прогноз** можно получить, взяв соответствующие квантили прогнозного распределения:  
Здесь – это двусторонняя квантиль распределения , соответствующая некоторой достаточно большой вероятности средней части распределения (50%, 75%, 90% и т. п.).

## Приложение. Вывод вспомогательной регрессии для получения прогноза

Рассмотрим следующую вспомогательную регрессию:  
или

Применим к этой регрессии МНК. Сумма квадратов остатков как функция от и равна

Удобно минимизировать эту функцию в два этапа – сначала найти минимум по при фиксированном , а затем подставить полученное значение и минимизировать по . Поскольку первое слагаемое не зависит от , следует минимизировать по второе слагаемое . Минимум по достигается при .

Далее после подстановки получим  
то есть сумму квадратов остатков из регрессии по , откуда  
Это означает, что – коэффициент при добавочной переменной будет равен прогнозу.

Осталось вывести стандартную ошибку для . Мы видим, что во вспомогательной регрессии остатки будут равны  
где – остатки из исходной регрессии по . Вспомогательная регрессия будет иметь  
степеней свободы, и несмещенная остаточная дисперсия во вспомогательной регрессии равна  
как и в исходной.

Обозначим  
Нам требуется найти правый нижний элемент матрицы . Здесь

По правилам обращения блочной матрицы (см. Математическое приложение) получим

(Можно проверить, что при умножении на указанную матрицу получается .)

Таким образом, правый нижний элемент матрицы равен  
Стандартная ошибка коэффициента во вспомогательной регрессии – это квадратный корень из этой величины:  
Это формула стандартной ошибки прогноза .

## Контрольные вопросы

1. Вывести формулу дисперсии ошибки прогноза.
2. Вывести формулу для оценки стандартной ошибки прогноза при
3. Покажите, что точечный прогноз не смещен относительно ожидаемого значения, т. е. ошибка прогноза в среднем равна нулю.
4. Докажите, что дисперсию , которая минимальна в классе линейных оценок .
5. Считается вспомогательная регрессию : зависимая переменная , матрица регрессоров Докажите, что при регрессоре коэффициент будет равен прогнозу , а стандартная ошибка этого коэффициента – стандартной ошибке прогноза

## Экзаменационные вопросы

1. Прогнозирование по модели регрессии: формула точечного прогноза, ошибка прогноза, свойства точечного прогноза (несмещенность, средний квадрат и дисперсия ошибки)
2. Прогнозирование по модели регрессии: нормальность распределения ошибок и интервальный прогноз. Зависимость ширины интервала от значения факторов.
3. Прогнозирование по модели регрессии. Вывод вспомогательной регрессии для получения прогноза

## Литература

1. Вербик М. Путеводитель по современной эконометрике. – М.: Научная книга,. 2008. [стр. 86-87]
2. Магнус Я. Р., Катышев П. К., Пересецкий А. А. Эконометрика. Начальный курс: – Учебник. – 6-е изд., перераб. и доп. – М.: Дело, 2004. [стр. 204-211]
3. Суслов В. И., Ибрагимов Н. М., Талышева Л. П., Цыплаков А. А. Эконометрия.– Новосибирск: СО РАН, 2005. [стр. 244-247, 569-572, 578-582]
4. Картаев Ф. С. Введение в эконометрику. Учебник : – Москва: Проспект, 2019. [стр. 58-61]

# Лекция: – Диагностика регрессии: введение

**Диагностика регрессии** – это статистические процедуры, которые направлены на то, чтобы проверить выполнение предположений, лежащих в основе модели регрессии. Также диагностика регрессии может проводиться для того, чтобы изучить влияние отдельных наблюдений и их групп на результаты. Более широко к диагностике регрессии можно отнести процедуры статистической проверки исходных данных на наличие ошибок.

## Изучение исходных данных

Диагностические процедуры могут быть как формальными, т. е. основанными на определенной модели и соответствующей теории, так и неформальными. Среди неформальных методов выделяются графические методы.

Хорошей практикой является проведение предварительного исследования исходных данных. Для отдельных переменных это, в частности:

- визуальный осмотр значений переменных (рис. 28),

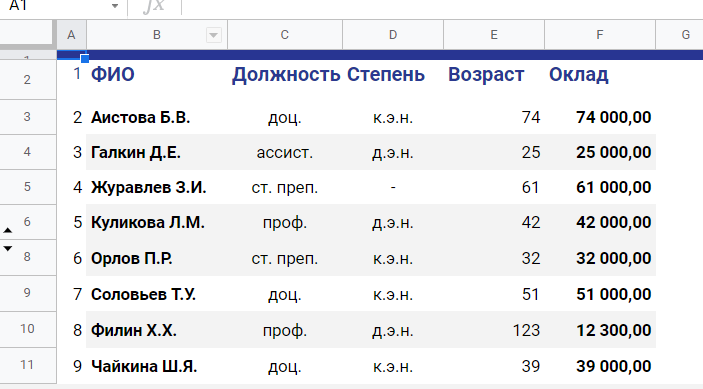


Рисунок 28. Таблица данных

- графики по номеру наблюдения (для временных рядов – по времени, рис. 29),

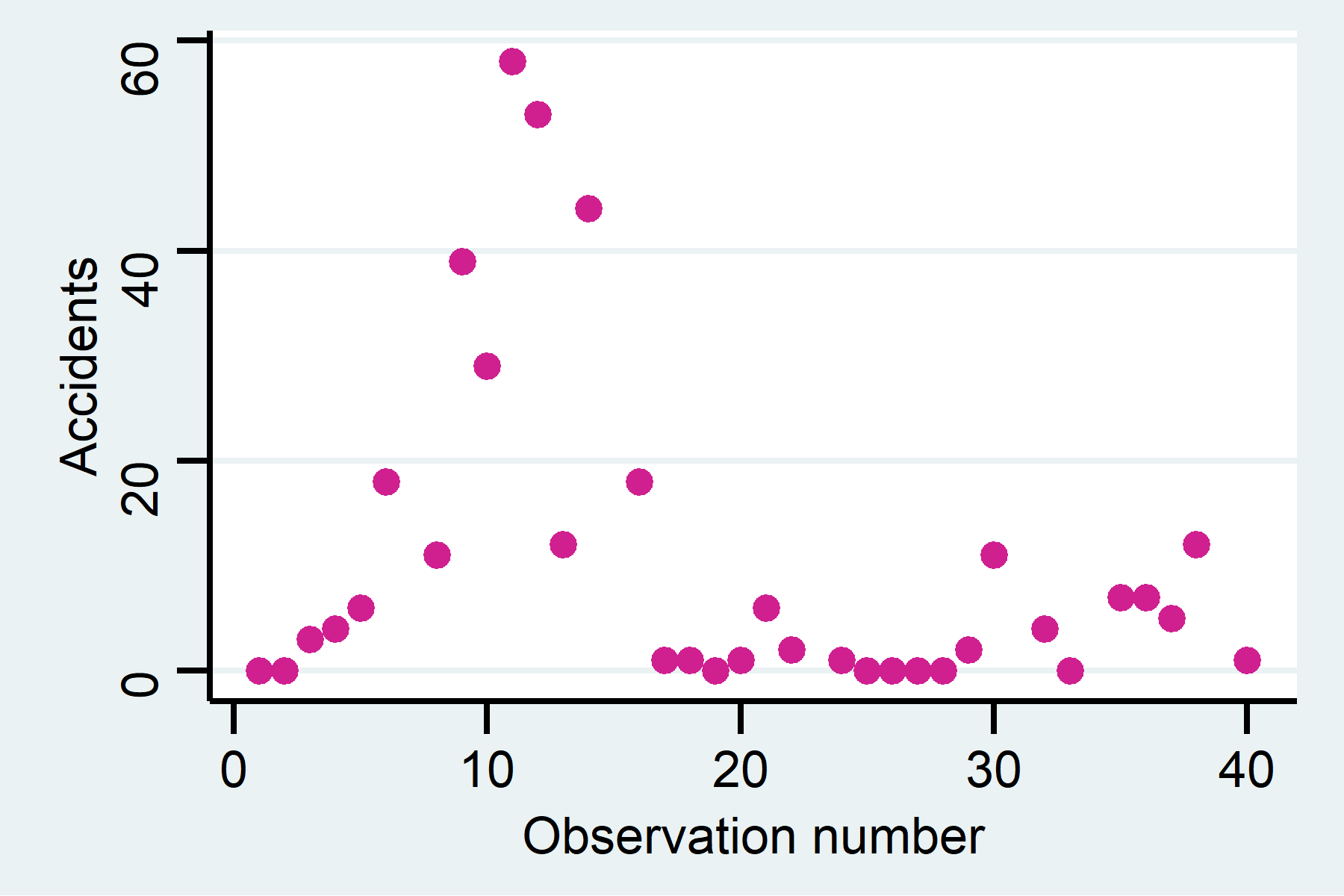
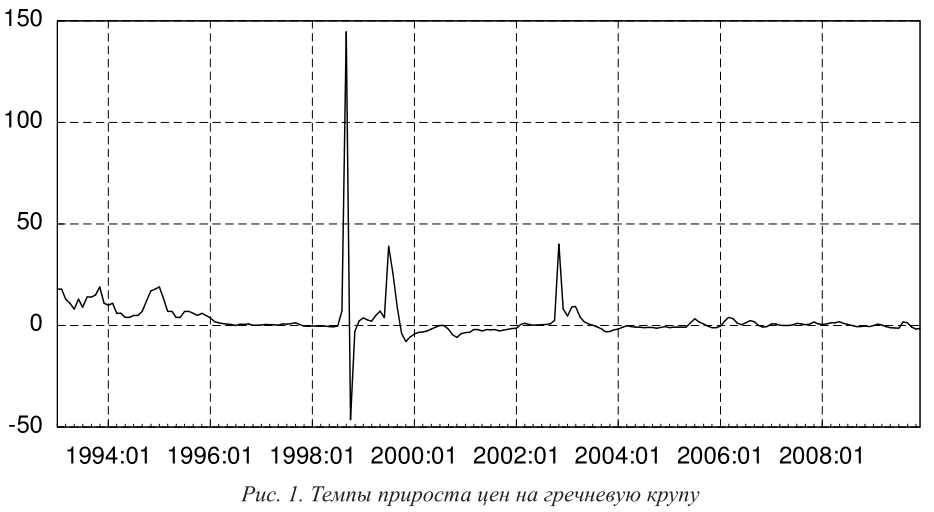
 

Рисунок 29. Графики наблюдений

- графики плотности, например, гистограммы (рис. 30),

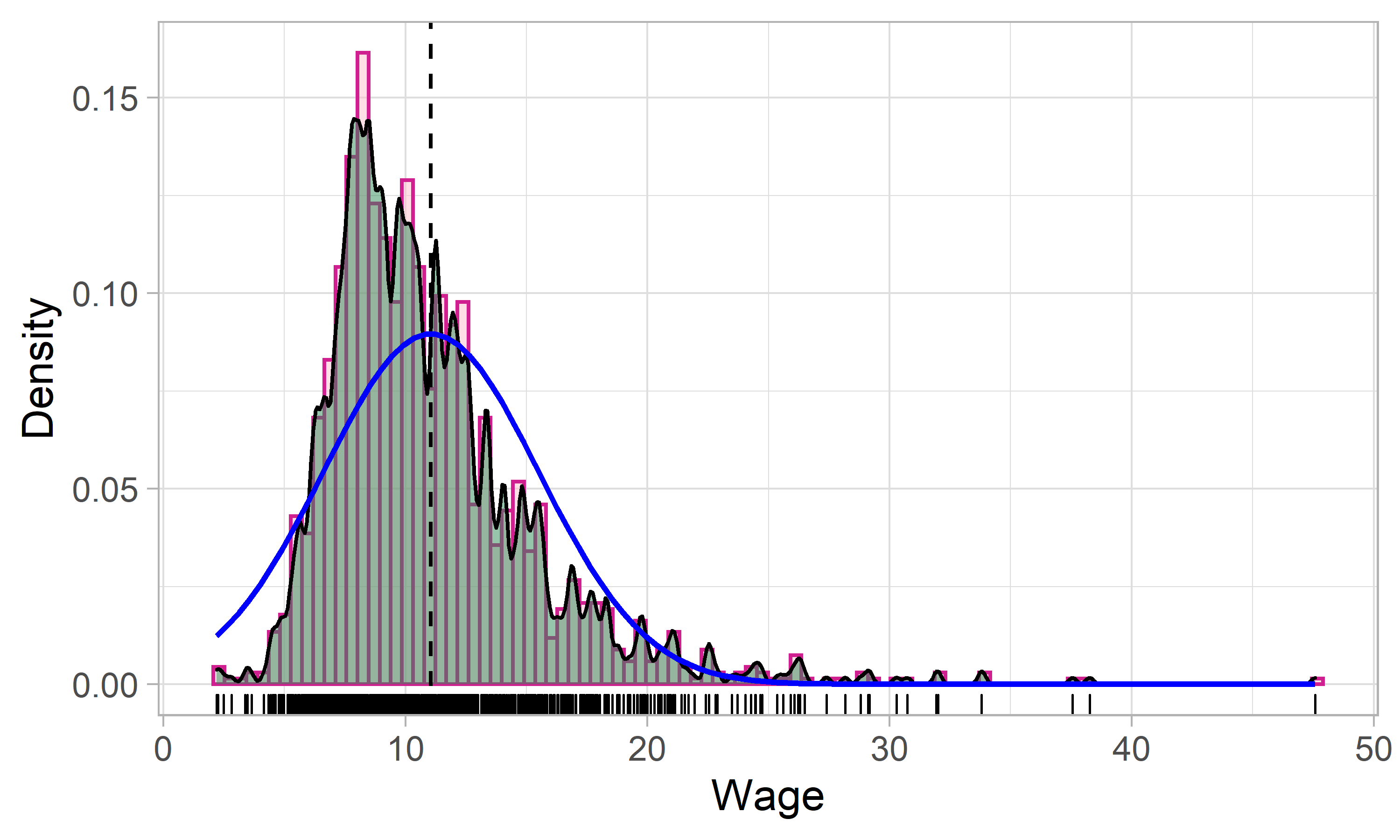


Рисунок 30. Оценки плотности

***Descriptive Statistics***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Obs** | **Mean** | **Std.Dev.** | **Min** | **Max** |
| capital | 566 | 11.5154 | 80.5188 | 0.0022 | 1786.8992 |
| labour | 559 | 201.0808 | 611.9959 | 1.0000 | 10661.0000 |
| output | 567 | 14.7192 | 62.7295 | 0.0263 | 1279.3717 |
| wage | 561 | 38.6329 | 14.4432 | 11.7337 | 188.4251 |

Рисунок 31. Таблица с описательными статистиками для набора данных

- таблицы с описательными статистиками (рис. 31).

Среди прочего следует обращать внимание на наличие в переменных выбросов (резко выделяющихся наблюдений) и/или неправдоподобных значений. Примеры неправдоподобных значений: нулевые или отрицательные цены, возраст больше 100 лет и т. п. Такие значения могут свидетельствовать о наличии ошибок в исходных данных.

После изучения отдельных переменных можно изучить совместные характеристики для пар переменных, например:

- коэффициенты парной корреляции (рис. 32),

***Matrix of correlations***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | (1) | (2) | (3) | (4) | (5) |
| (1) Price | 1.0000 |  |  |  |  |
| (2) MPG.city | -0.5946 | 1.0000 |  |  |  |
| (3) Horsepower | 0.7882 | -0.6726 | 1.0000 |  |  |
| (4) Fuel.tank | 0.6195 | -0.8131 | 0.7118 | 1.0000 |  |
| (5) Weight | 0.6472 | -0.8431 | 0.7388 | 0.8940 | 1.0000 |

Рисунок 32. Таблица парных корреляций

- точечные диаграммы для пар переменных (рис. 33).

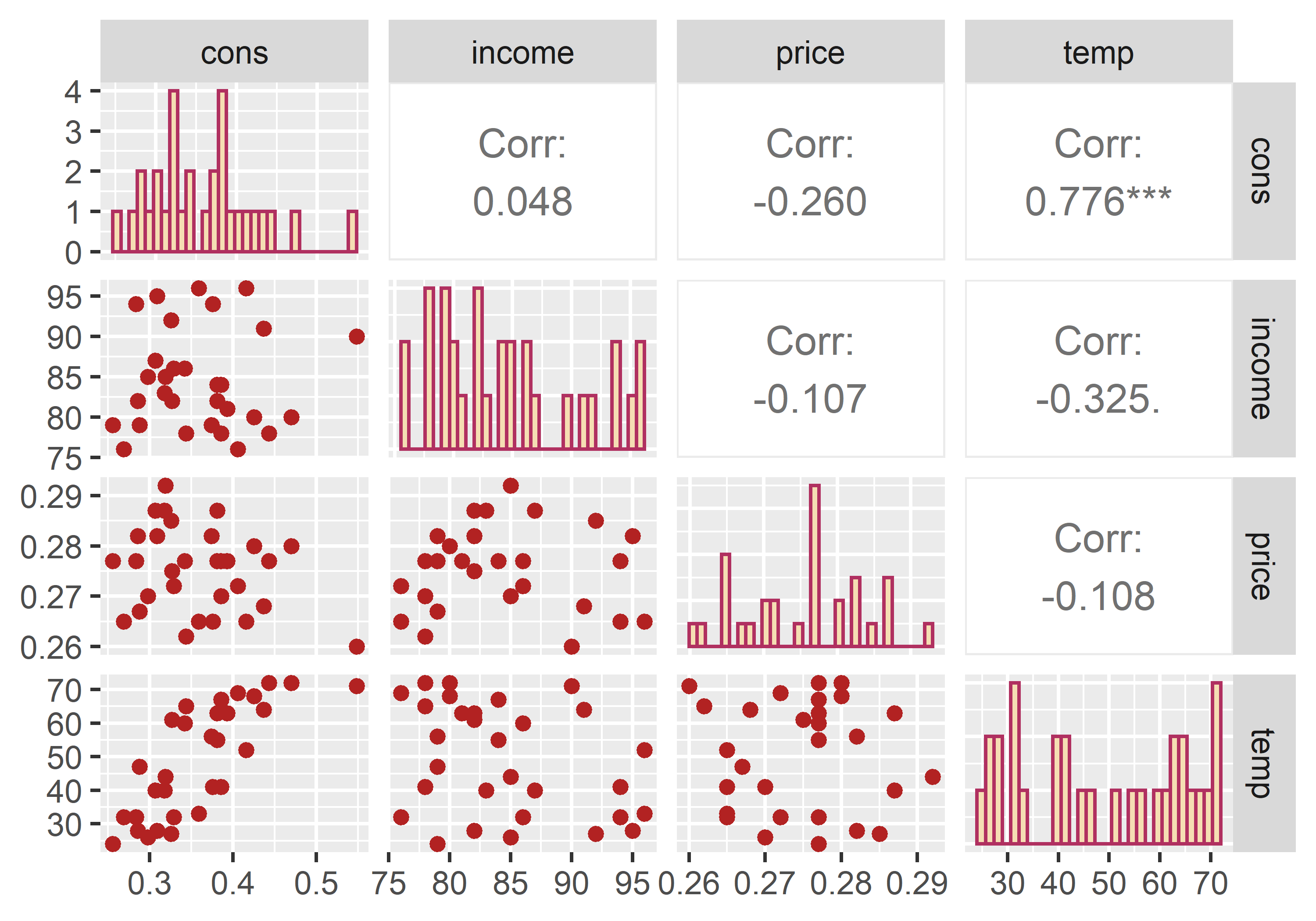


Рисунок 33. Точечные диаграммы и гистограммы

Изучение подобных графиков и показателей может как указать на ошибки и проблемы в данных, так и подсказать, какой должна быть модель регрессии, если она еще не сформулирована.

## Простейший графический анализ регрессии

После оценивания парной регрессии желательно посмотреть на точечную диаграмму данных с наложенной на нее линией регрессии (рис. 34).



Рисунок 34. Точечная диаграмма с линией регрессии

У парных регрессий, результаты которых выглядят как совершенно одинаковые с точки зрения цифр, могут быть совершенно разные «картинки».

Можно рассмотреть искусственно созданный пример, в котором для 7 разных наборов данных по 50 наблюдений каждый получаются регрессии с одинаковыми результатами. В этом примере две переменные имеют одни и те же средние и дисперсии, корреляции одинаковые, таблицы оцененных регрессий идентичные (рис. 36), но расположение точек на каждой из диаграмм имеет свои особенности (рис. 35). Только один из наборов данных выглядит «нормально». Для остальных наборов диаграммы указывают на возможное наличие каких-то проблем с моделью регрессии. В некоторых случаях есть основание подозревать, что нарушены предположения модели регрессии.

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Рисунок 35. Одинаковые линии регрессии для 7 разных наборов точек

Estimate Std. Error t value Pr(>|t|)   
(Intercept) 1.6000 0.5654 2.83 0.00678 \*\*  
x 2.2500 0.1573 14.30 < 2e-16 \*\*\*  
Residual standard error: 2.202 on 48 degrees of freedom  
Multiple R-squared: 0.81, Adjusted R-squared: 0.806  
F-statistic: 204.6 on 1 and 48 DF, p-value: < 2.2e-16

Рисунок 36. Одинаковая выдача программы R для регрессий по 7 разным наборам точек

Такие графики надо еще уметь интерпретировать. Некоторые возможные проблемы, которые можно диагностировать с помощью подобных графиков, будут рассмотрены в последующих темах.

Для множественной регрессии аналогом может служить точечная диаграмма фактических значений зависимой переменной по расчетным значениям . Для такой диаграммы аналогом линии регрессии может служить линия с наклоном 1, проходящая через начало координат.



Рисунок 37. Диаграмма расчетных и фактических значений

После оценивания любой регрессии важно, в числе прочего, провести разнообразные диагностические процедуры для остатков , которые в определенном смысле являются оценками ненаблюдаемых ошибок . Тут используются различные графики (рис. 38).

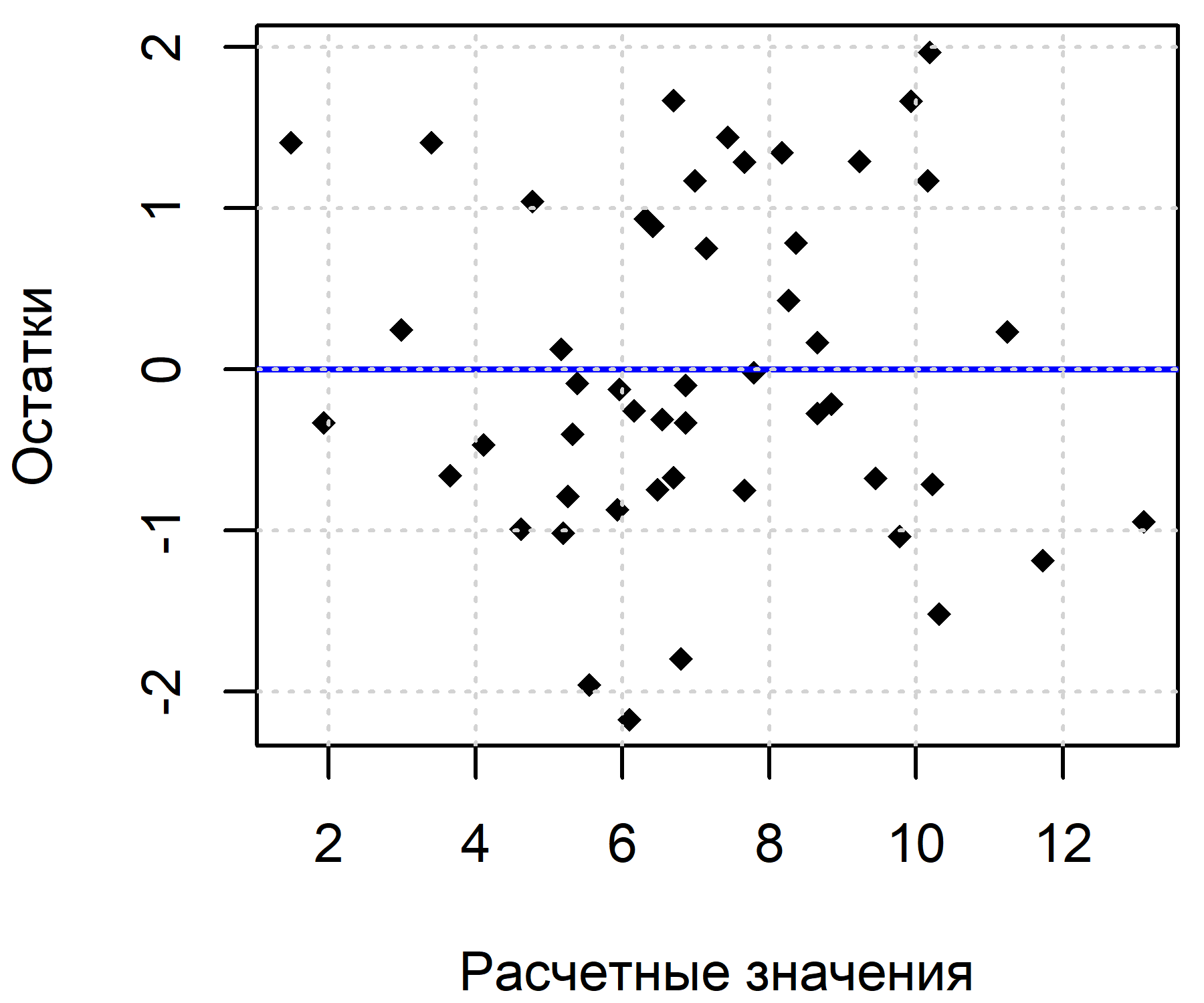
        

Рисунок 38. Простейшие графики для остатков регрессии

## Принципы проведения диагностических тестов

В целом в литературе предложено множество различных формальных и неформальных **диагностических тестов**, которые служат для проверки соответствия наблюдаемых данных предположениям модели. Практически все подобные тесты построены по одной и той же схеме:

- Идет проверка в сторону усложнения модели, то есть от частного к общему.

- В качестве нулевой гипотезы принимается предположение о корректности исходной модели, т. е. об отсутствии нарушения той или иной лежащей в основе модели гипотезы. Например, в тестах функциональной формы в качестве нулевой гипотезы принимается предположение, что функциональная форма модели верна (в частности, что модель линейна).

- Нулевая гипотеза о корректности проверяемой модели отклоняется, если *p*-значение тестовой статистики маленькое.

- Если нулевая гипотеза о корректности проверяемой модели отклоняется, то это может говорить о том, что следует каким-то образом модифицировать модель. Однако сам характер теста дает только самую общую подсказку относительно того, как это сделать. (Например, если проведен тест на линейность зависимости и найдена нелинейность, то видимо, следует как-то учесть эту нелинейность. Однако, как конкретно это делать, не всегда понятно.) В целом моделирование – это в значительной степени искусство, которому надо учиться, и в котором сложно избежать ошибок.

При проведении диагностических тестов удобно пользоваться различного рода **вспомогательными регрессиями**. Это такие регрессии, которые носят чисто технический характер, и нужны только для расчета тестовой статистики (и, возможно, ее *p*-значения). Удобство вспомогательных регрессий заключается в том, что если вы используете регрессионный анализ, то у вас наверняка есть техническая возможность оценить какую-то дополнительную регрессию, то есть вычислительные средства у вас, как говорится, под рукой.

Часто (но не всегда) вспомогательная регрессия выглядит как исходная проверяемая регрессия , в которую *добавляются* некоторые искусственно созданные регрессоры :

С «технической» точки зрения нулевая гипотеза тогда заключается в том, что *коэффициенты при добавленных переменных одновременно равны нулю*:  
Содержательно, как уже говорилось, нулевая гипотеза заключается в том, что исходная регрессия корректна. Смысл заключается в том, что если регрессия была корректной, то добавленные переменные в ней не нужны и коэффициенты при них должны быть нулевыми.

Например, для проверки линейности можно добавлять в регрессию нелинейные члены и проводить тесты на их значимость. В дальнейших темах мы еще рассмотрим различные подобные регрессии.

Пусть – количество наблюдений, – количество регрессоров в проверяемой регрессии, – сумма квадратов остатков в проверяемой регрессии, – оценки коэффициентов при добавленных переменных, – количество добавленных переменных, – сумма квадратов остатков в расширенной регрессии. Тогда для проверки гипотезы можно использовать стандартную *F*-статистику с соответствующими подстановками и переобозначениями:

где – блок ковариационной матрицы расширенной регрессии, относящийся к . Регрессия имеет степеней свободы и остаточную дисперсию  
Если проверяемая модель корректна, то *F*-статистика имеет распределение Фишера с и степенями свободы. Если *F*-статистика превышает выбранную границу, то нулевая гипотеза отклоняется. В отсутствие автокорреляции и гетероскедастичности можно использовать удобную альтернативную формулу для *F*-статистики:

При проведении диагностических тестов важно учитывать, что статистическая значимость того или иного эффекта – это не то же самое, что его практическая значимость с точки зрения влияния на оценки и статистики. Например, в регрессии может иметь место статистически значимая нелинейность, при том что с практической точки зрения соответствующая «кривизна» линии пренебрежимо мала. Обычно статистическая значимость несущественных эффектов проявляется в регрессиях с большим числом наблюдений.

## Контрольные вопросы

1. Какие проблемы побуждают проводить предварительные исследования исходных данных?
2. Какие выводы приводят к необходимости проводить анализ точечной диаграммы фактических значений зависимой переменной по расчетным значениям? Примеры.
3. Основная схема диагностических тестов для проверки соответствия наблюдаемых данных предположениям модели.
4. Зачем нужнавспомогательная регрессия при проведении диагностических тестов?

## Экзаменационные вопросы

1. Диагностика регрессии – формальные методы. Основная схема диагностических тестов
2. Диагностика регрессии – неформальные методы. Основная схема диагностических тестов

## Литература

1. Вербик М. Путеводитель по современной эконометрике. – М.: Научная книга, 2008. [стр. 121-130]
2. Картаев Ф. С. Введение в эконометрику. Учебник : – Москва: Проспект, 2019. [стр. 116-134]
3. Носко В. П. Эконометрика для начинающих Учебник : – Москва: 2000. [стр. 180-194]

# Лекция: – Функциональная форма модели

## Регрессия как линейная модель

### Что такое «линейная регрессия»?

Ранее мы записали уравнение линейной регрессии в виде  
В этой записи имеет место линейность как по коэффициентам , так и по регрессорам . Что же мы имеем в виду, когда говорим «линейная регрессия»?

С точки зрения получения оценок коэффициентов по формуле МНК важна *линейность по коэффициентам (параметрам)*. Поэтому, когда без уточнения говорят о линейности регрессии, то, скорее всего, речь идет именно о линейности по коэффициентам. Если модель линейна по коэффициентам, то ее всегда можно сделать линейной по переменным заменой переменных.

Поясним сказанное на примерах. Можно рассмотреть парную регрессию, связывающую и с объясняющей переменной , но такую, что обе переменные входят в нее в виде логарифмов:

Эта модель станет обычной линейной регрессией, если произвести замену , :

В качестве примера множественной регрессии рассмотрим производственную функцию Кобба—Дугласа:

в которой – выпуск, – капитал, – труд. Если данную формулу прологарифмировать, сделать замену и добавить в уравнение член ошибки, то получится уравнение регрессии, которое линейно по коэффициентам:

Замена переменных , , дает уравнение, которое линейно по переменным:

Коэффициенты , и тут можно находить через обычный МНК.

Еще один пример – квадратичная зависимость между переменными:  
Здесь можно обозначить , ,

Если модель регрессии **нелинейна по параметрам**, то иногда такую модель можно свести к линейной, применив замену переменных к параметрам. Например, уравнение

можно превратить в

,

обозначив , и .

В то же время, следующая модель регрессии:  
существенно нелинейна по параметрам и не сводится к линейной регрессии. (Здесь в правой части стоит так называемое преобразование Бокса–Кокса с параметром .)

Для применения метода наименьших квадратов важно, чтобы ошибка была **аддитивной**, то есть, чтобы зависимая переменная являлась суммой своего математического ожидания и ошибки. Об этом следует помнить, производя преобразования модели. Например, модель нельзя преобразовать в линейную по параметрам с аддитивной ошибкой. Аналогичную модель с мультипликативной ошибкой , где , и положительны, можно преобразовать к виду   
или   
где , , , . Здесь ошибка должна принимать положительные значения. Например, для годится логнормальное распределение – тогда ошибка в преобразованной регрессии оказывается нормально распределенной. Если мы возьмем ту же модель с мультипликативной ошибкой, но с предположением, что , то ее не получится свести к линейной регрессии (и она будет моделировать не математическое ожидание переменной , а ее дисперсию).

Если с помощью преобразований не получается получить линейную по параметрам модель с аддитивной ошибкой, то следует использовать какие-то нелинейные методы. В частности, для оценивания нелинейной по параметрам регрессии с аддитивной ошибкой можно использовать нелинейный МНК. Для оценивания более сложных моделей чаще всего используют метод максимального правдоподобия или (обобщенный) метод моментов.

### Линейные преобразования модели

Когда модель линейна, разного рода линейные преобразования не приводят к принципиально иной модели. Важно только, чтобы зависимая переменная не повлияла на правую часть регрессии. Можно рассмотреть два вида преобразований:

1) Замена переменных в параметрах регрессии. Вместо коэффициентов мы используем коэффициенты такие что , где – невырожденная квадратная матрица. Получаем регрессию , где , .

2) Линейное преобразование переменных регрессии. Вместо и мы строим регрессию для преобразованных данных и :

где , , – невырожденная квадратная матрица, – вектор-столбец, – скаляр.

Для нас здесь важно, что если используется простая линейная замена переменных, то коэффициенты и статистики для преобразованной модели можно однозначно пересчитать в коэффициенты и статистики исходной модели и наоборот. Подробнее об этом в Приложении к главе.

Приведем примеры простых преобразований (с некоторыми из них мы уже встречались в предыдущих темах):

- добавить константу к зависимой переменной (или вычесть),

- добавить константу к объясняющей переменной (или вычесть, например, произвести центрирование),

- поменять единицы измерения зависимой переменной (скажем, рубли вместо тысяч рублей),

- поменять единицы измерения объясняющей переменной

- вычесть объясняющую переменную из левой и правой части регрессии.

Как меняются те или иные статистики (оценки коэффициентов, t-статистики, F-статистики, и т. д.) в каждом из этих случаев мы не будем здесь разбирать. Но это, как правило, не очень сложно понять, если выписать в явном виде формулы, связывающие исходную и преобразованную регрессии и вспомнить определения статистик.

## Разные функциональные формы и их интерпретация

### Некоторые простые функциональные формы

Рассмотрим некоторые простые функциональные формы модели регрессии, которые чаще всего встречаются на практике.

Мы уже рассмотрели пример парной регрессии c зависимой переменной и объясняющей переменной :  
Подобного рода зависимости достаточно часто используют, если переменные принимают положительные значения, особенно если значения переменной могут различаться в несколько раз для разных наблюдений. (Например, может быть ценой квартиры, а – ее площадью; см. такой пример ниже.) Данную модель называют линейной в логарифмах или **логлинейной**. Можно также использовать удобное обозначение **log-log**. (Соответственно, обычная линейная регрессия – это **lin-lin**).

Кроме линейных в логарифмах существуют полулогарифмические модели. Для парной регрессии это

и

(**lin-log** и **log-lin** соответственно). Эти модели, как и логлинейная, превращаются в линейные заменой переменных.

По понятным причинам «кандидатами» на логарифмирование в первую очередь служат те переменные, которые заведомо могут принимать только положительные значения. Один из их признаков – это то, что, как правило, интересуются относительными приростами таких переменных (в процентах), а не абсолютными приростами (например, в рублях). Примеры из экономики: объем производства товара, объем продаж, цена, потребительские доходы и расходы, фондовый индекс, ВВП.

Одной из причин логарифмирования зависимой переменной является то, что прогноз положительной переменной не обязательно будет положительным, если переменная входит в регрессию в исходном виде. Кроме того, при положительности зависимой переменной ошибка регрессии, строго говоря, не может подчиняться нормальному распределению, поскольку нормально распределенная случайная величина может принимать любые действительные значения, в том числе сколь угодно малые отрицательные. Хотя на практике встречаются ситуации, когда нормальность распределения ошибки можно принять в качестве неплохой аппроксимации в регрессии с положительной зависимой переменной, но это бывает далеко не всегда. Также логарифмирование помогает ослабить гетероскедастичность. Об использовании логарифмирования для решения указанных проблем (отсутствие нормальности, сильная гетероскедастичность) речь пойдет в соответствующих главах.

Если положительные переменные колеблются не очень сильно, лишь на несколько процентов, то с точки зрения моделирования не принципиально, брать их в логарифмах или в исходном виде. Это следует из следующего приближения, известного из математического анализа ():  
Приближение применимо, когда величина мала.

Применяя это и аналогичные ему линейные приближения мы можем показать, что рассмотренные выше функциональные формы похожи друг на друга при не очень больших процентных отклонениях.

Рассмотрим только приближенное соотношение между линейной в логарифмах (log-log) и линейной (lin-lin) регрессиями. Перепишем log-log регрессию  
в видеЕсли ошибка мала, то и

Если – некое «типичное» значение переменой и процентные отклонения от не очень большие, то  
Отсюда видим, что линейная в логарифмах регрессия похожа на lin-lin регрессию  
где , и .

Заметим, что если в log-log регрессии ошибка примерно гомоскедастична, то в lin-lin регрессии мы будем иметь дело с гетероскедастичностью, поскольку среднеквадратическое отклонение ошибки будет примерно пропорционально . Такая ситуация достаточно часто наблюдается на практике.

В регрессиях можно также использовать переменные, представляющие собой функции от двух или более исходных переменных. В частности, довольно часто используют относительные показатели. Если наблюдения относятся к странам или регионам, то многие показатели, которые примерно пропорциональны населению, берут в расчете на душу населения. Например, если  – потребление вина в стране,  – ВВП страны, а  – численность населения, то можно оценивать регрессию вида  
Это только один пример использования относительных показателей. Возможны разные другие варианты, например, в правой части регрессии , в левой . Можно сочетать использование относительных показателей с логарифмированием:  
и т. д.

#### Пример: Зависимость цены квартиры от площади

В этом примере используются данные из объявлений о продаже квартир в г. Бердск (410 наблюдений). *Price* – цена квартиры в тыс. руб., *area* – общая площадь квартиры в м2.

Оценивание регрессии в исходных переменных дает уравнение

а в логарифмах  
Как эти две регрессии выгладят на графиках, можно увидеть на рис. 39. Забегая вперед, скажем, что на левом графике (для регрессии lin-lin) заметна довольно сильная гетероскедастичность.

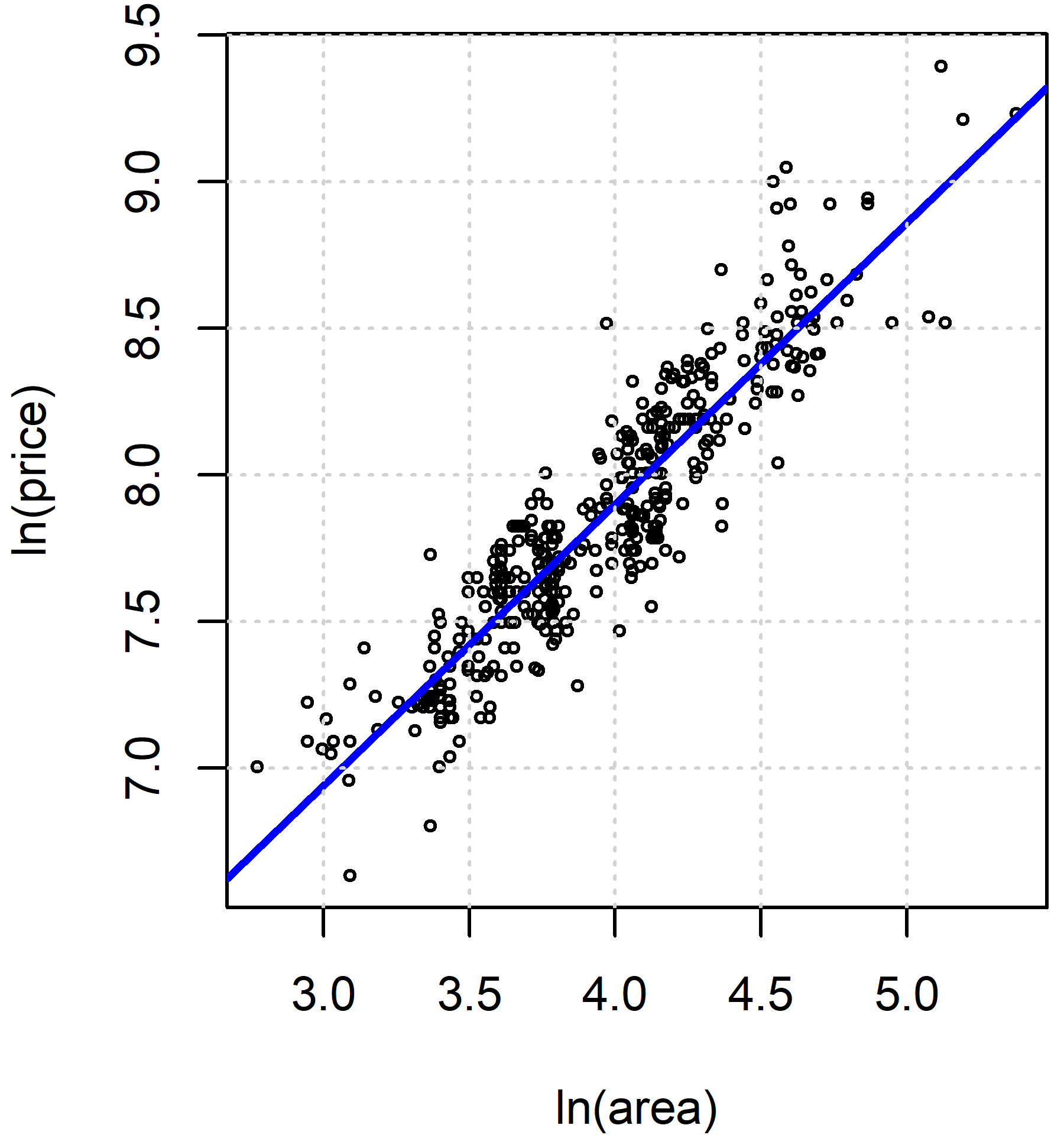
 

Рисунок 39. Две регрессии цены квартиры от площади

Еще один вариант постановки модели – это зависимость цены за квадратный метр общей площади от общей площади (рис. 40):

Аналогичная регрессия в логарифмах  
по сути дела совпадает с регрессией логарифма цены от логарифма площади, являясь ее линейным преобразованием. Коэффициент наклона на 1 меньше (). Коэффициенты детерминации в двух регрессиях разные из-за того, что зависимые переменные разные.

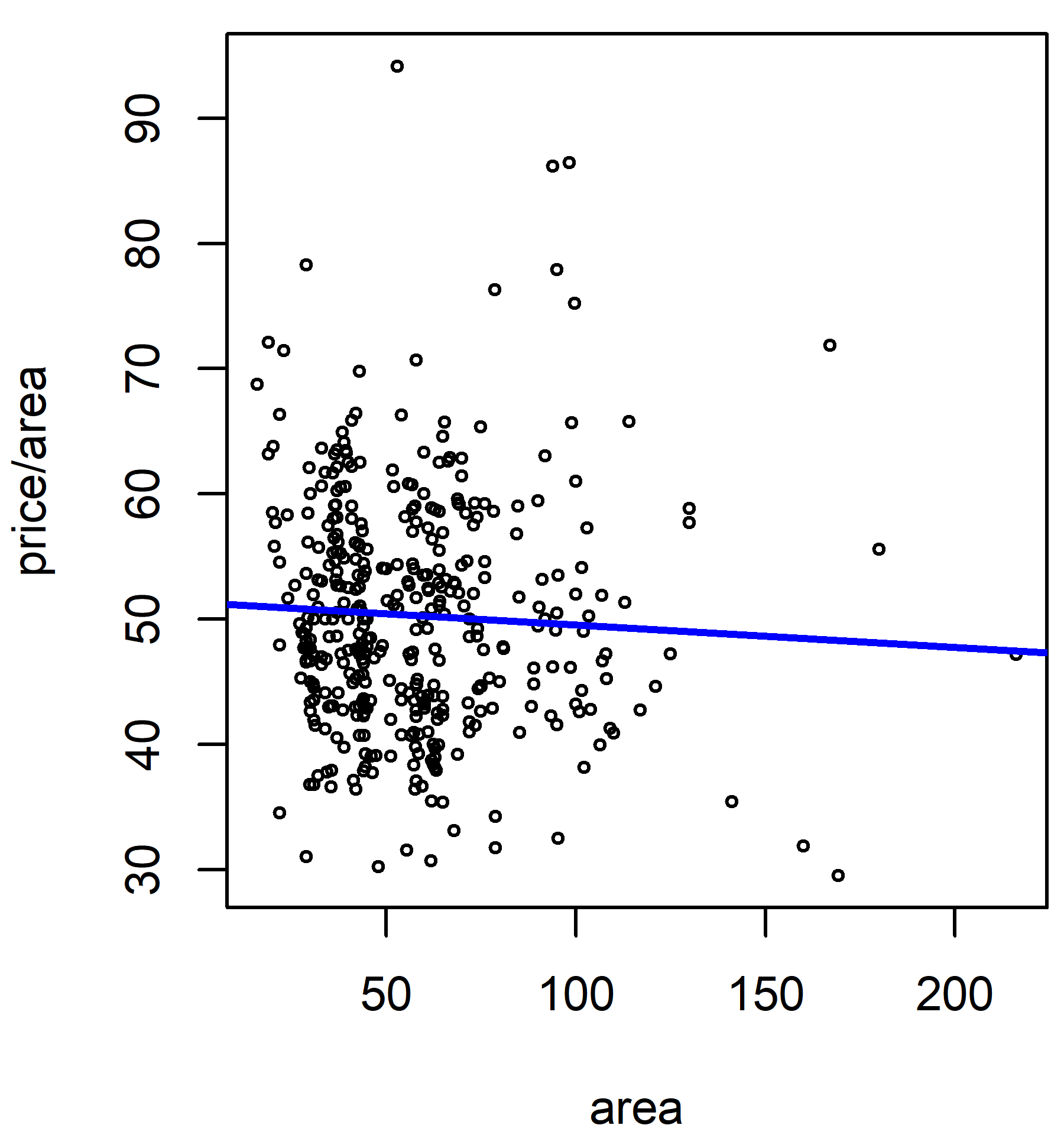
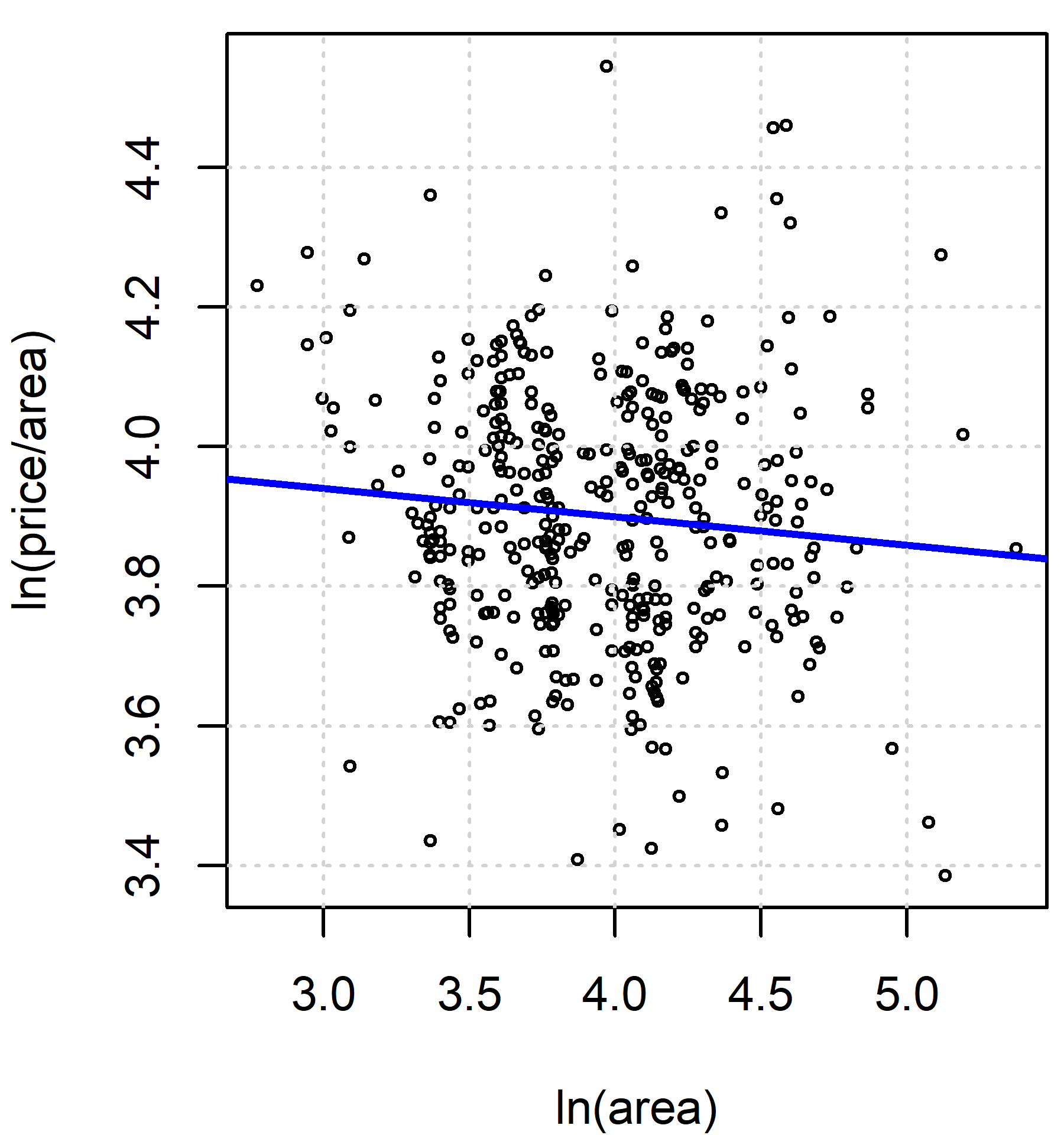
 

Рисунок 40. Две регрессии цены за м2 от площади

### Интерпретация коэффициентов в парной регрессии

После построения линейной регрессии возникает вопрос интерпретации полученных результатов. От этого зависит экономический смысл оценок по модели и выводы исследования.

Возьмем для начала простейшую парную регрессию. Ее теоретическое уравнение имеет вид

Коэффициенту (коэффициенту наклона) можно дать следующую интерпретацию: при приросте регрессора на 1 единицу зависимая переменная прирастает на единиц. При этом единицей измерения будет единица измерения , деленная на единицу измерения . В целом, если вырастет на , то вырастет на

Можно выразить эту зависимость производной:

Также можно рассматривать здесь не саму зависимую переменную, а ее математическое ожидание. Это уравнение теоретической линии регрессии

Тогда интерпретация коэффициента следующая: при приросте объясняющей перемененной на 1 единицу прирастает на единиц. (Если регрессор случайный, то в этой формуле следует использовать условное математическое ожидание .) Через производную

Геометрически это наклон (т. е. тангенс угла наклона) теоретической линии регрессии (см. рис. 41). В линейной по переменным парной регрессии производная одинаковая во всех точках .

Рисунок 41. Коэффициент парной регрессии как наклон линии регрессии

Производную, связанную с влиянием объясняющей переменной на объясняемую переменную, принято называть **маргинальным эффектом** (предельным эффектом) переменной .

Возможная интерпретация коэффициента (константы регрессии) состоит в том, что это математическое ожидание при условии равенства нулю :  
Однако не всегда значение является осмысленным, поэтому в общем случае проще смотреть на как на свободный член в линейной функции . Единица измерения совпадает с единицей измерения .

Пусть мы оценили парную регрессию и получили оцененное уравнение

Здесь – оценка , а – оценка . У оценок те же единицы измерения, что и у теоретических аналогов. Интерпретация оценок такая же, как и выше, только с добавлением слова «примерно». Например, для мы можем дать следующую интерпретацию: при приросте объясняющей перемененной на 1 единицу прирастает примерно на единиц.

Можно рассматривать оцененное уравнение как прогноз:  
Тогда при изменении значения на 1 единицу наш прогноз для меняется на единиц:

Для приведенного выше примера с ценами квартир  
коэффициент наклона имеет следующую интерпретацию: «Если бы площадь квартиры была на 1 м2 больше, то цена квартиры была бы примерно на тыс. рублей больше». При этом единицей измерения коэффициента будет тыс. руб./м2. Константу (тыс. руб.) можно интерпретировать как постоянную добавку к цене квартиры, не зависящую от ее площади и отражающую сам факт наличия жилья.

Для нелинейной по переменным регрессии меняется интерпретация всех коэффициентов. Размерность коэффициентов тоже не такая, как в линейном случае. Рассмотрим в качестве типичного примера линейную в логарифмах модель

Здесь можно рассчитать маргинальный эффект в виде производной

но величина эффекта в данном случае будет зависеть от конкретной точки (), в которой берется производная. Что касается величины , то ее значение зависит от точки и от распределения ошибки :

В то же время сам коэффициент здесь постоянен и имеет понятную интерпретацию. Это *эластичность* переменной по переменной . Напомним, что эластичность определяется как

При интерпретации в этом случае используются относительные приращения вида или процентные приращения . Интерпретация основана на правиле из математического анализа: если приращение мало́ по сравнению с (где ), то приближенно

Таким образом, относительное приращение положительной величины примерно равно приращению логарифма этой величины.

В линейной в логарифмах регрессии

процентный прирост зависимой переменной, соответствующий процентному приросту объясняющей переменной на , равен примерно  
В частности, при увеличении на 1 % переменная меняется примерно на %.

В полулогарифмических моделях вида log-lin и lin-log коэффициенты – это так называемые полуэластичности и соответственно.

В табл. 8 перечислены интерпретации по четырем разным парным регрессиям.

Таблица 8. Интерпретация коэффициента для 4 разных видов парной регрессии

|  |  |  |  |
| --- | --- | --- | --- |
| Модель | Уравнение | Словесная интерпретация | Формула  в приростах |
| lin-lin |  | увеличение на единицу приводит к изменению на единиц |  |
| log-log |  | увеличение на 1% приводит к изменению примерно на |  |
| lin-log |  | увеличение на 1% приводит к изменению примерно на единиц |  |
| log-lin |  | увеличение на единицу приводит к изменению примерно на | 100 |

### Интерпретация коэффициентов множественной регрессии

При рассмотрении множественной регрессии:

интерпретацию коэффициента при отдельной переменной можно дать по аналогии с интерпретацией коэффициента парной регрессии, если считать остальные регрессоры неизменными. Коэффициент имеет такую интерпретацию: «Приросту регрессора на 1 единицу соответствует прирост на единиц,*при прочих равных условиях*».

Если регрессоры и зависимая переменная – это простые функции некоторых исходных переменных (например, логарифмы), то при интерпретации можем использовать те же правила, о которых говорилось при обсуждении парной регрессии. Например, в линейной в логарифмах регрессии

процентный прирост зависимой переменной , соответствующий процентному приросту *j*-й объясняющей переменной на , равен примерно при неизменности всех остальных объясняющих переменных.

Такие стандартные интерпретации отдельных коэффициентов хорошо работают, если регрессоры независимы между собой. Однако в экономических приложениях независимость регрессоров встречается редко. Особенно аккуратно следует подходить к интерпретации регрессий, в которых регрессоры являются функциями некоторого набора исходных переменных.

#### Пример: Зависимость расходов на питание от доходов

Рассмотрим, используя данные обследования домохозяйств[[3]](#footnote-3), как зависят расходы на питание российских семей от их доходов . Для переменной используем ответ на вопрос «Вспомните, сколько примерно денег все члены Вашей семьи израсходовали на питание дома и вне дома в течение последних 30 дней?» Для переменной – ответ на вопрос «Каким был денежный доход всей Вашей семьи в течение последних 30 дней?»

Нас, скорее всего, интересует здесь зависимость от долгосрочного дохода, а не дохода за 30 дней, поскольку доход за отдельный месяц подвержен сильным случайным колебаниям. Чтобы несколько сгладить случайные колебания и сделать графики более наглядными, мы объединили домохозяйства в группы по 20 в зависимости от размера указанного дохода за 30 дней и использовали средние по группам.[[4]](#footnote-4) Кроме того, мы изъяли из выборки наблюдения по семьям, которые сообщили доход выше 200 тыс. руб.

Будем моделировать зависимость как квадратичную. Объяснение нелинейности состоит в том, что более обеспеченные домохозяйства склонны тратить на продукты питания меньшую часть дохода. Оцененное уравнение регрессии имеет следующий вид:

При изменении регрессора меняется и регрессор , поэтому стандартная интерпретация, что «коэффициент при измеряет, как при прочих равных условиях меняется при изменении » здесь неприменима. Для расчета маргинального эффекта переменной здесь можно использовать соответствующую производную:

Как видим, при нелинейности уравнения регрессии по переменным маргинальный эффект переменной не является постоянной величиной и зависит от выбранной точки (рис. 42). Например, при доходе (тыс. руб.) маргинальный эффект будет равен и человек с таким доходом из дополнительный тысячи рублей в среднем рублей истратит на питание. При доходе маргинальный эффект будет равен , так что из дополнительный тысячи рублей в среднем рублей пойдет на питание.

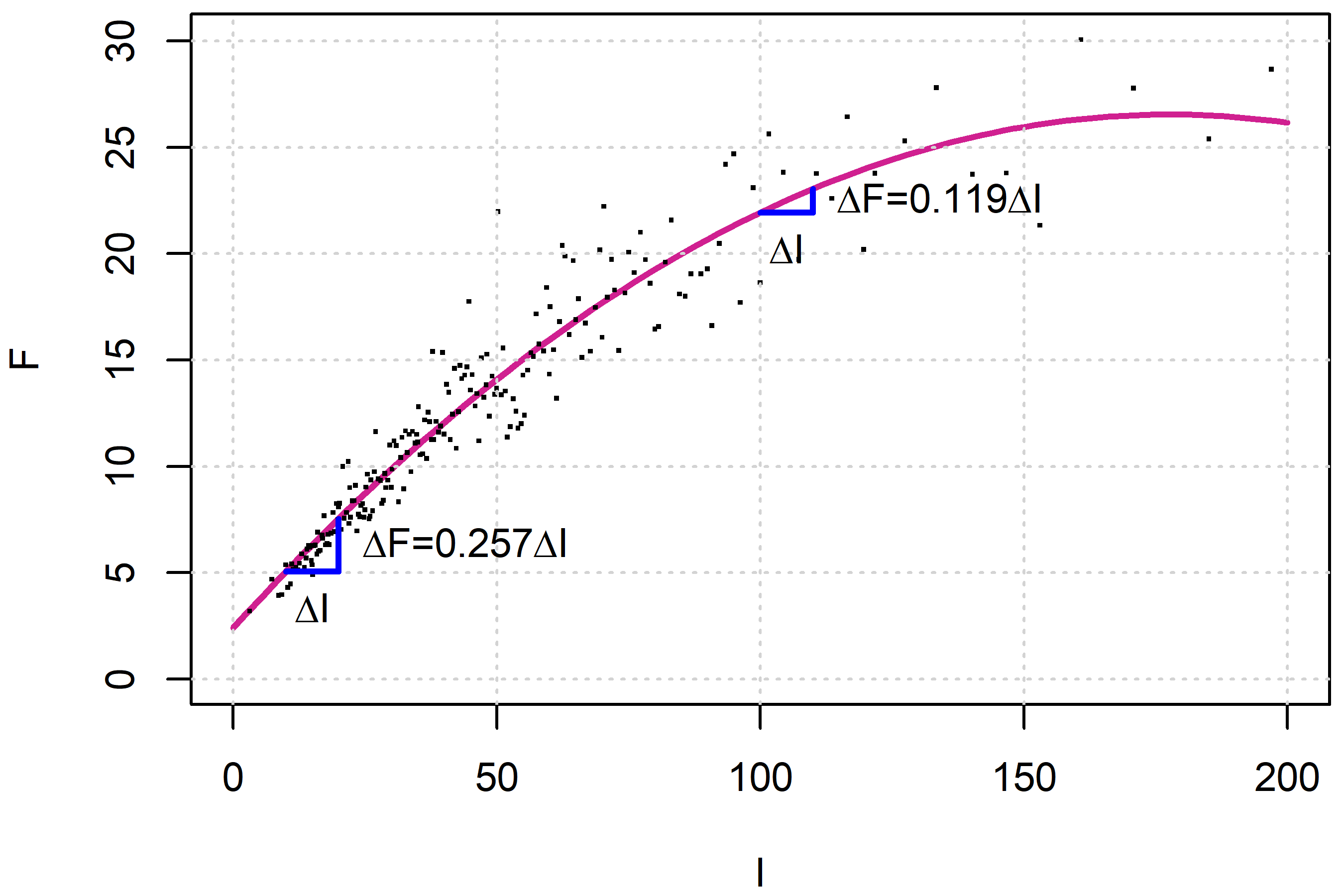


Рисунок 42. Изменение маргинального эффекта при квадратичной зависимости

Даже если нет такой жесткой зависимости между разными регрессорами, как в рассмотренном примере, все равно к интерпретации отдельного коэффициента следует относиться крайне осторожно. Строго говоря, в уравнении множественной регрессии  
в выполнении свойства  
мы можем быть уверены, только если  
для всех коэффициентов ,  
для всех остальных регрессоров , , и  
Именно поэтому в интерпретацию множественной регрессии следует добавлять слова «при прочих равных условиях». Даже для парной регрессии  
в выполнении свойства  
мы можем быть уверены, только если  
для коэффициентов и и

Если интерпретация дается в терминах предсказания того, как чье-то вмешательство в уровень переменной при неизменном уровне остальных регрессоров повлияет на уровень переменной , то требуется, чтобы оцененное уравнение имело соответствующее причинное содержание, а не было просто отражением корреляций между переменными. Это тесно связано с темой эндогенности, которая обсуждается в отдельной главе.

### Взаимодействия переменных

Частный случай нелинейности по переменным – это модели, в которых учитываются **эффекты взаимодействия** между разными переменными. Рассмотрим, например, регрессию с двумя объясняющими переменными:  
К этой модели мы можем добавить произведение переменных и , чтобы учесть возможную нелинейность из-за взаимодействия между этими переменными:

Подобный эффект может возникать по разным причинам. В частности, можно указанную модель рассматривать как парную регрессию от :  
в которой константа и коэффициент наклона меняются в зависимости от :  
Можно рассматривать ту же модель, наоборот, как парную регрессию от , в которой константа и коэффициент наклона меняются в зависимости от :

Для подобных моделей маргинальные эффекты зависят от точки, в которой считаются производные. Маргинальный эффект отдельной переменой , вообще говоря, будет зависеть от значений всех объясняющих переменных – эти значения нужно зафиксировать на некотором выбранном уровне. Например, для рассмотренной только что модели  
маргинальный эффект переменной будет равен  
а маргинальный эффект переменной будет равен

Эффекты взаимодействия между качественными переменными мы уже рассматривали в главе про фиктивные переменные. Возможны также эффекты взаимодействия между качественными и количественными переменными. Рассмотрим, например, зависимость переменной от количественной переменной и фиктивной переменной ( или ). С учетом возможного взаимодействия между и запишем модель   
В данном простом случае удобно записать модель в виде двух уравнений, связывающих и . Для наблюдений с уравнение регрессии имеет вид  
а для наблюдений с уравнение регрессии имеет вид  
Таким образом, при маргинальное значение равно сумме двух коэффициентов:

#### Пример: Зависимость цены квартиры от площади и района

Добавим к рассмотренным ранее данным о ценах квартир в г. Бердск (410 наблюдений) аналогичные данные по Октябрьскому району г. Новосибирска (1433 наблюдений). Пусть для бердских квартир. По-прежнему *price* – цена, а *area* – площадь.

Оценивание регрессии в исходных переменных дает уравнение

Согласно полученным оценкам, дополнительный квадратный метр жилой площади добавляет к цене квартиры в Октябрьском районе тыс. руб., а к цене квартиры в Бердске только тыс. руб. На рис. 43 оценки по модели представлены в виде двух линий регрессии – для Октябрьского района и Бердска. Видно, что линии имеют разный наклон.

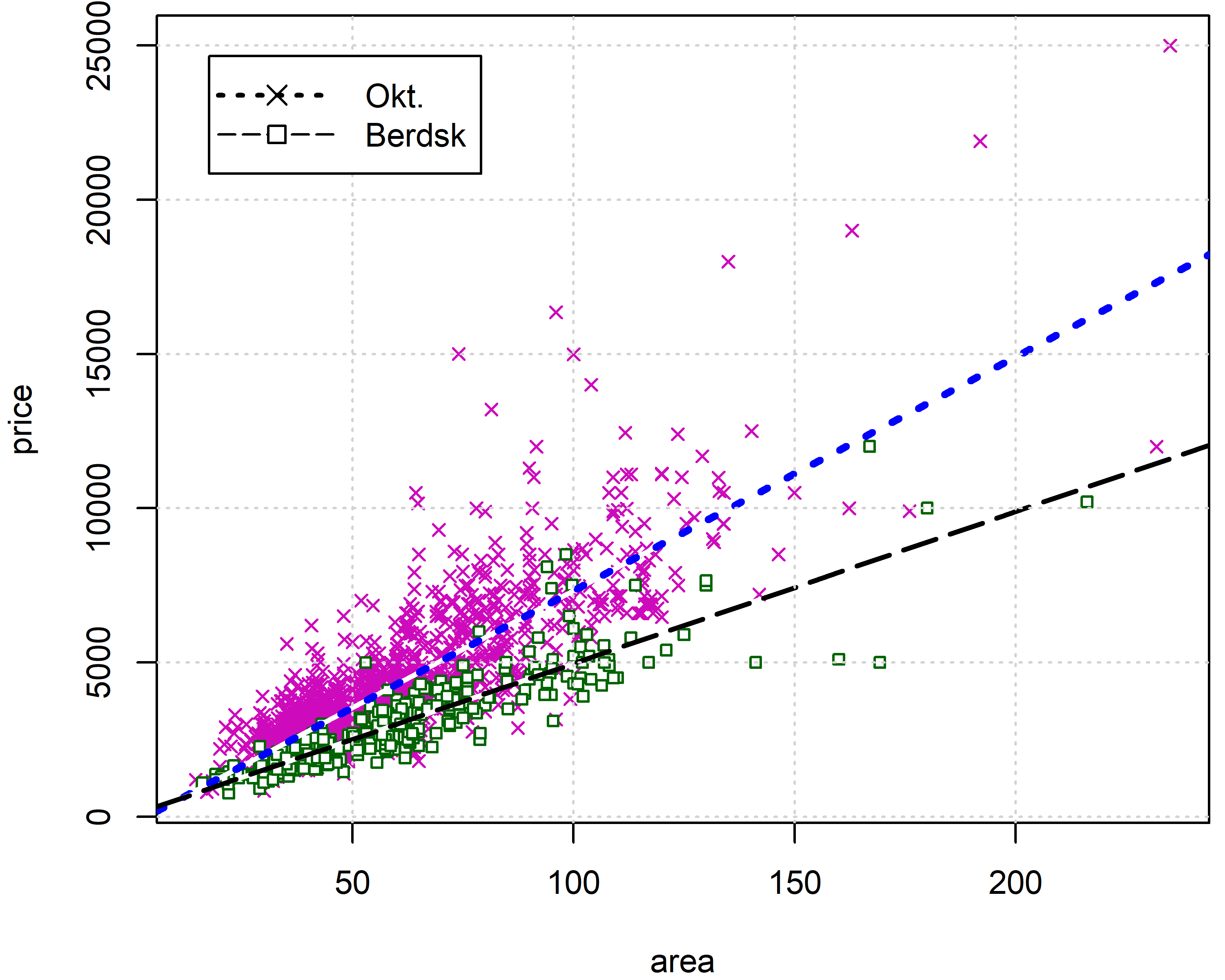


Рисунок 43. Зависимость цены квартиры от площади в г. Бердск и Октябрьском районе г. Новосибирска

Рассмотрим в заключение модель с двумя непрерывными количественными переменными, причем такую, что модель по одной из переменных является квадратичной.

#### Пример: Регрессия для цен алмазов

Пусть *price* – стоимость алмаза в долл. США, *carat* – вес алмаза в каратах, , – некоторые размеры алмаза в мм (условно «длина» и «ширина»).[[5]](#footnote-5) Всего имеется 953 наблюдения. Введем переменную («пропорции алмаза») и оценим модель зависимости логарифма цены алмаза (в долл. за карат) от логарифма веса, квадрата логарифма веса, переменной , а также от сочетания с логарифмом веса и квадратом логарифма веса. Получим следующее оцененное уравнение:

Поскольку в модели две объясняющие переменные, регрессию сложно показать на графике, но можно построить график зависимости логарифма цены от логарифма веса при фиксированном значении . На рис. 44 изображены линии регрессии при двух характерных значениях . Также на график добавлены точки наблюдений с разными маркерами. Два облака точек, соответствующие случаям и , расположены по-разному из чего можно заключить, что использование эффектов взаимодействия было оправданным.

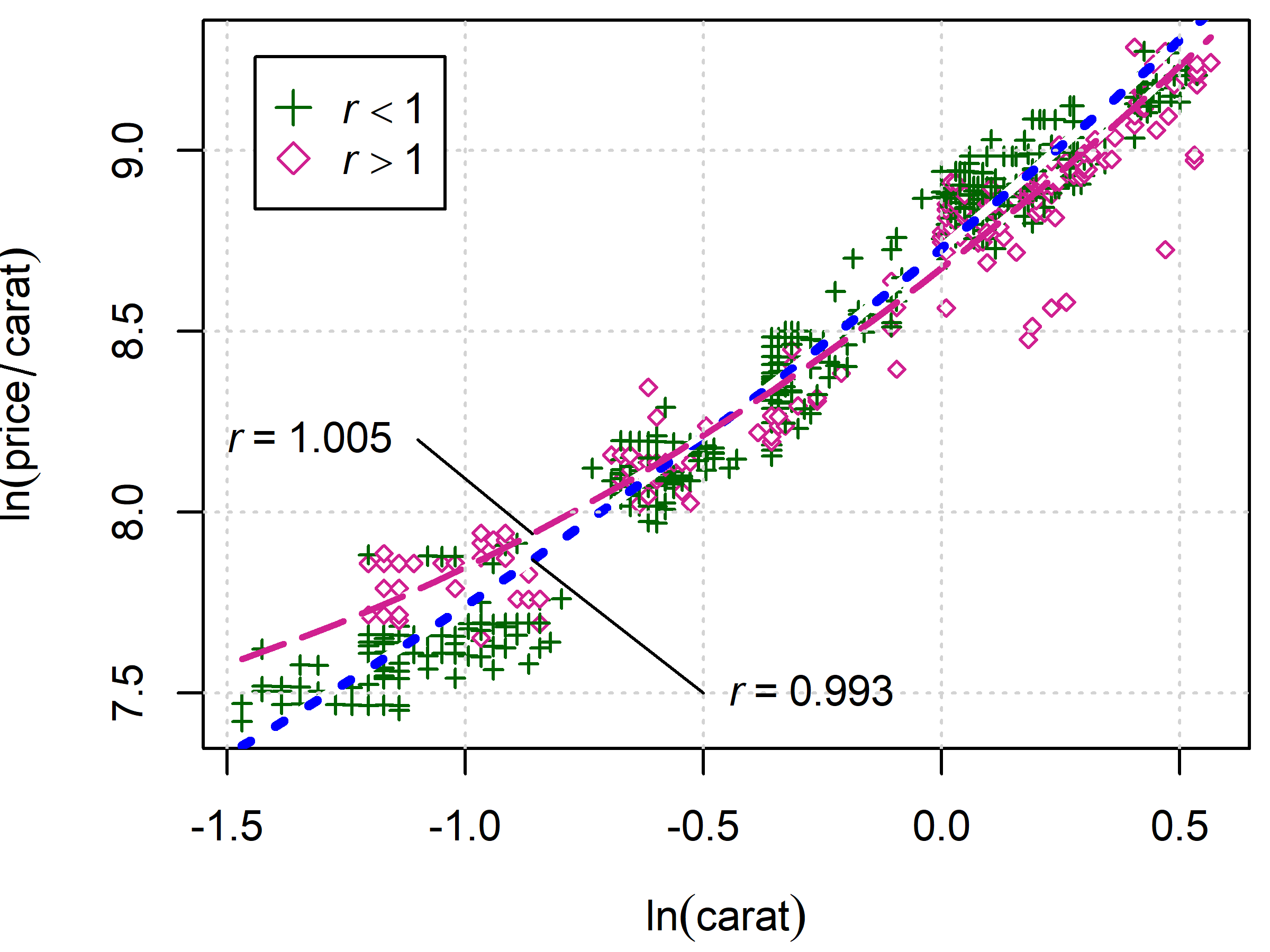


Рисунок 44. Модель для логарифма цены алмазов – взаимодействие переменных и

## Ошибки спецификации функциональной формы

### Последствия неверной спецификации функциональной формы

Раньше мы рассмотрели квадратичную зависимость (на примере связи между расходами на питание и доходами). Для подобной модели маргинальный эффект переменной непостоянен и зависит от точки, в которой берется производная (см. рис. 42). Если зависимость между теми же переменными задать как линейную, то маргинальный эффект уже будет постоянным. Отсюда ясно, что если зависимость на самом деле квадратичная, а оценивается вместо нее линейная, то возникает *смещение в оценке маргинального эффекта*. Кроме того, в этом случае возникает *смещение в прогнозах по модели* (какими будут расходы на питания у домохозяйства с данным доходом).

В целом смещения в оценке маргинального эффекта и в прогнозах по модели возникают почти всегда при неправильном выборе функциональной формы.

Продемонстрируем смещение в прогнозах в общей ситуации, когда есть зависимая переменная и детерминированные регрессоры . Пусть, как обычно, для построения прогноза получена оценка МНК по наблюдениям , . Прогноз для наблюдения c регрессорами рассчитывается по стандартной формуле . Предположим, что между математическим ожиданием зависимой переменной и регрессорами существует однозначная зависимость , причем она верна для всех возможных наблюдений (и для прогнозов тоже). Прогноз будет несмещенным, если математическое ожидание ошибки прогноза равно нулю, т. е. если . Таким образом, для несмещенности прогнозов требуется, чтобы  
где . Отсюда видно, что прогнозы по модели линейной регрессии от будут несмещенными для любых только если между и есть линейная зависимость. Если же зависимость нелинейна по , то, вообще говоря, будет наблюдаться смещение.

### Диагностика неверной функциональной формы

Из предыдущего обсуждения ясно, что, как правило, можно предложить много альтернативных функциональных форм для зависимостей между одними и теми же переменными. Конечно, исследователь не всегда может быть уверен, что правильно сформулировал модель. Важно уметь обнаруживать нарушения функциональной формы, чтобы не получать смещенные оценки маргинальных эффектов и смещенные прогнозы.

В случае парной регрессии  
можно использовать графический метод диагностики. Используется точечная диаграмма в координатах с наложенной на нее оцененной линией регрессии. Если облако точек неровно расположено вдоль линии регрессии, имеет какие-то заметные изгибы, то это признак неудачно выбранной функциональной формы.

Рассмотрим пример зависимости расходов на питание российских домохозяйств от их доходов. Выше мы моделировали зависимость как квадратичную. Что если использовать линейную по доходу модель? На точечной диаграмме с наложенной на нее линией регрессии видно, что прямая линия плохо соответствует конфигурации точек (рис. 45).

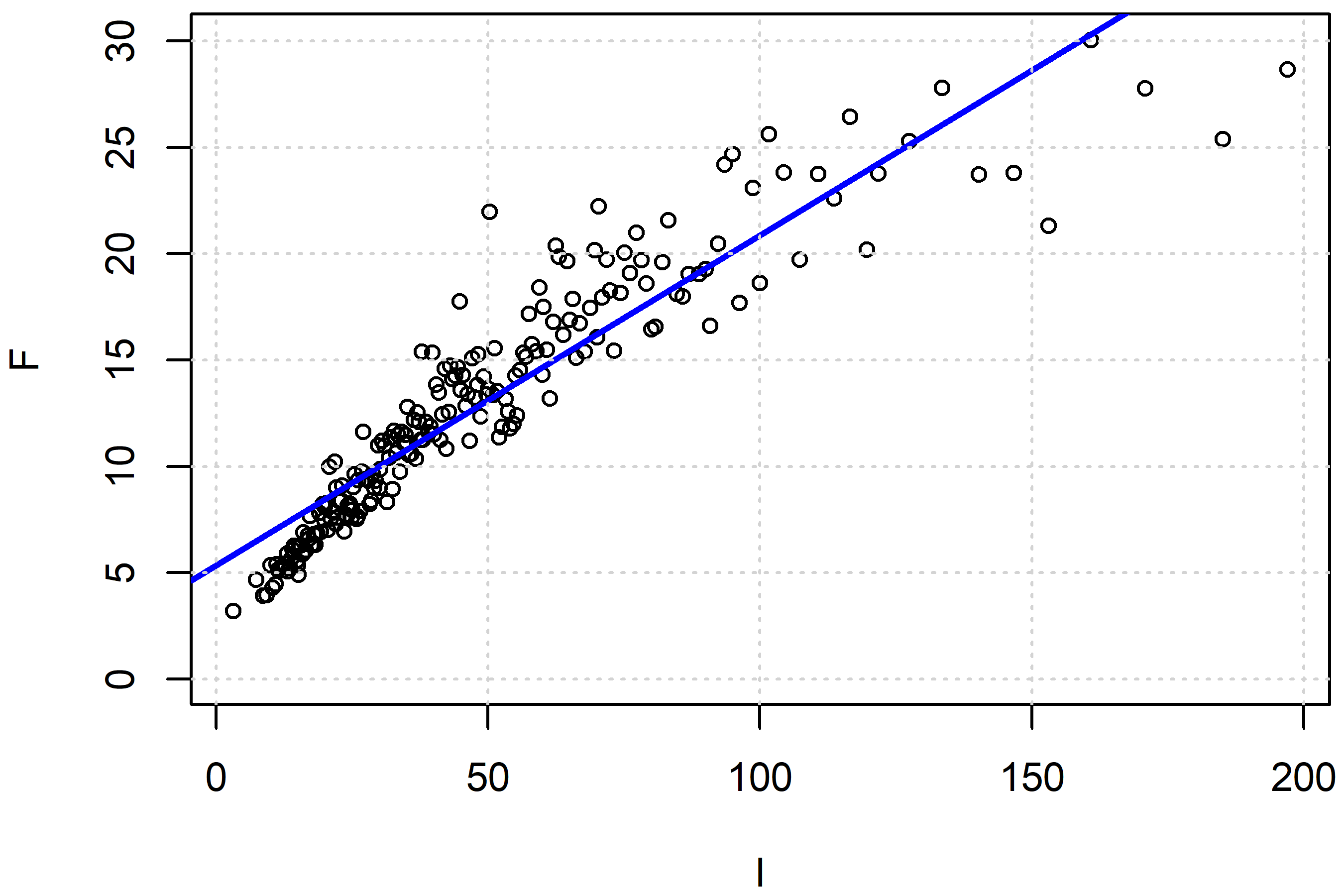


Рисунок 45. Зависимость расходов на питание от доходов, линейная по доходам модель

Похожим образом можно использовать график остатков регрессии по объясняющей переменной . В среднем остатки равны нулю. Если облако точек имеет кривизну, так что при некоторых значениях остатки в среднем заметно выше нуля, а при других значениях остатки в среднем заметно ниже нуля, то это симптом неудачно выбранной функциональной формы.

В примере с линейной зависимостью расходов на питание от доходов точечная диаграмма остатков от доходов ясно указывает на нелинейность (рис. 46). При низких доходах остатки оказываются в среднем отрицательными, при доходах в районе 50 тыс. руб. – в среднем положительными, а при высоких – опять в среднем отрицательными. Облако точек изгибается в виде дуги (на рисунке дуга показана штрихами).

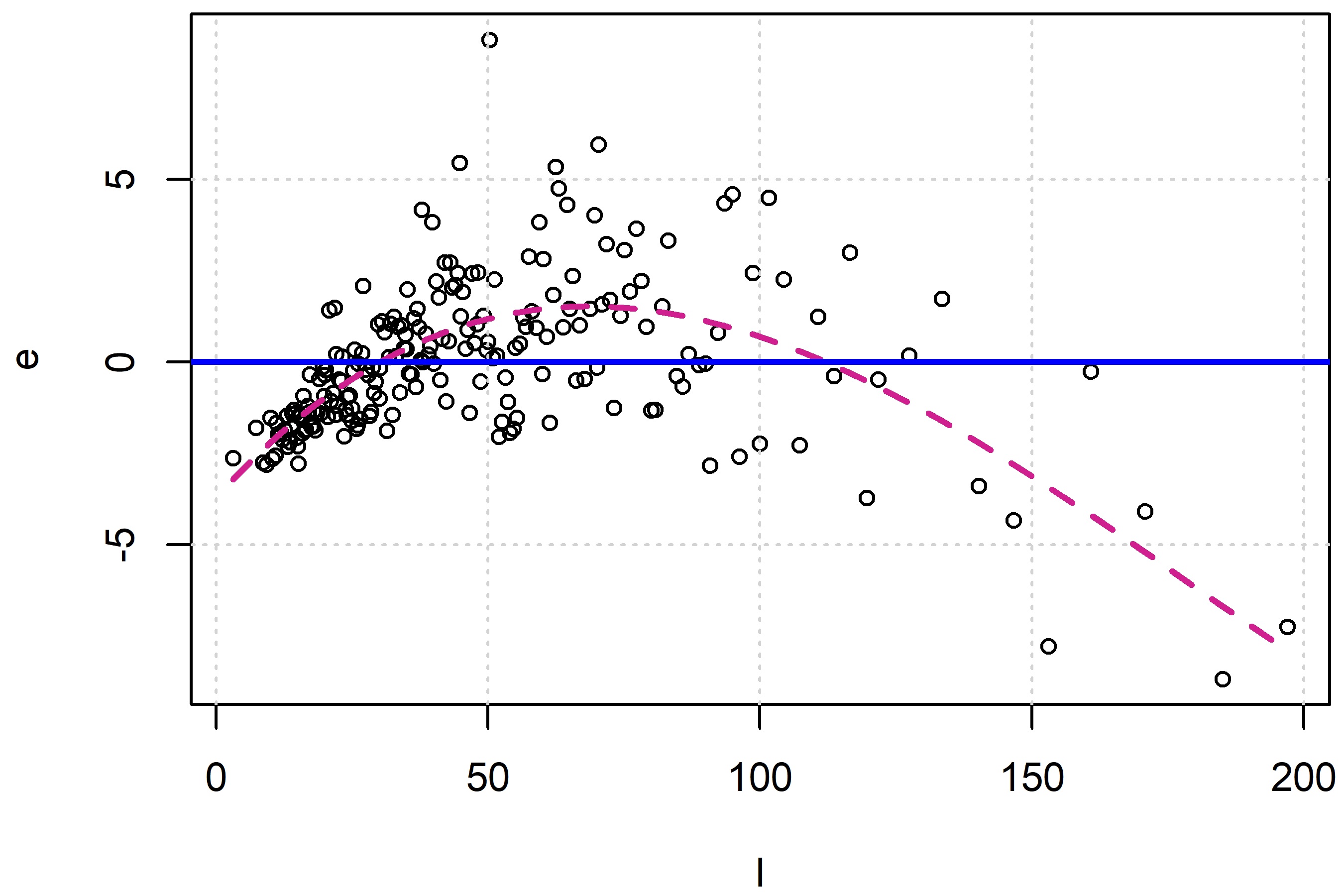


Рисунок 46. Остатки в линейной по доходам модели расходов на питание

Если мы имеем дело с моделью множественной регрессии, то подобные графики можно построить для каждой из объясняющих переменных. Можно также использовать какие-то функции и комбинации объясняющих переменных. По горизонтальной оси откладывается выбранная переменная , которая строится на основе объясняющих переменных, а по вертикальной оси зависимая переменная или остатки . В качестве такой переменной часто используются расчетные значения . График остатков можно дополнить горизонтальной прямой . График зависимой переменной по можно дополнить прямой .

Если облако точек ( от или от ) по форме изогнутое, то это может свидетельствовать о нелинейности. На графике остатков ( от ) нелинейности и другие нарушения функциональной формы могут проявиться также в том, что есть области значений , на которых точки лежат в среднем заметно выше или ниже нуля.

Рассмотрим, например упрощенную модель цен алмазов без дополнительных нелинейных членов:  
На графике остатков по расчетным значениям (рис. 47) заметны существенные недостатки функциональной формы модели. К точкам добавлена воображаемая линия, показывающая среднее остатков для разных интервалов расчетных значений. В облаке наблюдений видна кривизна. Также, например, при малых значениях все остатки положительны.

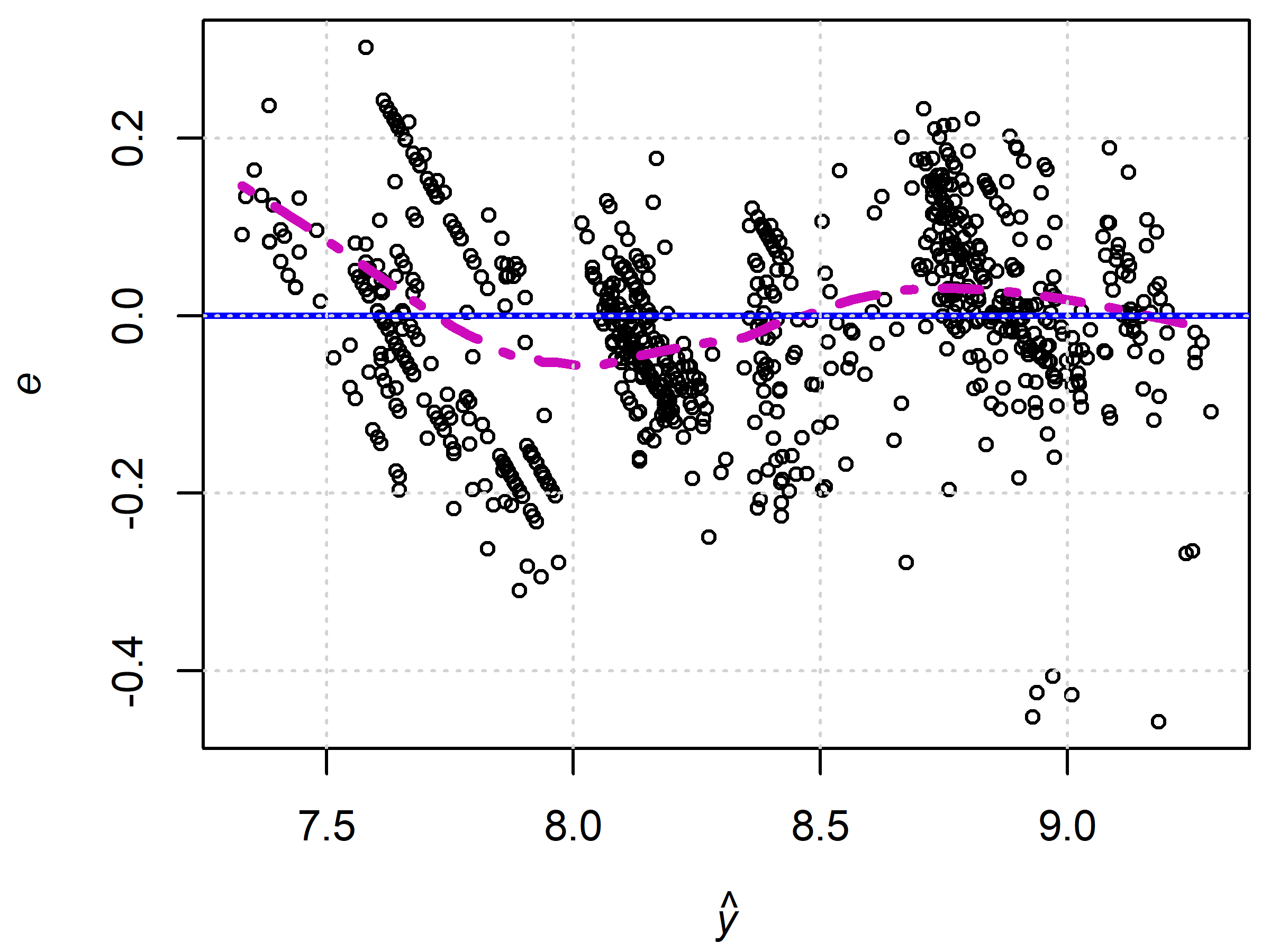


Рисунок 47. Остатки по расчетным значениям для упрощенной модели цен алмазов

Простейший формальный способ диагностики состоит в использовании вспомогательной регрессии. К проверяемой регрессии добавляются некоторые регрессоры , которые призваны отразить недостатки существующей функциональной формы. *Если добавленные регрессоры оказываются значимыми, то это может указывать на неудачную функциональную форму*. Укажем на некоторые возможные добавки.

**Тест RESET Рамсея** (Ramsey RESETtest) основан на добавлении в проверяемую регрессию степеней расчетных значений из проверяемой регрессии. Добавляются переменные .

**J-тест** Дэвидсона и Маккиннона (Davidson and MacKinnon J Test) основан на проверке модели с помощью альтернативной модели с другой функциональной формой (например, в одной модели используются сами объясняющие переменные, а в другой их логарифмы). В проверяемую модель добавляются расчетные значения из альтернативной модели .

Можно также, исходя из тех или иных соображений, добавлять произведения объясняющих переменных, их степени и другие нелинейные функции. При этом следует проявлять здравый смысл, поскольку при проведении большого количества тестов на добавление переменных велика вероятность случайно обнаружить эффект, который на самом деле отсутствует.

Заметим, что если добавленные переменные оказались статистически значимыми, это не обязательно означает, что в модели имеется большой потенциал для усовершенствования. Важно понять, насколько большим является соответствующий эффект, а не только убедиться в его статистической значимости. Можно сравнить, к примеру, сумму квадратов остатков и величину предсказанных по модели значений – насколько они сильно меняются при добавлении переменных .

Продолжая пример с ценами алмазов, проверим только что рассмотренную упрощенную модель с помощью добавления в нее квадрата и куба расчетных значений, т. е. проведем тест RESET. Для этого теста *F*-статистика равна 78.4 при очень маленьком *p*-значении, т. е. нелинейность значима. Поскольку зависимой переменной является логарифм цены, то стандартную ошибку регрессии можно выразить в процентах, умножив на 100%. В данном случае она уменьшается с 11.3% в проверяемой регрессии до 10.5% во вспомогательной регрессии. На рис. 48 показано изменение расчетных значений за счет добавления в регрессию новых переменных (). Видно, например, что для самых дешевых алмазов (у которых расчетные значения самые маленькие) прогноз цены меняется почти на 20% в сторону увеличения. По-видимому, с практической точки зрения этот эффект значителен и желательно поискать другую функциональную форму (или использовать ту более сложную, что мы оценивали до этого).

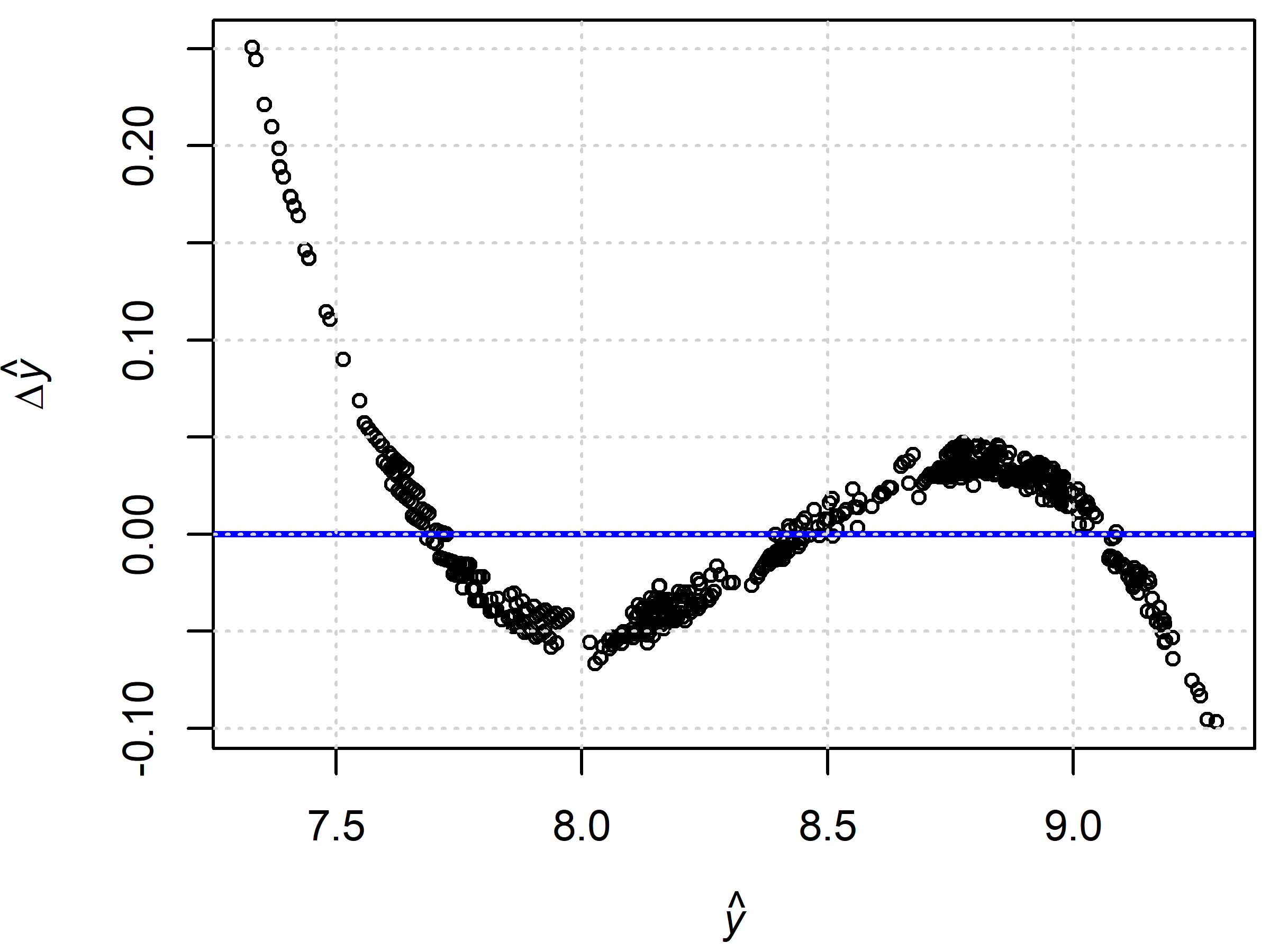


Рисунок 48. Изменение расчетных значений в тесте на функциональную форму

## Непостоянство коэффициентов модели. Тесты Чоу

### Непостоянство коэффициентов модели

Часто возникает сомнение в том, что для всех наблюдений модель неизменна, в частности, что параметры модели неизменны. Пусть все наблюдения в регрессии разбиты на две группы. В группе имеется наблюдений, в группе – наблюдений, так что . Запишем модель регрессии с разными коэффициентами:

Ситуация регрессии с разными коэффициентами в случае парной регрессии проиллюстрирована на рисунке. Здесь как константа, так и коэффициент наклона линии регрессии для двух групп наблюдений различаются.

A

Рисунок 49.

Без ограничения общности можно считать, что сначала идут наблюдения из группы , а потом – наблюдения из группы . Тогда исходную модель можно записать как единую регрессию в блочном виде:

Вообще говоря, если нет каких-то дополнительных ограничений на коэффициенты, то можем оценивать регрессию не в блочном виде, а отдельно по наблюдениям из групп и . Действительно, остатки равны  
а сумма квадратов остатков равна

Два вектора коэффициентов и не связаны между собой. Минимизируя эту сумму по и мы, тем самым, будем минимизировать по и по . Минимум достигается при

В то же время, данную модель можно представить в альтернативном виде, добавив в регрессию с постоянными коэффициентами набор дополнительных переменных. Обозначим через разницу между двумя векторами коэффициентов:

Тогда можем записать модели для двух групп наблюдений в виде

В блочном матричном виде

Очевидно, что две регрессии (с параметрами и параметрами ) эквивалентны с точностью до замены переменных.

Первая часть матрицы регрессоров этой модели соответствует модели с постоянными коэффициентами (т. е. , ):  
а во второй части матрицы регрессоров стоит блок

Для создания соответствующих переменных второй части можно использовать фиктивную переменную , такую что

Если в регрессии с постоянными коэффициентами использовались переменные  
то к ним надо добавить еще переменную  
Это все возможные сочетания исходных переменных с (включая 1, которой соответствует ).

Подход на основе фиктивной переменной отличается гибкостью. Если мы считаем, что меняется только константа, то добавляем только переменную (такое мы уже делали раньше). Если считаем, что меняются только константа и коэффициент при переменной , то добавляем и . И так далее.

### Тест Чоу на постоянство коэффициентов

Можно формально проверить, что коэффициенты регрессии для наблюдений из группы и наблюдений из группы одинаковы. Тест на равенство коэффициентов регрессии для двух групп наблюдений называют **тестом Чоу** (так называемый 1-й тест Чоу**)**. Нулевая гипотеза имеет вид

F-статистику для теста Чоу можно рассчитать через общую формулу по суммам квадратов остатков. Если нулевая гипотеза верна, то наблюдения описываются моделью с постоянными коэффициентами . Оценки МНК равны

Соответствующие остатки обозначим через (так как они получены по единой модели для групп и ):

Сумму квадратов остатков обозначим

С другой стороны, сумма квадратов остатков в регрессии без ограничений равна

Здесь – остатки регрессии, полученные по наблюдениям , а – остатки регрессии, полученные по наблюдениям , как если бы по двум группам оценивали регрессии по-отдельности.

В модели без ограничений оценивается коэффициентов (удвоенное количество). Всего мы имеем ограничений (столько же, сколько коэффициентов в модели с ограничением). Таким образом, *F*-статистика для теста Чоу равна

Если эта статистика больше критического значения, то отвергается и делаем вывод, что коэффициенты разные.

Для того чтобы применить этот тест, нужно оценить модель по двум группам наблюдений. Формально это можно сделать, когда количество наблюдений не меньше количества параметров в каждой из групп, т. е. , , и хотя бы одно неравенство строгое. Но, конечно, желательно иметь существенно больше наблюдений.

Альтернативно тест Чоу можно представить как тест на добавление переменных  
в регрессию с постоянными коэффициентами, т. е. как тест на добавление блока

Здесь мы хотим проверить гипотезу, что коэффициенты при добавленных переменных равны нулю:  
Тем самым мы проверим гипотезу, что коэффициенты исходной модели постоянны (). Используется F-статистика для проверки значимости добавляемых переменных.

Заметим, что приведенная здесь статистика теста Чоу имеет распределение только при выполнении стандартных предположений, в частности, предположения о гомоскедастичности ошибок (о равенстве дисперсий ошибок всех наблюдений). Но это не очень правдоподобное предположение, особенно если принять во внимание возможное различие моделей для частей и (дисперсии в двух частях могут быть разными). В связи с этим рекомендуется использовать формулу *F*-статистики не на основе сумм квадратов остатков, а на основе ковариационной матрицы коэффициентов с поправкой на гетероскедастичность (см. следующую главу). В данном случае проведение теста Чоу при помощи вспомогательной регрессии с добавлением переменных может быть более предпочтительным.

### Тест Чоу на точность прогноза (2-й тест Чоу)

То, что модель для двух групп наблюдений одна и та же, можно проверить и по-другому. Пусть сначала рассчитывается регрессия по наблюдениям, а затем по всем наблюдениям. Если полученные результаты существенно отличаются, то это должно означать, что во второй части модель каким-то образом поменялась.

Реализуем эту идею с помощью вспомогательной регрессии, к которой можно применить тест добавления переменных:  
где единичная матрица размерности

Эта регрессия с ограничением совпадает с регрессией по всем наблюдениям по первоначальной модели , и остатки равны .

Если же оценить вспомогательную регрессию без ограничений, то, как можно показать, оценки совпадут с оценками по наблюдениям, т. е. по регрессии . Остатки будут иметь вид .

Суммы квадратов остатков в двух моделях равны и , соответственно. Количество ограничений . Таким образом, получаем следующую статистику:  
Статистика имеет указанное распределение, если выполнена нулевая гипотеза Если нулевая гипотеза принимается, это означает, что модель не менялась.

Второй тест Чоу можно интерпретировать также как тест на точность прогноза. В этом контексте группу наблюдений принято называть **тренировочной**, а группу наблюдений **тестовой**. Поскольку — прогнозы, полученные для наблюдений на основе оценок по наблюдениям , то из второго уравнения системы следует, что оценки равны ошибкам такого прогноза:

Таким образом, проверяя гипотезу , мы проверяем, насколько точны прогнозы. Если модель по второй части выборки отличается от модели по первой части, то ошибки прогноза будут большими, и мы отклоним нулевую гипотезу.

## Приложение. Линейные преобразования в регрессии

Рассмотрим сначала замены переменных в параметрах регрессии. Пусть вместо коэффициентов мы используем коэффициенты такие что  
Если мы подставим это в исходную модель, то получим  
или  
Это линейная регрессия с зависимой переменной , матрицей регрессоров и коэффициентами . Если мы найдем в преобразованной регрессии с помощью МНК оценки коэффициентов , скажем , то легко можем получить из них оценки коэффициентов :  
Ковариационная матрица оценки связана с ковариационной матрицей оценки уравнением

Остатки в обеих регрессиях одинаковые, но, например показатель при в общем случае разный из-за того, что зависимые переменные разные. Это говорит о том, что надо осторожно относиться к величине .

Рассмотрим теперь однозначные линейные преобразования переменных регрессии. Пусть вместо и мы строим регрессию для и :

где  
 – невырожденная квадратная матрица, – вектор-столбец, – скаляр. Заметим, что здесь не входит в формулу для , иначе это будут принципиально разные регрессии. Перейдя к исходным переменным получим  
или  
Видно, что коэффициенты и связаны уравнением  
Оценки связаны с оценками аналогичным уравнением

Остатки соотносятся как , так что в общем случае суммы квадратов остатков разные. Если , то коэффициенты детерминации в двух регрессиях могут быть разными, но при они одинаковые, поскольку множитель сократится в формуле коэффициента детерминации.

## Контрольные вопросы

1. С помощью какой регрессии можно проверить правильность функциональной формы уравнения регрессии?
2. Докажите, что в модели  
   остатки будут иметь вид .

## Экзаменационные вопросы

1. Основные типы нелинейных моделей и их оценка.
2. Оценка и интерпретация коэффициентов модели типа: **lin-lin, log-log, lin-log и log-lin.**
3. Выбор между альтернативными регрессионными моделями.
4. Выбор функциональной формы зависимости с помощью теста Рамсея.
5. Чоу тесть на постоянство модели.
6. Чоу тесть на точность прогноза.

## Литература

1. Вербик М. Путеводитель по современной эконометрике. – М.: Научная книга, 2008. [стр.93-98, 110-134]
2. Картаев Ф. С. Введение в эконометрику. Учебник : – Москва: Проспект, 2019. [стр. 116-127]
3. Магнус Я. Р., Катышев П. К., Пересецкий А. А. Эконометрика. Начальный курс: Учебник. – 6-е изд., перераб. и доп. – М.: Дело, 2004 [стр. 78-88,112-118]
4. Суслов В. И., Ибрагимов Н. М., Талышева Л. П., Цыплаков А. А. Эконометрия. – Новосибирск: СО РАН, 2005. [стр. 572-582]
5. Цыплаков А. А. Некоторые эконометрические методы. Метод максимального правдоподобия в эконометрии. —Новосибирск: НГУ, 1997. [стр. 6-17]
6. Wooldridge J. M. Introductory econometrics. A modern approach. 6th ed. – Cengage Learning, 2016. [стр.36-40,63-73,166-180]

# Лекция: – Регрессия с ковариационной матрицей ошибок общего вида

## МНК в случае ковариационной матрицы ошибок общего вида

Мы исходили из выполнения ряда предположений: (A0)–(A5). Очень часто те или иные предположения плохо согласуются с имеющимися данными. Если не выполнены предположения (A0), (A1) и (A2), то это может привести к серьезным последствиям – смещенности и несостоятельности оценки МНК, поэтому от этих предположений редко отказываются при использовании регрессионного анализа. В то же время отказ от предположений (A3), (A4) и (A5) сам по себе не означает, что оценка МНК непригодна. Более того, есть основания считать, что эти предположения слишком часто не согласуются с имеющимися данными.

Здесь мы рассмотрим, к чему приводит отказ от предположений (A3) (нет гетероскедастичности ошибок) и (A4) (нет автокорреляции ошибок).

Считаем, что (A0), (A1) выполнены, то есть наша модель имеет вид  
где . Как и ранее, для упрощения теоретических рассуждений вместо предположения (A2) исходим из того, что матрица регрессоров детерминированная.

Как мы видели, оценку МНК  
можно представить в виде  
Взяв математическое ожидание от обеих частей, мы доказывали *несмещенность* оценки МНК:  
При этом используются только предположения (A0), (A1) и (A2).

Можно сделать вывод, что наличие гетероскедастичности и автокорреляции ошибок в регрессии само по себе *не приводит к смещенности оценки МНК*.

Рассмотрим другие свойства оценки МНК. Вычислим ковариационную матрицу коэффициентов МНК:  
Видим, что в общем случае ковариационная матрица коэффициентов МНК равна  
где мы обозначили ковариационную матрицу ошибок через :  
Таким образом, ковариационная матрица коэффициентов МНК в общем случае имеет вид **сэндвича**: матрица умножается слева и справа на .

Как мы видели, при выполнении предположений (A3) и (A4), т. е. при , выражение-сэндвич упрощается, и мы получаем стандартную формулу

## Последствия нарушения предположений

С учетом предыдущего перечислим последствий невыполнения гипотез (A3) и (A4) при том условии, что предположения (A0), (A1) и (A2) выполняются.

* Оценка МНК остается несмещенной.
* Оценка МНК в общем случае не является наилучшей в классе линейных несмещенных оценок (не обладает свойством BLUE). Таким образом, в классе линейных несмещенных оценок *найдется более точная оценка*, чем оценка МНК. Далее мы рассмотрим соответствующий метод – обобщенный метод наименьших квадратов.
* В общем случае ковариационная матрица коэффициентов МНК имеет вид, отличный от . Как следствие обычная оценка является некорректной оценкой ковариационной матрицы коэффициентов МНК (формально – смещенной и несостоятельной).
* Следствием некорректности обычной оценки ковариационной матрицы коэффициентов МНК является то, что обычные *t*- и *F*-статистики также являются некорректными. Аналогичные проблемы возникают и с доверительными интервалами. Распределение *t*- или *F*-статистик при нулевой гипотезе может сильно отличаться от обычного *t* (соответственно, *F*) распределения. Таким образом, привычные тесты при наличии автокорреляции и гетероскедастичности могут давать ложные результаты – отклонять нулевую гипотезу чаще или реже, чем требуется. Например, при номинальном уровне значимости 5% на самом деле вероятность отклонения нулевой гипотезы может быть 1% или 20%.

Последнее обстоятельство является наиболее серьезным. Из-за него проведение статистических тестов в модели регрессии оказывается под вопросом.

## Обобщенный МНК

Если ковариационная матрица ошибок известна (с точностью до множителя), то эту информацию можно использовать для получения оценки, которая обладает свойством BLUE. Будем считать, что  
где матрица нам известна. Здесь – некоторый положительный множитель (например, , если матрица точно известна). Мы здесь используем обозначение , но это пока что просто некоторый множитель, а не дисперсия.

Из теории матриц известно, что из любой положительно (полу)определенной симмет­ричной матрицы () можно извлечь квадратный корень, то есть представить А в виде  
где – квадратная матрица . Есть очень много способов выбрать матрицу . Самый простой способ – использовать так называемое разложение Холецкого, когда ищется в виде верхней (или нижней) треугольной матрицы. Это такая матрица, в которой ниже главной диагонали стоят нули (соответственно, в нижней треугольной нули выше главной диагонали). Можно также найти корень на основе так называемого спектрального разложения матрицы .

Поскольку является положительно определенной и симметричной, то тоже является положительно определенной и симметричной. Таким образом, существует матрица (), которая является квадратным корнем из :

Несложно проверить, что при этом будет выполнено и

Домножим теоретическое уравнение регрессии слева на матрицу . Получим  
или  
где мы обозначили

Ясно, что для преобразованного уравнения регрессии выполнены предположения (A0), (A1) и (A2). Найдем ковариационную матрицу ошибок в данной регрессии:  
Видим, что в регрессии с преобразованными переменными ошибка удовлетворяет предположениям (A3) и (A4), то есть эти ошибки не автокоррелированы и у них одинаковая дисперсия .

Поскольку выполнены предположения (A0)–(A4), то оценки МНК в регрессии с преобразованными переменными *будут обладать свойством BLUE*. С точки зрения исходной регрессии это будут так называемые оценки **обобщенного метода наименьших квадратов** (**ОМНК**, англ. *generalized least squares*, *GLS*):

Если подставим в эту формулу , и учтем, что , то получим

Это формула аналогична формуле для обычного МНК, но между перемножаемыми матрицами вставляется некоторая взвешивающая матрица , которая с точностью до множителя является обратной к ковариационной матрице ошибок .

Оценку ОМНК можно получить, решая задачу минимизации обобщенной суммы квадратов остатков:

Если матрица была корректно задана, то, как видно из регрессии с преобразованными переменными и , коэффициенты ОМНК будут иметь ковариационную матрицу  
или

Если множитель неизвестен, то для получения оценки этой матрицы вместо требуется подставить оценку. В качестве такой оценки следует взять несмещенную остаточную дисперсию из регрессии с преобразованными переменными:  
где

Можно также использовать формулу  
где  
Таким образом, в ОМНК используется формула

## Доступный обобщенный МНК

Выше мы предполагали, что ковариационная матрица ошибок нам известна (возможно, с точностью до положительного множителя). В большинстве случаев матрица на самом деле неизвестна. Можно попытаться построить для матрицы некоторую дополнительную параметрическую модель. Тогда  
где – неизвестный параметр. (Это может быть один параметр или вектор.) В этом случае часто использует **доступный обобщенный метод наименьших квадратов** (англ. *feasible generalized least squares*, *FGLS*).

Для разных моделей детали доступного обобщенного метода наименьших квадратов могут различаться, но его общая схема выглядит следующим образом:

1) Оцениваем коэффициенты регрессии обычным МНК. Получаем оценку и соответствующие остатки .

2) С помощью остатков оцениваем параметр . Получаем оценку .

3) Подставляем вместо в и вычисляем соответствующие оценки обобщенного МНК

При использовании доступного ОМНК мы, строго говоря, уже не получаем оценки со свойством BLUE, но во многих случаях это достаточно хорошее приближение.

Существует также вариант данного метода – **итеративный доступный обобщенный метод наименьших квадратов**. Смысл его состоит в том, что оценки с шага 3) можно использовать для вычисления остатков  
а затем использовать эти остатки для шага 2) – оценки параметра . Далее шаги 2) и 3) чередуются до тех пор, пока не будет достигнута сходимость:

## Подходы к моделированию в случае ковариационной матрицы ошибок общего вида

Имеется два основных подхода:

1) Использование обычного МНК. При таком подходе для ковариационной матрицы коэффициентов МНК строится некоторая состоятельная оценка типа «сэндвич». Эта оценка используется при построении скорректированных *t*- и *F*-статистик.

2) Использование одного из вариантов обобщенного МНК.

На практике используют как тот, так и другой подход. Преимуществом второго подхода является то, что он позволяет получить несколько более точные оценки. К преимуществам первого подхода можно отнести то, что для его использования, как правило, не требуется предполагать определенную параметрическую модель для мат­рицы . Кроме того, следует учитывать, что если ошибиться в спецификации модели для во втором подходе, то оценки ОМНК уже не обладают свойством BLUE. В целом с ними возникают такие же сложности, как и с оценками обычного МНК. Поэтому имеет смысл для оценки обобщенного МНК также использовать ковариационную матрицу типа «сэндвич».

## Контрольные вопросы

1. В чем смысл обобщенного МНК (ОМНК)?
2. При нарушении какой гипотезы применятся ОМНК?
3. Какие свойства МНК-оценок коэффициентов регрессии теряются, если ошибки по наблюдениям коррелированы и/или имеют разные дисперсии?
4. Построить оператор ОМНК-оценивания, вывести формулу для матрицы ковариации оценок параметров в этом случае.
5. Проверьте несмещенность оценки ОМНК :
6. Как надо преобразовать уравнение регрессии, чтобы применение МНК (после преобразования) было эквивалентно применению ОМНК к исходному уравнению?
7. Зачем вводится матрица B при ОМНК и какую размерность имеет данная матрица?
8. Проверить равенство
9. Доказать, что
10. Итеративный доступный обобщенный метод наименьших квадратов : в чем его смысл?

## Экзаменационные вопросы

1. Обобщенный метод наименьших квадратов.
2. Показать, что ОМНК-оценки относятся к классу **BLUE**.

## Литература

1. Вербик М. Путеводитель по современной эконометрике. – М.: Научная книга,. 2008. [стр. 138-142]
2. Магнус Я. Р., Катышев П. К., Пересецкий А. А. Эконометрика. Начальный курс: – Учебник. – 6-е изд., перераб. и доп. – М.: Дело, 2004. [стр. 158-166]
3. Суслов В. И., Ибрагимов Н. М., Талышева Л. П., Цыплаков А. А. Эконометрия.– Новосибирск: СО РАН, 2005. [стр. 257-258]
4. Картаев Ф. С. Введение в эконометрику. Учебник : – Москва: Проспект, 2019. [стр. 159-163]

# Лекция: – Гетероскедастичность ошибок

## Гетероскедастичность ошибок – определение и последствия

Мы предполагали, что ошибки во всех наблюдениях регрессии имеют одинаковые дисперсии, т. е. . Это предположение (A3) – гомоскедастичность. Кроме того, мы предполагали отсутствие корреляции между ошибками разных наблюдений. Это предположение (A4) – отсутствие автокорреляции.

Будем по прежнему предполагать, что автокорреляция отсутствует. Таким образом, ковариационная матрица ошибок является диагональной. Однако теперь мы будем предполагать, что дисперсии ошибок могут быть разными.

Обозначим дисперсию ошибки *i*-го наблюдения через :

Таким образом, мы предполагаем, что в нашей модели  
вектор ошибок имеет диагональную ковариационную матрицу с элементами :

Если не все дисперсии ошибок совпадают между собой, то имеет место **гетероскедастичность** ошибок.

Как правило, различие дисперсий бывает связано с зависимостью от каких-то факторов. Пусть, к примеру, дисперсия в парной регрессии от зависит от . А именно, пусть дисперсия тем больше, чем больше . Этот случай изображен на рис. 50. Для некоторых на этом графике изображены небольшие графики плотности зависимой переменной . Кроме различия средних, определяемых линией регрессии, в этих распределениях мы видим, что чем больше , тем шире разброс распределения. Здесь видна гетероскедастичность.

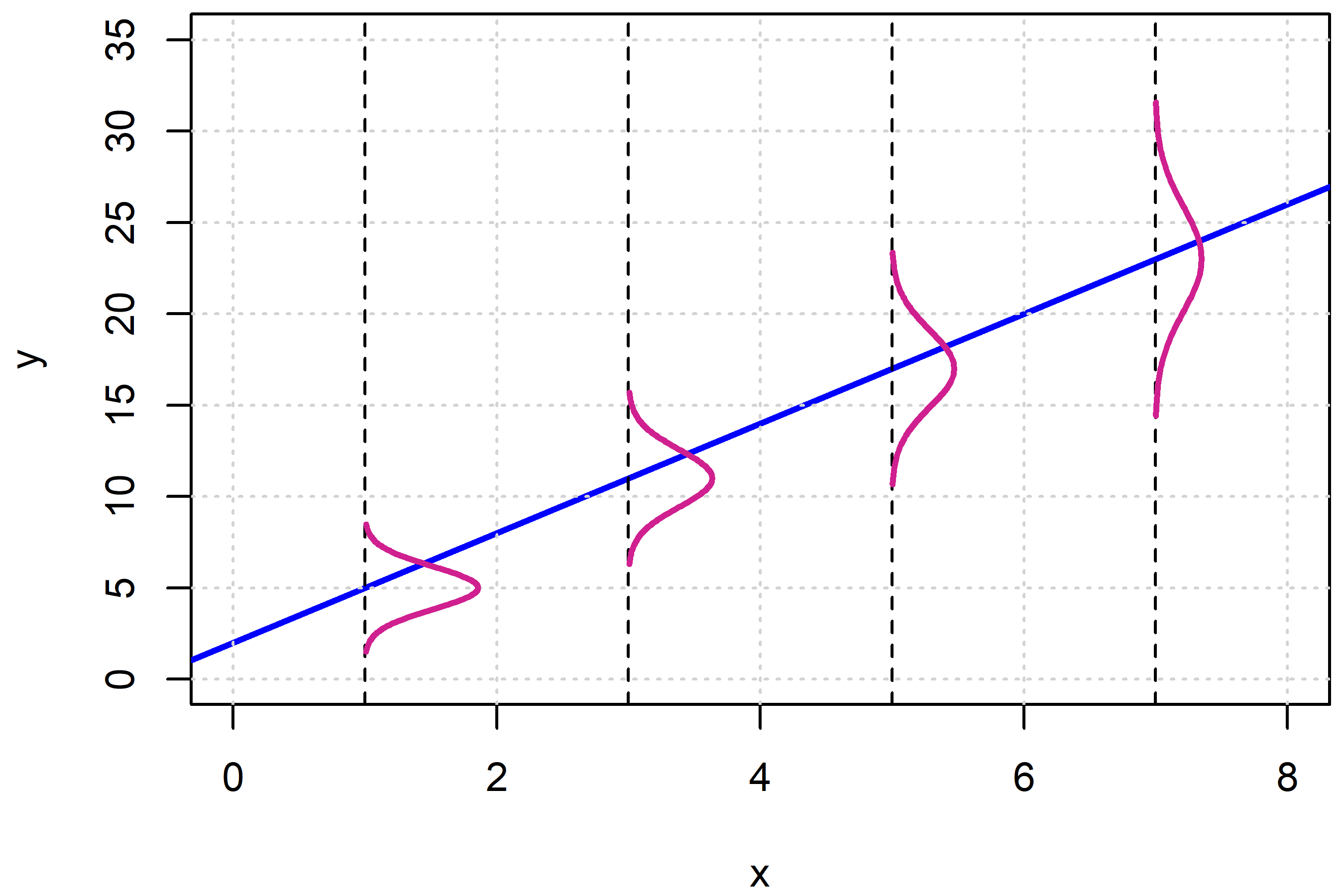


Рисунок 50. Иллюстрация гетероскедастичности для парной регрессии

Последствия гетероскедастичности для коэффициентов МНК такие же, как в общем случае нарушения гипотез (A3) и (A4):

* Гетероскедастичность сама по себе не приводит к смещенности оценки МНК.
* Оценка МНК, вообще говоря, не самая точная из линейных по несмещенных оценок (т. е. она не BLUE). Существуют более точные оценки.
* Обычная оценка матрицы ковариаций коэффициентов МНК является некорректной оценкой ковариационной матрицы коэффициентов МНК. В случае гетероскедастичности теоретическая ковариационная матрица равна  
  или
* Обычные *t*- и *F*-статистики и доверительные интервалы также являются некорректными.

## Диагностика гетероскедастичности

Самый доступный метод диагностики гетероскедастичности – это графический метод. Если дисперсии разные, то это может быть связано с тем, что величина дисперсии зависит от каких-то переменных.

Для парной регрессии

можно изобразить на графике облако наблюдений и оцененную линию регрессии. При наличии гетероскедастичности для одних значений разброс точек вокруг линии регрессии может быть шире, чем для других (рис. 51).

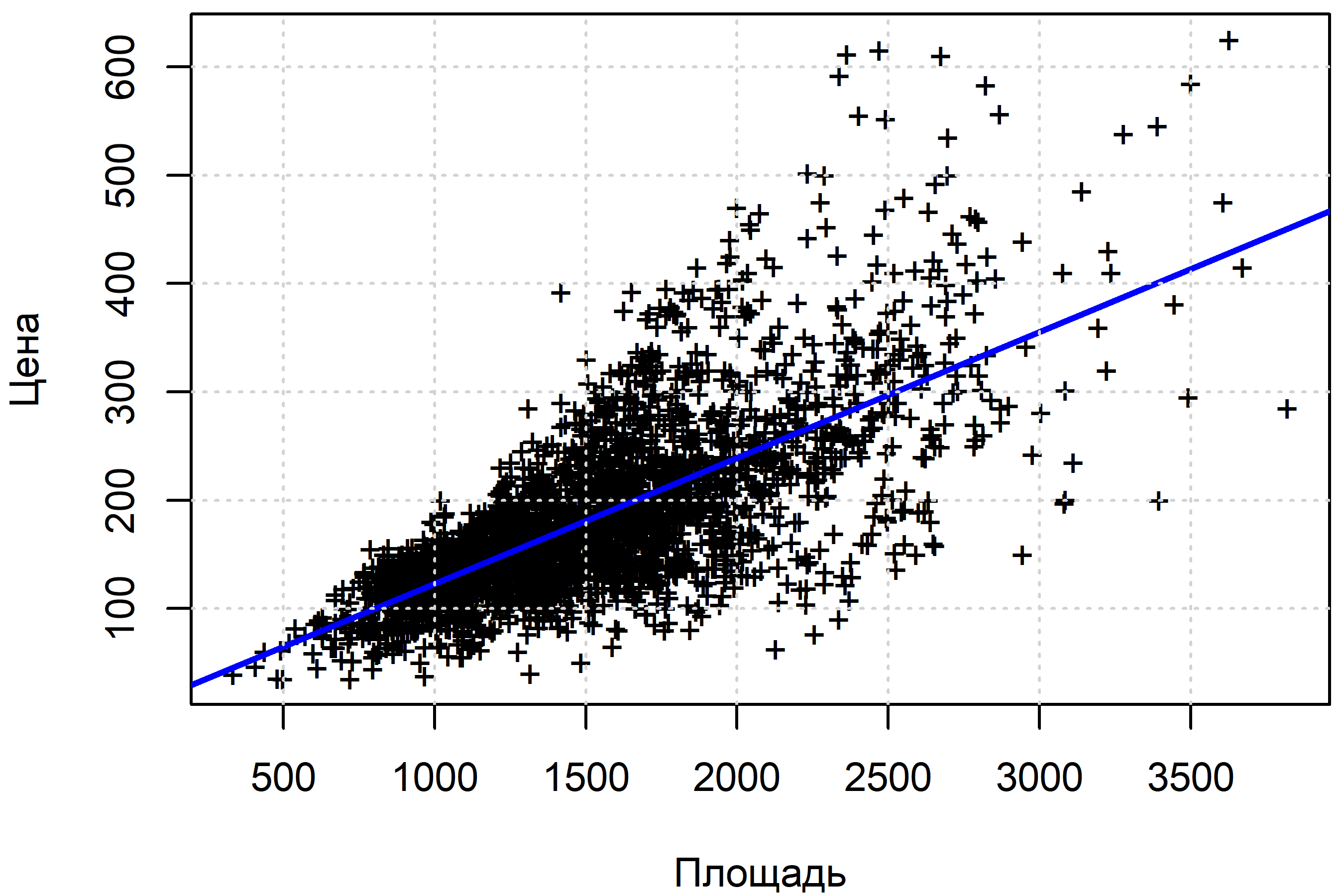


Рисунок 51. Гетероскедастичность в парной регрессии

В более общем случае удобнее использовать остатки от регрессии. Пусть мы подозреваем, что дисперсия зависит от некоторой переменной . Тогда можно использовать следующий графический метод диагностики:

- Сохраняем остатки из оцененной регрессии.

- Строим точечную диаграмму для (). В среднем остатки колеблются около нуля. Но разброс остатков вокруг нуля может быть разным для разных значений – это и есть проявление гетероскедастичности (рис. 52).

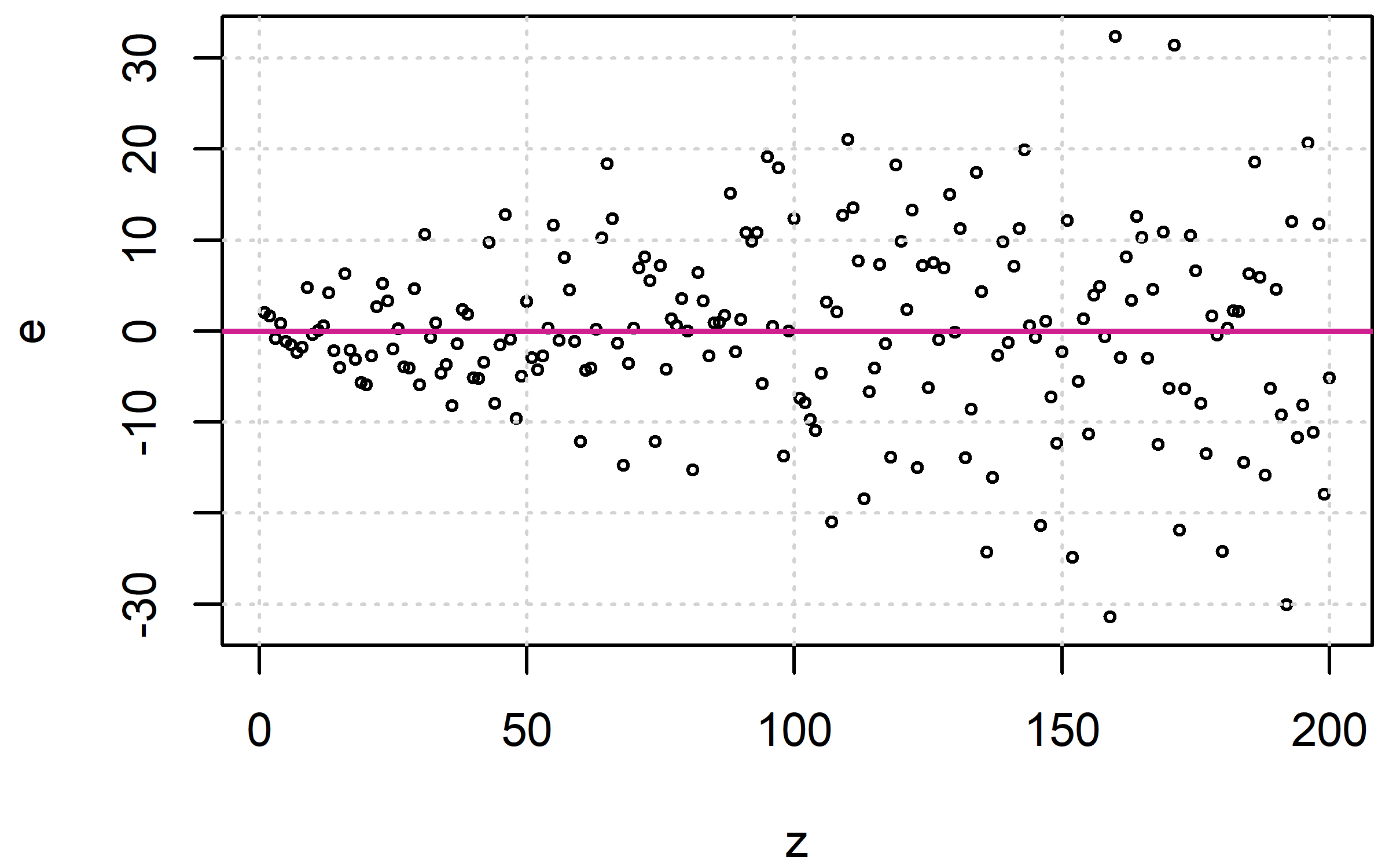


Рисунок 52. Гетероскедастичность на графике остатков

Может быть, например, так, что дисперсия тем больше, чем больше (как на графике), или тем больше, чем меньше .

При использовании данной процедуры мы исходим из того, что остатки регрессии служат некоторыми грубыми оценками для неизвестных ошибок (), поэтому проблемы с ошибками можно обнаружить, анализируя остатки.

Какую переменную взять в качестве ? В первую очередь здесь стоит обратить внимание на регрессоры, но это может быть и какая-нибудь экзогенная переменная, отсутствующая в регрессии. Можно также использовать в качестве переменной расчетные значения из регрессии .

К неформальным методам диагностики гетероскедастичности можно отнести сравнение остаточных дисперсий по некоторым подвыборкам. Пусть, например, дисперсия ошибок может зависеть от переменной . Можно разбить общий интервал значений переменной на несколько частей, в каждой из которых достаточное количество точек, и найти выборочные дисперсии остатков, попавших в каждую из частей (рис. 53). Если дисперсии сильно отличаются, то это может свидетельствовать о наличии гетероскедастичности.

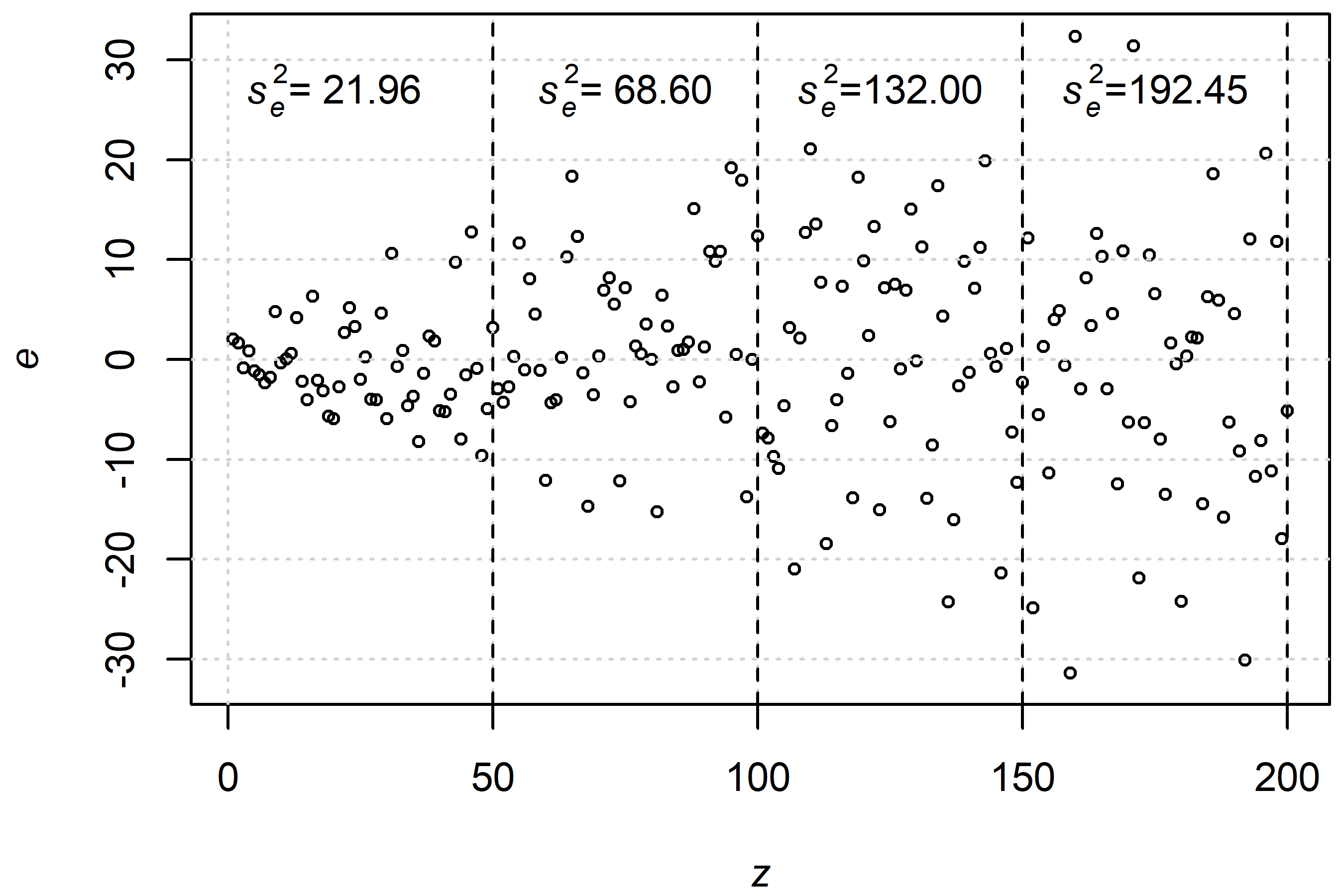


Рисунок 53. Остаточные дисперсии по подвыборкам

По этой схеме построен формальный **тест Голдфелда–Квандта** (англ. *Goldfeld-Quandt test*). Сначала наблюдения сортируются по переменной . Затем берутся наблюдений с самыми малыми значениями . По ним строится регрессия и вычисляется несмещенная остаточная дисперсия . Затем берутся наблюдений с самыми большими значениями . По ним опять строится регрессия и вычисляется несмещенная остаточная дисперсия . Можно выбрать в качестве , например, треть или четверть от общего числа наблюдений . С целью повышения мощности теста средних наблюдений в этой процедуре не используются. Если выполнены стандартные предположения модели регрессии то статистика  
имеет *F*-распределение с степенями свободы. Если статистика попадает далеко в левый или правый хвост распределения , то есть одна из дисперсий заметно больше другой, то нулевую гипотезу о гомоскедастичности отклоняем. Если не строим две отдельные укороченные регрессии, а берем остатки из полной проверяемой регрессии, отсортированные в порядке возрастания переменной , то *F*-статистика Голдфельда–Квандта будет иметь указанное *F*-распределение только приближенно, при большом количестве наблюдений.

Как почти во всех диагностических тестах нулевая гипотеза состоит в том, что модель корректна и стандартные предположения выполнены. Таким образом, *в тестах на гетероскедастичность нулевая гипотеза состоит в том, что ошибки гомоскедастичны*.

Следует понимать, что, как и в случае других диагностических тестов, сильно статистически значимая гетероскедастичность (p-значение теста очень маленькое) не означает, что гетероскедастичность количественно большая. При очень большом числе наблюдений нулевая гипотеза о гомоскедастичности будет уверенно отклоняться даже тогда, когда дисперсии по наблюдениям различаются только на несколько процентов. Поэтому о степени различности дисперсий следует судить непосредственно по отношению дисперсий , а не по p-значению.

Если дисперсия ошибки является некоторой функцией от переменной , то можно представить данную зависимость в виде регрессии:

Здесь по определению ошибка имеет нулевое математическое ожидание:

Если функция является гладкой и монотонной по , то можно приблизить ее линейной функцией. Вместо неизвестных ошибок естественно использовать остатки . Рассуждая так, приходим к следующей вспомогательной регрессии:

Если переменная в данной регрессии окажется значимой, то делаем вывод, что дисперсия ошибки зависит от . Можно использовать как *t*-статистику для переменной , так и *F*-статистику для регрессии в целом. Во вспомогательной регрессии мы проверяем гипотезу . Тем самым для исходной регрессии мы проверяем нулевую гипотезу о гомоскедастичности.

Данный подход удобен тем, что можно использовать во вспомогательной регрессии несколько объясняющих переменных:

Когда переменных несколько, следует использовать *F*-статистику для регрессии в целом, чтобы проверить значимость всех этих переменных одновременно. Статистика будет приближенно распределена как при выполнении нулевой гипотезы о гомоскедастичности. Следовательно, можно использовать соответствующее *p*-значение, которое выдают почти все компьютерные программы, рассчитывающие регрессии. *Если регрессоры оказываются значимыми, то нулевая гипотеза о гомоскедастичности отклоняется*. В качестве переменных можно, например, использовать переменные из исходной регрессии.

Как вариант, здесь можно использовать в качестве тестовой статистики , где  – коэффициент детерминации из вспомогательной регрессии. При выполнении нулевой гипотезы о гомоскедастичности статистика распределена приблизительно как . Нулевая гипотеза отклоняется, когда статистика большая.

Описанный тест обычно называют **тестом Бройша–Пейгана** (англ. *Breusch–Pagan test*). Он, фактически, исходит из того, что если дисперсия ошибки каких-то наблюдений большая, то и *квадраты остатков* для этих наблюдений в среднем будут большими.

Похожий **тест Глейзера** (англ. *Glejser test*). исходит из того, что если дисперсия ошибки каких-то наблюдений большая, то и *модули остатков* для этих наблюдений в среднем будут большими. В тесте Глейзера используется вспомогательная регрессия

Здесь также можно использовать соответствующую *F*-статистику для регрессии в целом и *p*-значение из этой статистики. Если регрессоры оказываются значимыми, то нулевая гипотеза о гомоскедастичности отклоняется. Тест Глейзера имеет теоретическое обоснование только при нормальности ошибок, поэтому его следует отнести к разряду неформальных (но полезных) диагностических процедур.

Повторим снова, что не следует судить о степени различности дисперсий по *p*-значению диагностических тестов. В случае рассматриваемых тестов Бройша–Пейгана и Глейзера можно ориентироваться на расчетные значения из вспомогательных регрессий, поскольку они дают грубые оценки дисперсий (корней из дисперсий с точностью до множителя в случае теста Глейзера). Если полученные оценки для разных наблюдений различаются во много раз, то, скорее всего, имеет место серьезная гетероскедастичность, которая приводит к заметной неэффективности оценок коэффициентов.

## Взвешенный МНК

Каким способом можно скорректировать гетероскедастичность? Пусть мы знаем ковариационную матрицу ошибок с точностью до множителя. Другими словами, знаем дисперсии ошибок с точностью до множителя:

В обозначениях предыдущей главы здесь

Матрица , такая что  
находится как диагональная матрица с элементами , т. е.  
Преобразованная регрессия для нахождения оценки ОМНК:

Дисперсии преобразованных ошибок равны  
Значит, преобразованные ошибки гомоскедастичны. При этом, как несложно увидеть, если исходные ошибки были неавтокоррелированными, то и преобразованные ошибки будут обладать этим свойством. Таким образом, преобразованная регрессия удовлетворяет стандартным предположениям (A0)–(A4). Оценка МНК из этой регрессии обладает свойством BLUE.

В преобразованной регрессии мы каждое из наблюдений делим на число, пропорциональное корню из дисперсии ошибки этого наблюдения. Заметим, что константа также подвергается преобразованию – типичным элементом преобразованного вектора из единиц будет . Поэтому, если оценивать эту регрессию с помощью обычного МНК, то компьютерная программа может выдавать предупреждение об отсутствии константы и некорректно вычислять *F*-статистику.

Данный метод оценивания называют **взвешенным МНК**. Это разновидность обобщенного МНК, поэтому будем для оценки, полученной этим методом, использовать обозначение .

Оценку взвешенного МНК можно получить из задачи минимизации **взвешенной суммы квадратов**:  
 здесь играют роль **весов** для квадратов остатков.

Даже если веса точно неизвестны и после взвешивания гетероскедастичность пропадет не полностью, но станет менее заметной, это может повысить точность оценок и т. п.

Если выразим оценку взвешенного МНК через исходные переменные, то получим следующую формулу:  
Теоретическая ковариационная матрица оценки взвешенного МНК равна

На графике (рис. 54) показана зависимость расходов на продукты питания от дохода для бельгийских домохозяйств[[6]](#footnote-6). Здесь можно заметить гетероскедастичность (и некоторую нелинейность зависимости).

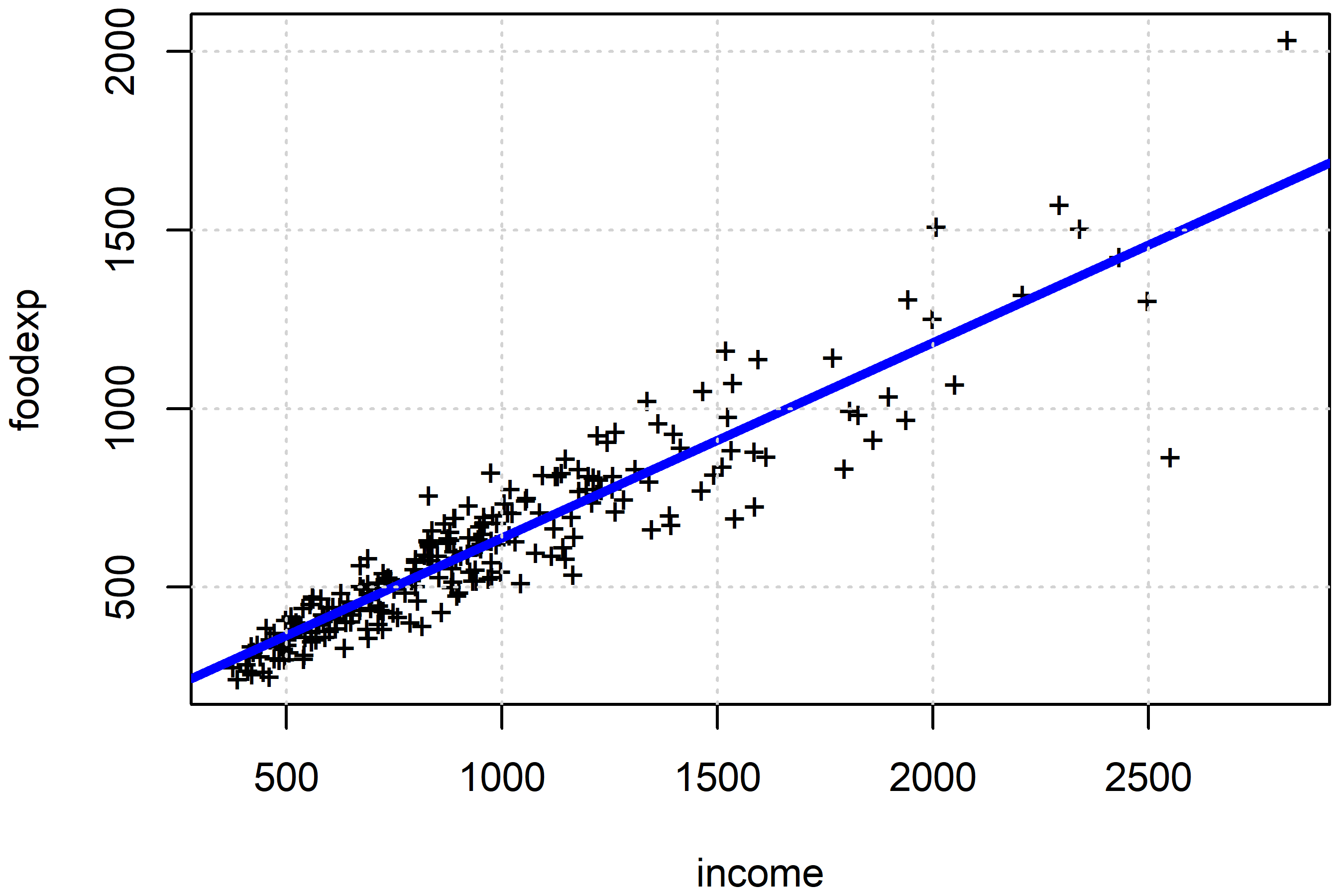


Рисунок 54. Зависимость расходов на продукты питания от дохода

График остатков от дохода подтверждает наличие гетероскедастичности. Чем больше доход, тем шире разброс остатков (рис. 55).

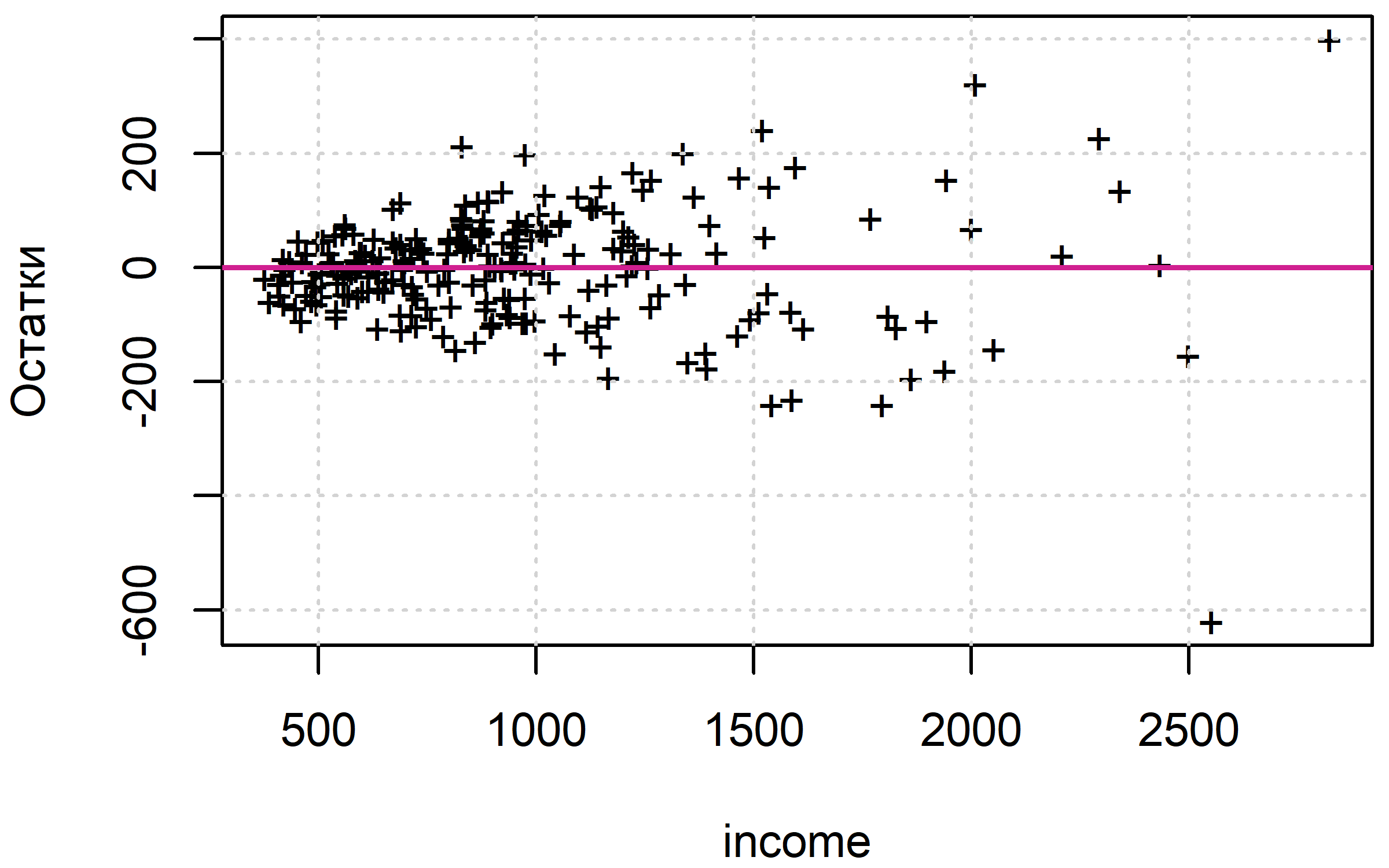


Рисунок 55. График остатков от дохода

Сделаем предположение, что среднеквадратическое отклонение ошибок примерно пропорционально доходу, то есть дисперсия примерно пропорциональна квадрату дохода. Тогда квадрат дохода можно использовать в качестве .

Обычный МНК дает уравнение  
а взвешенный МНК –  
В скобках здесь стоят стандартные ошибки. В обоих случаях это робастные стандартные ошибки, о которых речь пойдет далее. Видим, что стандартная ошибка коэффициента при доходе почти в 2 раза меньше в случае взвешенного МНК. Из-за повышения точности доверительный интервал для того же коэффициента примерно в 2 раза у́же: [0.482, 0.611] и [0.554, 0.616] соответственно.

Для диагностики гетероскедастичности во взвешенной регрессии можно использовать так называемые **нормированные остатки** (англ. *standardized residuals*), которые строятся так, что их дисперсия примерно равна 1. Нормированные остаткиво взвешенной регрессии определяются как  
где – обычные остатки, а – оценка коэффициента , получаемая из взвешенной суммы квадратов остатков:

В примере с расходами на продукты питания нормированные остатки не обнаруживают заметной гетероскедастичности (рис. 56).



Рисунок 56. График нормированных остатков из взвешенной регрессии от дохода

На практике чаще используют не взвешенный МНК, а обычный МНК. Дело в том, что мы, как правило, не знаем правильные веса . Проводить моделирование дисперсий в явном виде бывает довольно сложно. Это усложнение может быть неоправданным, если нас интересует зависимость только для математического ожидания изучаемой переменной , но не для ее дисперсии. С другой стороны, при сильной гетероскедастичности возможно существенно повысить точность оценок, как показывает вышеприведенный пример.

## Логарифмирование переменных и гетероскедастичность

Для многих положительных переменных, таких как население, стоимость, объем производства, доход и т. д. используется логарифмирование. Особенно это касается переменных, значения которых могут варьироваться в несколько раз. Причин для логарифмирования много, но одна из них состоит в том, что разброс ошибок, соответствующих таким переменным, как правило, пропорционален уровню переменных. Например, если объем производства большой, то и ошибка в соответствующем уравнении регрессии может иметь широкий разброс, а если объем производства маленький, то и разброс ошибки маленький. Это свойство реальных данных может приводить к гетероскедастичности.

Если в рассматривавшемся примере с зависимостью расходов на продукты питания от дохода построить регрессию в логарифмах, то гетероскедастичность практически пропадает (рис. 57). Кроме того, в отличие от исходной модели, здесь нет сколько-нибудь заметной нелинейности зависимости.

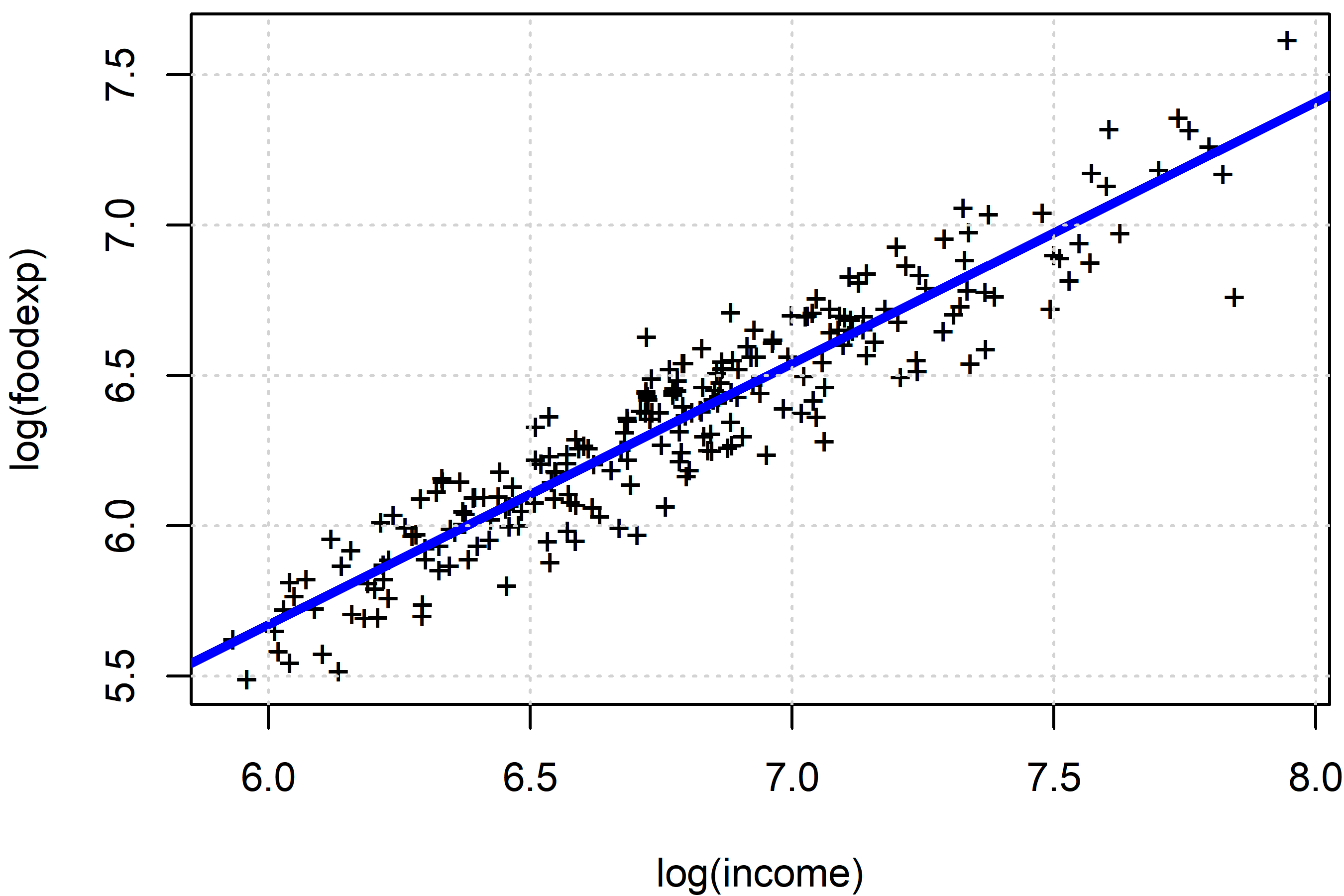


Рисунок 57. Зависимость логарифма расходов на продукты питания от логарифма дохода

Даже если после логарифмирования гетероскедастичность пропадет не полностью, но станет менее заметной, все равно имеет смысл применять этот прием для улучшения соответствия модели имеющимся данным, улучшения прогнозов и для устранения других проблем (несоответствия функциональной формы, сильной ненормальности ошибок и т. д.).

## Робастная оценка для ковариационной матрицы коэффициентов МНК

Как уже говорилось, гетероскедастичность приводит к некорректности обычной оценки матрицы ковариаций коэффициентов МНК. Поэтому требуется использовать какую-то корректную оценку. Самая простая оценка, не требующая информации о форме гетероскедастичности – это **оценка Уайта** (англ. *White’s estimator*). Также ее называют оценкой Хьюбера–Уайта (англ. *Huber–White estimator*).

Поскольку остатки – это приближения для ошибок (), то  
Если в теоретическую ковариационную матрицу оценки обычного МНК  
вместо подставим , то получим состоятельную оценку. Это и есть оценка Уайта:  
где  
В развернутом виде

Это **робастная к гетероскедастичности оценка** (устойчивая к гетероскедастичности). Она *относится к классу сэндвич-оценок*, поскольку используется трехслойная матрица.

Как использовать оценку Уайта и аналогичные ей сэндвич-оценки? Корни диагональных элементов в робастной ковариационной матрице – это так называемые **робастные стандартные ошибки** (или **состоятельные при гетероскедастичности стандартные ошибки**, англ. *Heteroskedasticity-consistent standard errors*, *HC standard errors*). Эти стандартные ошибки можно использовать для вычисления робастных *t*-статистик. Данные статистики вычисляются по обычной формуле  
но здесь робастная. Аналогично, *F*-статистику для линейного ограничения можно рассчитать по стандартной формуле  
с использованием в ней робастной ковариационной матрицы . (Представление *F*-статистики через суммы квадратов остатков здесь уже не работает!) Для данных *t*-статистик и *F*-статистик можно использовать обычные таблицы распределений и , но это только **асимптотические приближения**, работающие при достаточно большом количестве наблюдений.

Некоторые авторы *рекомендуют по умолчанию использовать робастную к гетероскедастичности оценку* при моделировании данных, в которых нет автокорреляции, потому что при этом не требуется производить диагностику гетероскедастичности и в целом это, в определенном смысле, безопасный подход.

Следует обратить особое внимание на гетероскедастичность при проведении тестов на функциональную форму. Вообще гетероскедастичность и ошибки в функциональной форме – это взаимосвязанные проблемы (см. напр. выше обсуждение логарифмирования). *При тестировании функциональной формы лучше использовать робастные к гетероскедастичности статистики*.

## Контрольные вопросы

1. Что такое гетероскедастичность?
2. В каком случае гетероскедастичность ухудшает оценки параметров регрессии (случай гетероскедастичности «с негативными последствиями») ?
3. Вычислите ковариационную матрицу вектора взвешенного МНК-оценок.
4. Проверьте для парной регрессии с гетероскедастичностью ошибок, что дисперсия оценки параметра b, полученная ОМНК, меньше, чем дисперсия МНК-оценки.
5. Как можно проверить наличие гомо-или гетероскедастичность ошибок?
6. Что нужно сделать со значениями переменных в регрессии в случае гетероскедастичности «с негативными последствиями»?
7. Какая матрица используется в случае гетероскедастичности «с негативными последствиями»?
8. Какое преобразование матрицы наблюдений перед оценкой регрессии полезно сделать, если среднеквадратические отклонения ошибок регрессии пропорциональны какому-либо фактору?
9. Перечислите основные причины для логарифмирования переменных при построения регрессионного управления.
10. Что такое оценка Уайта и как использовать эту оценку?

## Экзаменационные вопросы

1. Гетероскедастичность: Определение. Последствия гетероскедастичности. Диагностика (тесты на гетероскедастичность, графические методы диагностики).
2. ОМНК при гетероскедастичности (взвешенный МНК). Оценка Уайта ковариационной матрицы оценок обычного МНК.
3. Логарифмирование переменных и гетероскедастичность. Оценка Уайта ковариационной матрицы оценок обычного МНК.

## Литература

1. Вербик М. Путеводитель по современной эконометрике. – М.: Научная книга,. 2008. [стр.137-164 ]
2. Магнус Я. Р., Катышев П. К., Пересецкий А. А. Эконометрика. Начальный курс: – Учебник. – 6-е изд., перераб. и доп. – М.: Дело, 2004. [стр.167-203 ]
3. Суслов В. И., Ибрагимов Н. М., Талышева Л. П., Цыплаков А. А. Эконометрия.– Новосибирск: СО РАН, 2005. [стр.258-264 ]
4. Картаев Ф. С. Введение в эконометрику. Учебник : – Москва: Проспект, 2019. [стр. 143-149, 159-166 ]

# Лекция: – Автокорреляция ошибок

## Некоторые понятия из теории временных рядов

**Временной ряд** – это переменная (например, ), в которой наблюдения () соответствуют идущим по порядку моментам времени. Обычно наблюдения идут с приблизительно равными интервалами – дни, месяцы, годы и т. п.

**Лаг** ряда – это тот же ряд, но сдвинутый назад во времени. *k*-й лаг – это переменная, состоящая из элементов (), т. е. переменная вида

(В случае, если при использовании лагов «доисторические» данные отсутствуют, то все ряды приходится соответствующим образом укорачивать. Для 1-го лага – брать наблюдения , и т. д.)

Для временного ряда, так же как и для обычной переменной, можно рассчитать выборочные характеристики – среднее, выборочную дисперсию, и т. д. Но есть и особые характеристики.

**Автоковариация** *k*-го порядка – это ковариация между рядом и его -м лагом . С учетом того, что ряды будут укороченными (), а средние и по смыслу должны быть одинаковыми (берем ), выборочная автоковариация ряда рассчитывается как  
где – центрированный ряд.

**Автокорреляция** *k*-го порядка – это корреляция между рядом и его -м лагом . Выборочная автокорреляция рассчитывается как  
Здесь мы исходим из того, что дисперсии и по смыслу должны быть одинаковыми, и поэтому берем в качестве их выборочных дисперсий одинаковую оценку

Автокорреляцию во временных рядах также называют **серийной корреляцией** (англ. *serial correlation*).

Наблюдается тенденция, что чем больше лаг , тем автокорреляция слабее. (Одно из возможных исключений из общей тенденции – это сезонная автокорреляция. Например, для помесячных данных может наблюдаться заметная автокорреляция на лагах , и т. д. и она может быть больше по величине, чем автокорреляция на лагах , т. д.)

Если отвлечься от сезонности, то чаще всего в экономических временных рядах наблюдается сильная положительная автокорреляция 1-го порядка ( близко к 1).

Как выглядят типичные ряды с автокорреляцией 1-го порядка различного знака и без автокорреляции можно увидеть из рис. 58, рис. 59 и рис. 60.

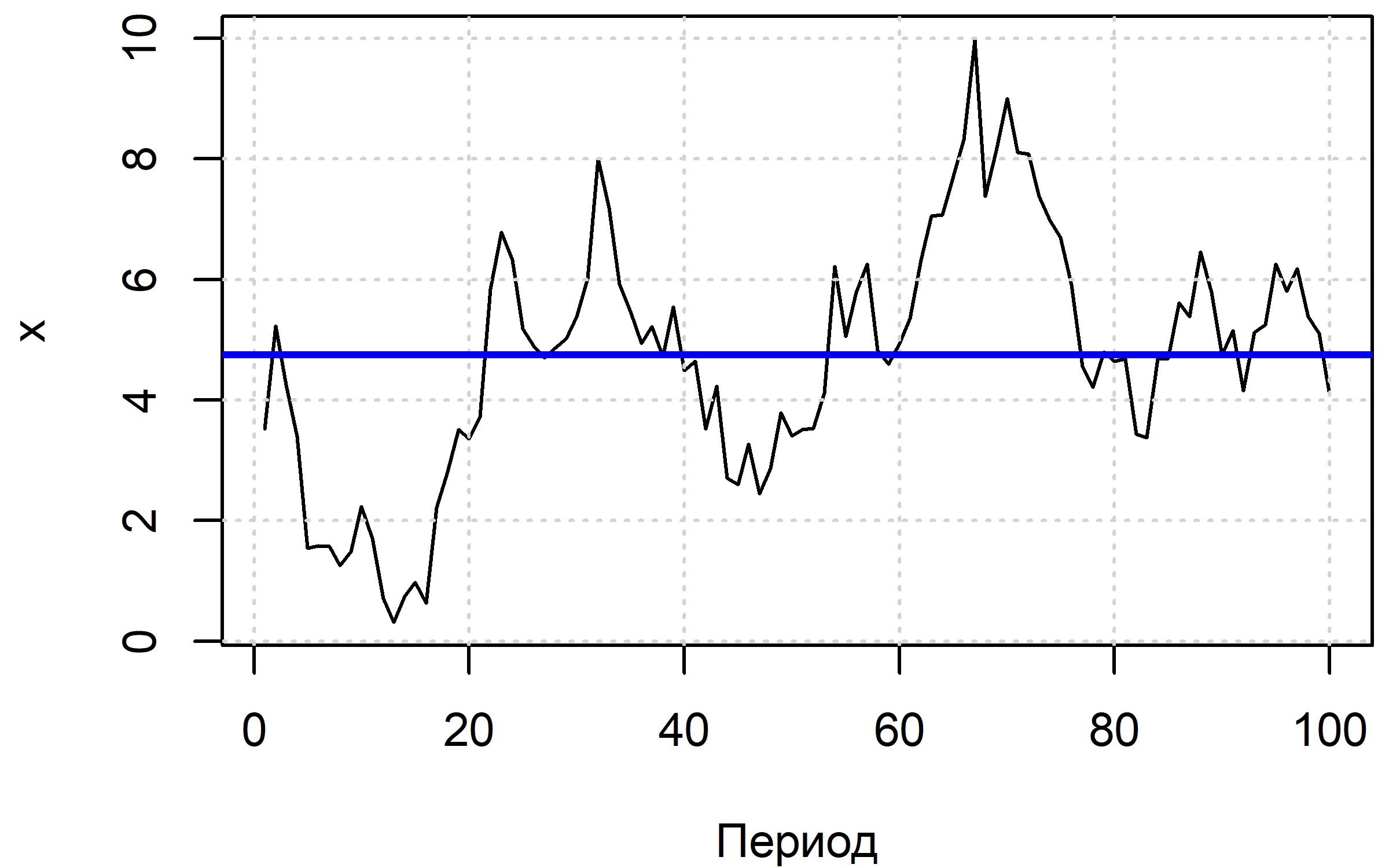


Рисунок 58. Ряд с

При (рис. 58) на графике видны колебания вокруг линии среднего в виде волн. Имеются периоды, когда подряд идущие наблюдения выше среднего и ниже среднего.

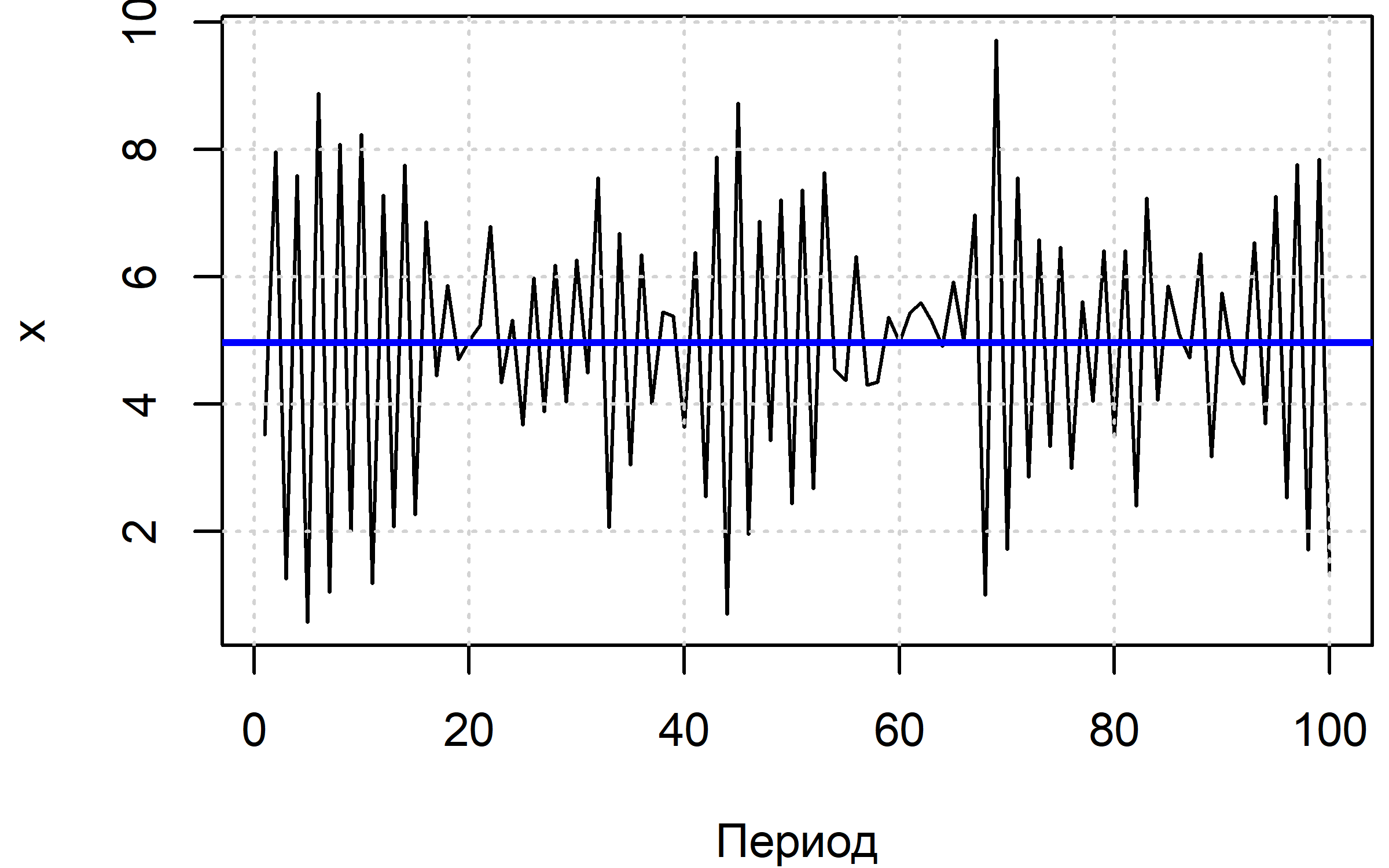


Рисунок 59. Ряд с

При (рис. 59) на графике видны колебания вокруг линии среднего в виде «пилы». Если значение ниже среднего, то следующее значение обычно выше среднего и наоборот.

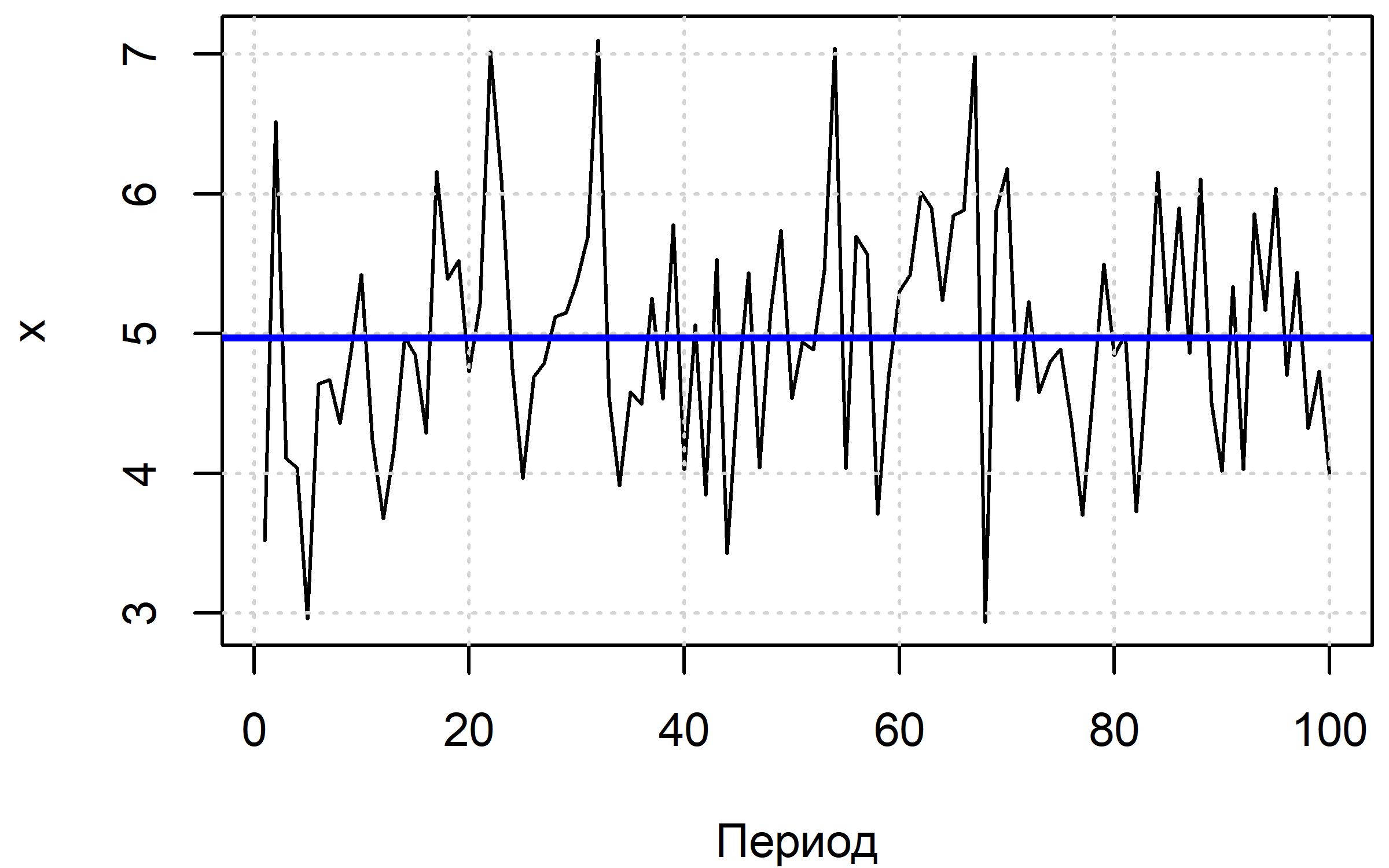


Рисунок 60. Ряд с

При (рис. 60) не заметно каких-то особых закономерностей. Ряд выглядит как чисто случайный шум.

Одна из моделей, позволяющих моделировать автокорреляцию во временных рядах – это **авторегрессия**. Это регрессия, в которой зависимая переменная зависит от своих лагов.

Например, авторегрессия 1-го порядка (обозначается AR(1)):

Временной ряд, рассматриваемый как многомерная случайная величина, называют **случайным процессом**. Среди случайных процессов выделяют **стационарные**. Если говорить неформально, то это такие процессы , свойства которых остаются неизменными с течением времени. В частности, у них неизменное математическое ожидание , неизменная дисперсия , а автоковариации и автоковариации зависят от лага , но не от индекса времени .

**Белый шум** – это стационарный процесс с нулевым математическим ожиданием и не автокоррелированный (т. е. такой, у которого все автоковариации равны нулю).

## Автокорреляция ошибок – определение и последствия

В целом **автокорреляция** (нарушение предположения, которое мы называли (A4)), состоит в том, что матрица ковариаций вектора ошибок  
*не является диагональной*. Это означает, что найдутся номера наблюдений и , где , такие что ковариация между ошибками не равна нулю:

Для временных рядов рассматривают ковариацию между рядом ошибок и его -м лагом , т. е. автоковариацию -го порядка. Ряд называется имеющим автокорреляцию *k*-го порядка, если

В предположении стационарности ряда ошибок можем определить для него **теоретическую автокорреляцию *k*-го порядка** (теоретическую **серийную корреляцию** *k*-го порядка) как  
где (Для стационарного ряда имеет место гомоскедастичность.)

Таким образом, для стационарных временных рядов нарушение предположения (A4) об отсутствии автокорреляции проявляется в том, что  
при некотором . По указанным выше причинам чаще всего мы можем ожидать, что в ошибках имеется положительная автокорреляция 1-го порядка:

Кроме серийной автокорреляции в экономических данных проявляются и другие разновидности автокорреляции, которые мы обсудим позже.

Последствия наличия автокорреляции в ошибках регрессии для оценок МНК в целом такие же, как в общем случае:

* Автокорреляция сама по себе не приводит к смещенности оценки МНК. (Но есть дополнительная проблема, о которой ниже.)
* Оценка МНК, вообще говоря, не самая точная из линейных по несмещенных оценок (не BLUE). Существуют более точные оценки.
* Обычная оценка матрицы ковариаций коэффициентов МНК является некорректной оценкой ковариационной матрицы коэффициентов МНК.
* Обычные *t*- и *F*-статистики также являются некорректными.

По поводу несмещенности оценок МНК следует сделать оговорку. Мы пока что используем предположение (A2) об экзогенности регрессоров в очень жестком виде – считаем что матрица регрессоров является детерминированной. Но ясно, что это жесткое предположение *не выполнено, если рассматривается авторегрессия*, и в принципе, если среди регрессоров присутствуют лаги () зависимой переменной .

Если среди регрессоров присутствуют лаги зависимой переменной, то оценки МНК в общем случае являются смещенными, но они при отсутствии автокорреляции ошибок и при выполнении ряда других предположений являются асимптотически несмещенными (т. е. смещение пропадает в пределе при большом количестве наблюдений) и состоятельными.

Если же в регрессии есть лаги зависимой переменной и есть автокорреляция, то оценки МНК в общем случае будут асимптотически смещенными и не состоятельными, т. е. непригодными к использованию.

## Диагностика серийной корреляции

Как и в случае гетероскедастичности при диагностике автокорреляции используется тот факт, что остатки регрессии напоминают ошибки .

Для временных рядов мы исследуем серийную корреляцию остатков, то есть корреляцию между рядом остатков и лагами остатков .

Для проведения графической диагностики строится график остатков по номеру наблюдения (то есть по времени). В типичном случае положительной автокорреляции 1-го порядка на графике наблюдаются «волны» и достаточно длительные периоды с подряд идущими положительными и отрицательными остатками.

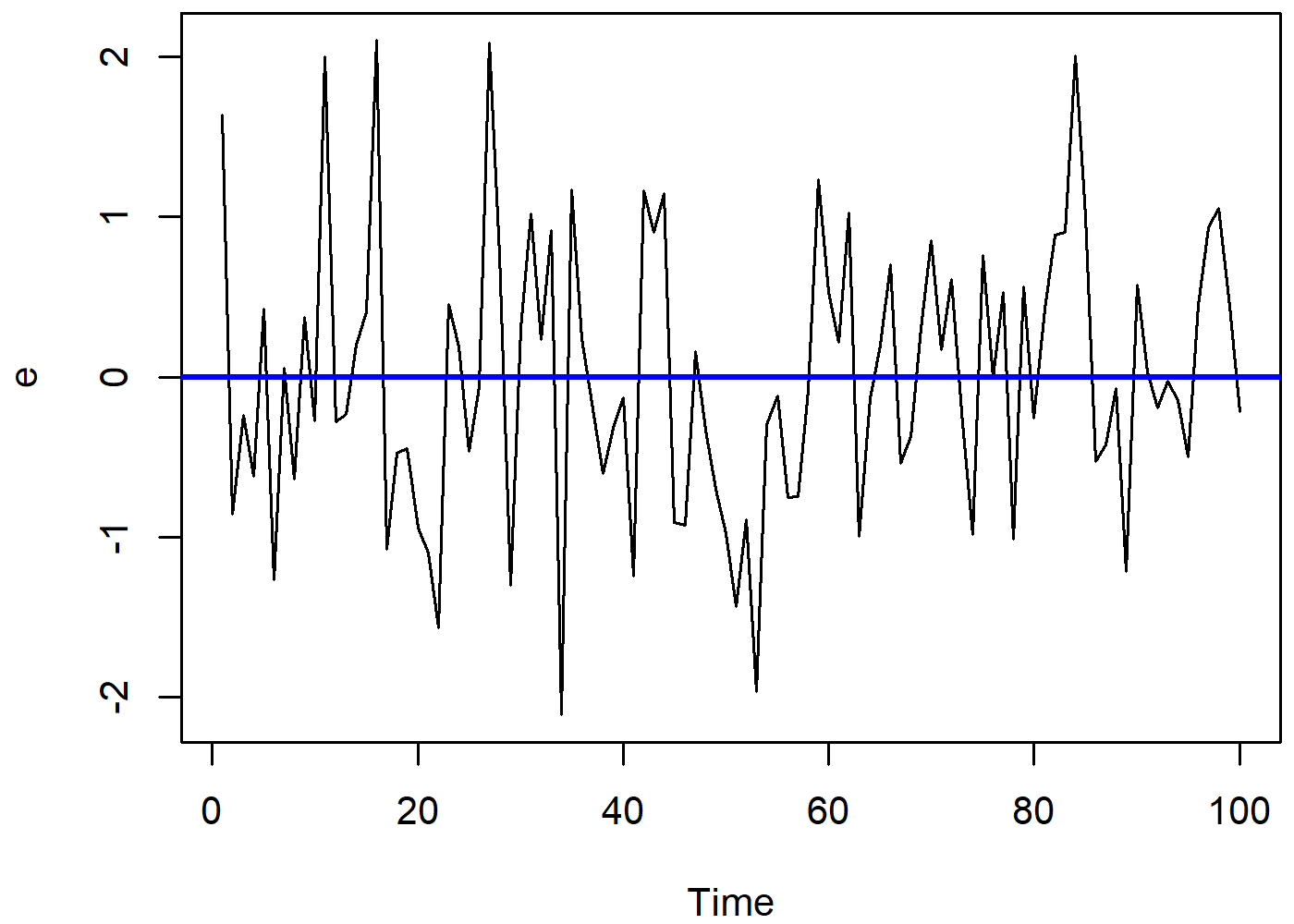
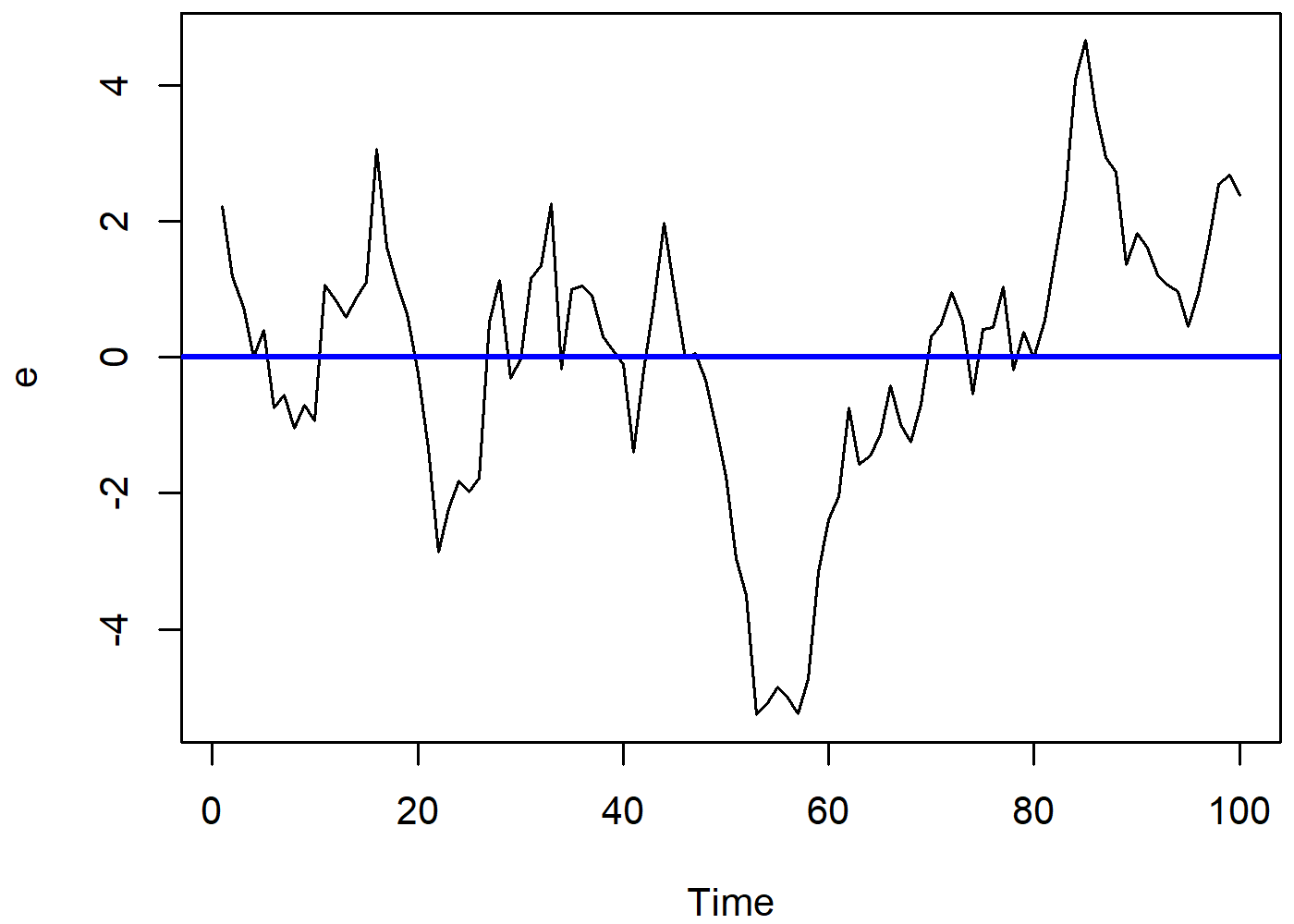
а) б)

Рисунок 61. а) Нет следов автокорреляции остатков; б) сильная положительная автокорреляция 1-го порядка

Кроме того, для неформальной диагностики можно использовать коэффициенты автокорреляции для остатков. Выборочный коэффициент автокорреляции *k*-го порядка vможно рассчитать по формуле

(Остатки здесь уже центрированные, их не требуется центрировать.) Можно рассчитать несколько первых коэффициентов () посмотреть на них в виде таблицы или графика (это так называемая **автокорреляционная функция**, **АКФ**).

Формально проверить наличие автокорреляции можно построив авторегрессию для остатков:

Отсутствующие «доисторические» значения остатков берем равными нулю или укорачиваем регрессию на первых наблюдений. В этой вспомогательной регрессии мы проверяем гипотезу при помощи *F*-статистики. Тем самым мы по факту *проверяем нулевую гипотезу об отсутствии автокорреляции*.

Данный тест, как и остальные рассматриваемые здесь тесты, являются асимптотическим, т. е. обосновывается переходом к пределу по количеству наблюдений (). Поэтому *F*-распределение при нулевой гипотезе будет иметь место только приближенно.

Для случая здесь можно использовать *t*-статистику для проверки . Оценка коэффициента будет приближенно равна коэффициенту автокорреляции 1-го порядка , поэтому по смыслу данная процедура очень близка к проверке значимости отличия от нуля. Если отрицательная автокорреляция рассматривается как маловероятная, то такой тест можно использовать как односторонний, т. е. отклонять нулевую гипотезу при попадании в правый хвост распределения Стьюдента.

(Для случая есть еще и *известный тест Дарбина*—*Уотсона*, но он требует использования специальных таблиц.)

В **тесте Бройша**—**Годфрея** (англ. *Breusch–Godfrey test*) используется примерно такая же вспомогательная регрессия, но к правой части добавляются еще регрессоры из проверяемой модели.

При этом в левой части можно вместо остатков поставить зависимую переменную – от этого *F*-статистика не поменяется:

Таким образом, *тест Бройша–Годфрея можно проводить как тест на добавление в исходную регрессию лагов остатков*. Проверяя по *F*-статистике нулевую гипотезу о том, что лаги не нужны в регрессии () мы, тем самым, проверяем, нулевую гипотезу об отсутствии автокорреляции. *F*-статистика здесь приближенно распределена как при нулевой гипотезе.

(Отметим, что если оставить в левой части регрессии остатки, то можно использовать в качестве тестовой статистики из вспомогательной регрессии, умноженный на количество наблюдений . Такая статистика приближенно имеет распределение )

Преимущество теста Бройша—Годфрея по сравнению с авторегрессией для остатков (и тестом Дарбина—Уотсона) состоит в том, что он дает корректные результаты даже в том случае, когда среди регрессоров имеются лаги зависимой переменной.

## Что делать в случае серийной корреляции?

Как уже упоминалось, существуют модели, которые могут описать автокорреляцию во временном ряду. Если ряд ошибок описывается подобной моделью, то можно использовать *обобщенный МНК*, чтобы исправить данную проблему.

Например, пусть ошибка подчиняется процессу авторегрессии первого порядка AR(1):  
где – белый шум с дисперсией . Можно преобразовать регрессию  
к виду  
где , , , . Ошибка здесь не автокоррелирована и гомоскедастична. Когда параметр неизвестен, можно использовать оценку , где – коэффициент автокорреляции остатков 1-го порядка. Это будет некоторый вариант доступного ОМНК.

Однако в случае, когда диагностический тест указывает на наличие автокорреляции, коррекция указанным способом – это, скорее всего, неправильный подход. Дело в том, что причиной автокорреляции может быть неправильная динамическая спецификация модели. Если мы механически используем обобщенный МНК, вместо того, чтобы привести модель к более подходящему виду, то мы, тем самым, можем спрятать проблему, вместо того, чтобы ее принципиально решить. Существуют различные **динамические модели регрессии**, и какую из них использовать – это отдельный вопрос, который требует серьезного анализа.

Альтернативный подход к проблеме серийной корреляции – использование оценки обычного МНК и робастной оценки для ее ковариационной матрицы. Обычно здесь используются **состоятельная при автокорреляции и гетероскедастичности оценка** (англ. *heteroskedasticity and autocorrelation consistent estimator*, *HAC*). Самая известная сэндвич-оценка такого рода – это **ковариационная матрица Ньюи**—**Уэста** (англ. *Newey–West covariance matrix estimator*):  
где , при и при .

## Другие виды автокорреляции. Пространственная автокорреляция

Автокорреляция – это просто ситуация, когда матрица не является диагональной, так что про автокорреляцию в самом общем случае мало что можно сказать. Следует наложить какую-то структуру на ошибки, чтобы сделать анализ более конкретным.

Предположим, что пары наблюдений связаны между собой каким-то образом и существует показатель , который отражает, насколько сильно наблюдение связано с наблюдением . Данные коэффициенты (их называют весами) можно расположить в виде **матрицы весов** . Диагональные элементы следует считать равными нулю. Вот несколько примеров:

Пример 1. Наблюдения упорядочены по определенному принципу (например по времени) и , если наблюдение непосредственно предшествует наблюдению , т. е. . Этот пример соответствует случаю серийной корреляции 1-го порядка для временных рядов. Для 5 наблюдений соответствующая матрица весов имеет вид

Пример 2. Наблюдения являются узлами в некоторой сети, и между парами узлов могут существовать связи (например, знакомство для людей, соседство для регионов страны). На рис. 62 для 5 наблюдений изображена сеть в виде графа. Рядом показана матрица весов.

2

1

3

4

5

Рисунок 62

Пример 3. Наблюдения – это пункты на географической карте. Каждая пара пунктов находится между собой на некотором расстоянии (это может быть, например, расстояние по автодорогам) и веса равны обратным расстояниям, т. е. .

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 2 | 3 | 4 | 5 |  |  |
| 1 | 31 км | 8 км | 31 км | 23 км |  |
| 2 |  | 24 км | 11 км | 15 км |  |
| 3 |  |  | 15 км | 22 км |  |
| 4 |  |  |  | 22 км |  |

Пример 4. Наблюдения разбиты на группы; , если наблюдения и из одной группы, и иначе. Например, наблюдения 1, 2, 3 из группы , а наблюдения 4, 5 из группы . Матрица весов имеет вид

Для диагностики автокорреляции требуется предположить, как именно каждая ошибка потенциально может быть связана со всеми остальными ошибками в векторе . Здесь можно использовать веса из матрицы . А именно, мы можем сопоставить ошибки и линейную комбинацию остальных ошибок , где – -я строка матрицы (поскольку , то не входит в ). Вместо ошибок , как обычно, приходится использовать остатки .

Неформальный графический способ диагностики состоит в изучении точечной диаграммы . Если, например, облако точек имеет положительный наклон, то это может служить признаком того, что ошибки близко связанных наблюдений коррелируют между собой (см. рис. 63).

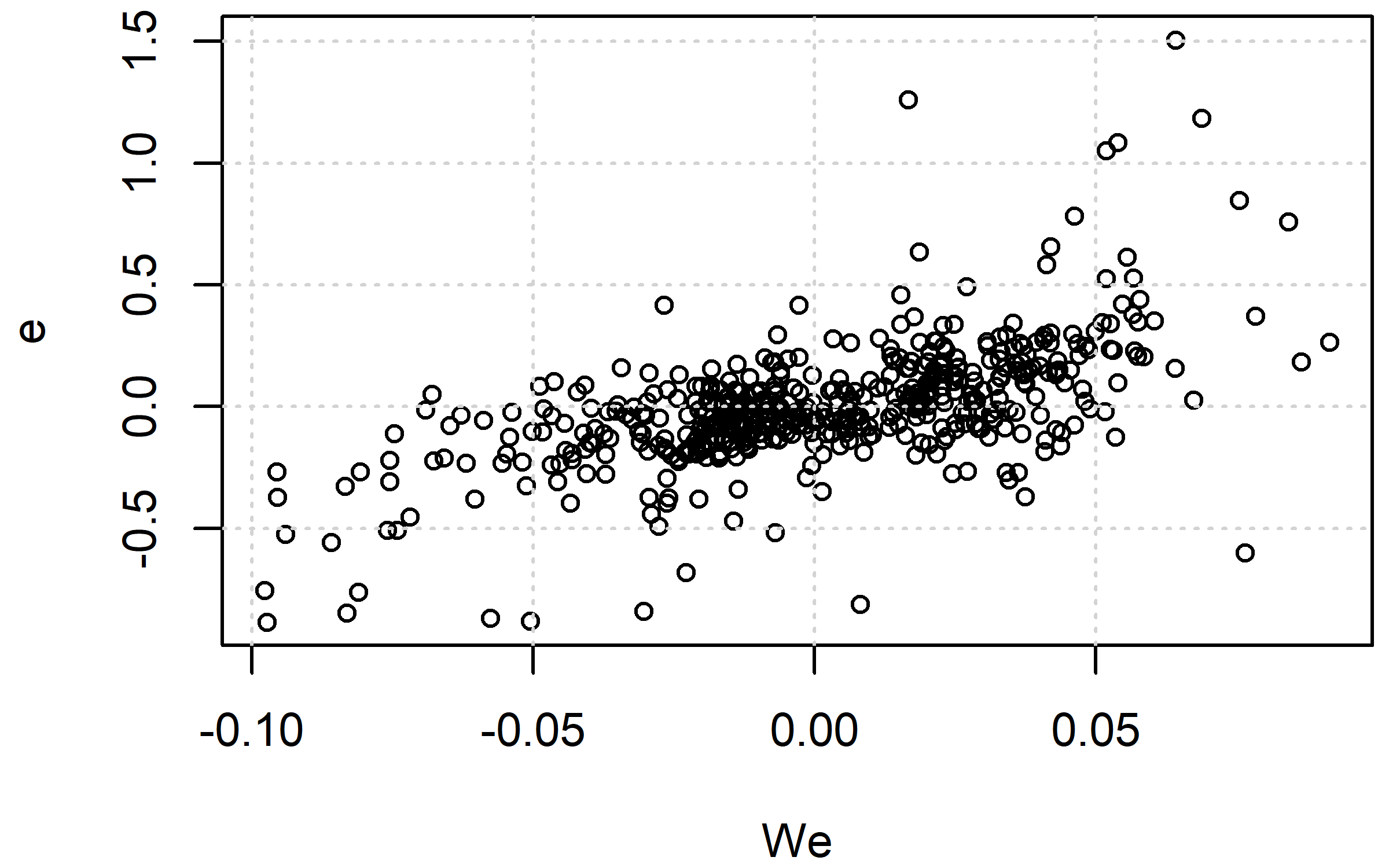


Рисунок 63

Для **пространственных данных** имеется выборочный показатель **пространственной автокорреляции** – так называемый **индекс Морана** ( Морана). Если матрица нормирована так, что сумма элементов каждого ряда равна 1, то индекс Морана на основе МНК-остатков регрессии можно рассчитать по формуле  
По конструкции этот показатель похож на автокорреляцию первого порядка остатков для временных рядов. Если автокорреляции в ошибках нет, то индекс близок к нулю.

Индекс Морана можно использовать для формального теста на автокорреляцию. При некоторых предположениях о виде распределения ошибок можно вычислить его математическое ожидание и дисперсию . Выражения для них достаточно сложные (см. Cliff, Ord, 1973), поэтому мы их здесь не приводим. Статистика  
при нулевой гипотезе об отсутствии автокорреляции в ошибках имеет приближенно стандартное нормальное распределение. Если отрицательная автокорреляция маловероятна, то можно использовать односторонний вариант этой статистики, то есть отклонять нулевую гипотезу при больших положительных значениях.

Если здесь использовать аппроксимацию и , то получится статистика LM-ERR, предложенная в Burridge (1980). Распределение при нулевой гипотезе тоже приближенно стандартное нормальное.

Если с помощью некоторой матрицы весов обнаружена автокорреляция, то дальше можно как-то учесть это и изменить модель. Однако как почти всегда с диагностикой эконометрических моделей, направление изменения не очевидно. Наличие пространственной автокорреляции, скорее всего, означает, что требуется использовать методы и модели такого раздела эконометрии как **пространственная эконометрия**.

## Контрольные вопросы

1. Что такое автокорреляция во временных рядах?
2. Как выглядят типичные ряды с автокорреляцией 1-го порядка различного знака и без автокорреляции
3. Что такое автокорреляция ошибок?
4. Что происходит с оценками параметров регрессии при автокорреляции ошибок?
5. Как можно проверить отсутствие автокорреляции ошибок при построении статистической регрессионной модели ?
6. В чем преимущество теста Бройша—Годфрея по сравнению с тестом Дарбина—Уотсона?
7. Что делать в случае серийной корреляции?
8. Что такое ковариационная матрица Ньюи—Уэста, и как эта матрица используется, если матрица ковариации ошибок не диагональна с равными элементами по диагонали?
9. Объясните понятие пространственной автокорреляции.
10. Приведите примеры пространственной автокорреляции.
11. Индекс Морана – что это такое и для чего используется?

## Экзаменационные вопросы

1. Понятие лага. Понятие автоковариации и автокорреляции. Определение автокорреляции ошибок. Последствия автокорреляции. Причины автокорреляции. Диагностика (тесты на автокорреляцию, Дарбина-Уотсона, Бройша–Годфрея, графические методы диагностики).
2. Понятие авторегрессии первого порядка (AR(1)) в ошибках. ОМНК в условиях AR(1)-ошибки. Ньюи-Уэст ковариационной матрицы оценок обычного МНК.
3. Пространственная автокорреляция : определения, примеры, диагностика.

## Литература

1. Вербик М. Путеводитель по современной эконометрике. – М.: Научная книга,. 2008. [стр.164-183 ]
2. Суслов В. И., Ибрагимов Н. М., Талышева Л. П., Цыплаков А. А. Эконометрия.– Новосибирск: СО РАН, 2005. [стр.265-269 ]
3. Картаев Ф. С. Введение в эконометрику. Учебник : – Москва: Проспект, 2019. [стр. 163-165 ]
4. Andrews, D. W. K. "Heteroskedasticity and autocorrelation consistent covariance matrix estimation". *Econometrica*. 59 (1991, 3). 817–858.
5. Burridge, P., On the Cliff-Ord Test for Spatial Autocorrelation, *Journal of the Royal Statistical Society B*, 42 (1980), 107–108.
6. Cliff, A. D. and Ord, J. K. *Spatial Autocorrelation*. London: Pion, 1973.
7. Davidson R., MacKinnon J. G. *Econometric Theory and Methods*, Oxford University Press, 2003. (7.7 Testing for Serial Correlation).

# Лекция: – Отсутствие нормальности ошибок. Выбросы

## Диагностика отсутствия нормальности и наличия выбросов

Обычно отсутствие нормальности в регрессии проявляется в том, что распределение ошибок имеет более **толстые хвосты** (другие термины – тяжелые, длинные хвосты), чем нормальное. Это приводит к тому, что в ошибках наблюдаются **выбросы** – аномальные наблюдения, которые по модулю больше нескольких среднеквадратических отклонений. Кроме того, в распределении ошибок достаточно часто наблюдается **скошенность** (асимметрия), когда один из хвостов существенно толще, чем другой, и в нем наблюдается повышенная доля выбросов. Как правило, это правая асимметрия (также называемая положительной).

Как и в других случаях регрессионной диагностики, диагностика отсутствия нормальности в ошибках регрессии проводится на основе изучения выборочного распределения остатков.

Самые доступные и наглядные методы диагностики – графические.

Можно, в частности, использовать график остатков по какой-либо переменной, например, просто по номеру наблюдения, по одному из регрессоров или по расчетным значениям зависимой переменной. Выбросы – это наблюдения, которые сильно отклоняются по сравнению с остальными остатками от нуля (на несколько стандартных ошибок). Здесь удобнее использовать **нормированные остатки** (англ. *standardized residuals*), т. е. остатки деленые на стандартную ошибку регрессии :  
поскольку их среднеквадратическое отклонение примерно равно 1, и можно непосредственно ориентироваться на их абсолютную величину.

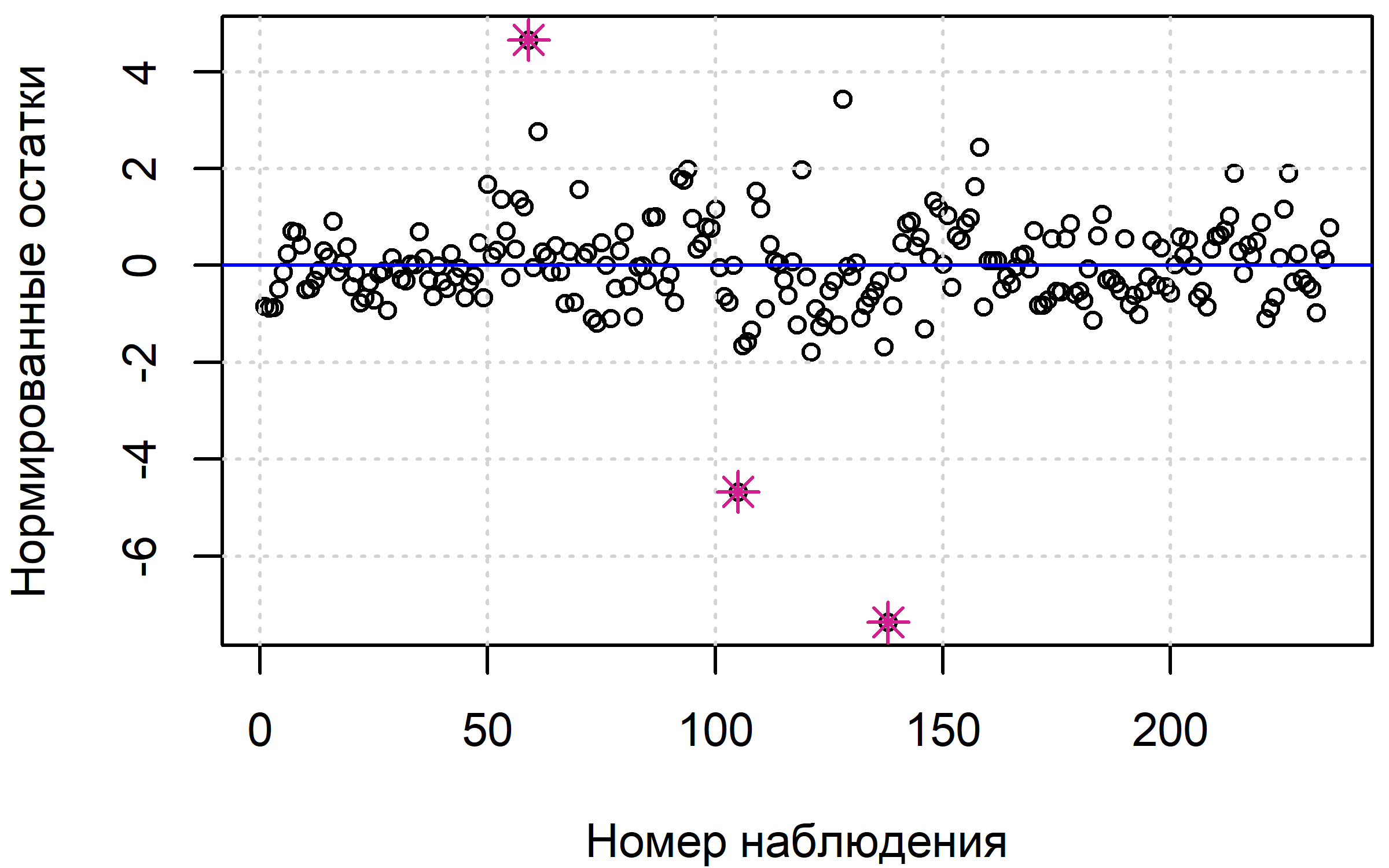


Рисунок 64. График нормированных остатков по номеру наблюдения. Выбросы показаны звездочками

Более точный анализ можно провести, исследуя так называемые **стьюдентизированные остатки**:  
где – нормированные остатки. Если выполнены стандартные предположения модели регрессии, то отдельный студентизированный остаток имеет распределение Стьюдента с  
 степенями свободы. График студентизированных остатков представляет собой как бы усовершенствованный вариант графика нормированных остатков – более чувствительный к выбросам и снабженный формальной критической границей.

(Есть альтернативная формула для расчета студентизированных остатков:  
где – диагональный элемент матрицы , а  
остаточная дисперсия, рассчитанная по всем остаткам, кроме .)

Заметим, что использование студентизированных остатков можно представить как *t*-тест на добавление фиктивной переменной вида  
в которой все элементы, кроме *i*-го, равны нулю.

При работе со студентизированными остатками следует помнить, что при использовании, скажем, уровня значимости 5 % примерно 5 % значений по модулю будут больше критической границы (при выполнении нулевой гипотезы о нормальности ошибок). Поэтому выбросами следует считать только очень большие по модулю студентизированные остатки, которым соответствует очень маленькие *p*-значения.

Другой графический метод диагностики – оценить каким-либо способом плотность распределения остатков и сравнить ее на глаз с нормальной плотностью. Если оцененная плотность существенно несимметрична или имеет толстые хвосты, то это обычно можно увидеть на графике. Самые простые и известные методы оценки плотности – это **гистограмма** и **ядерная оценка плотности** (англ. *kernel density estimator*).

В методах оценки плотности есть некоторый параметр, отвечающий за степень гладкость – его надо подобрать на основе компромисса между пересглаженностью, когда график слишком ровный и не видны детали, и недосглаженностью, когда график слишком неровный. Обычно степень гладкости несложно подобрать на глаз.

Для сравнения на том же графике удобно изобразить соответствующую нормальную плотность. Берется нормальное распределение с нулевым средним и той же дисперсией, т. е. остаточной дисперсией.

На рисунке изображены две оценки плотности – гистограмма и ядерная оценка. Видно, что плотность сильно отличается от нормальной – вершина более заостренная, а в хвостах имеются сильные выбросы, далеко отстоящие от остальных значений.

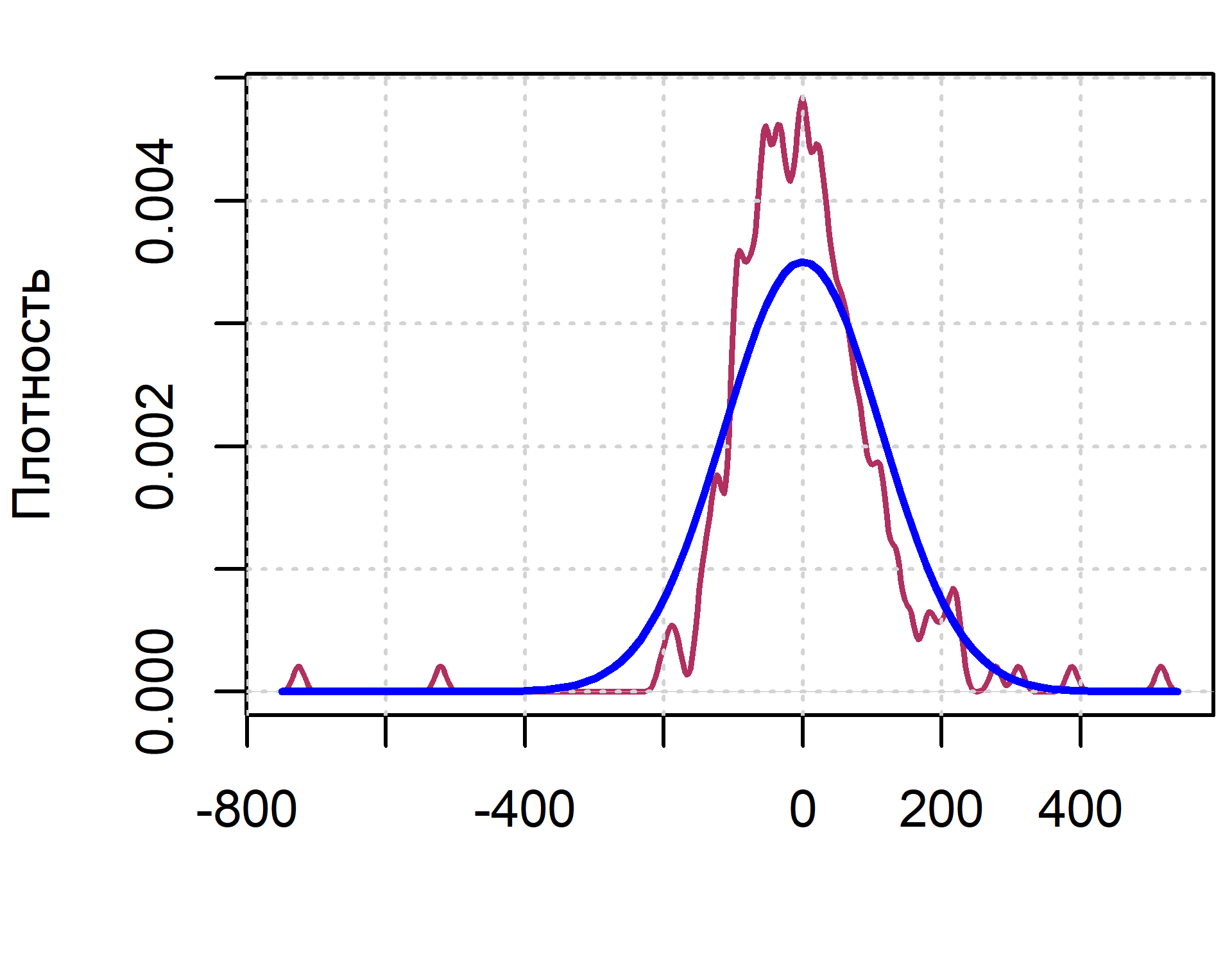
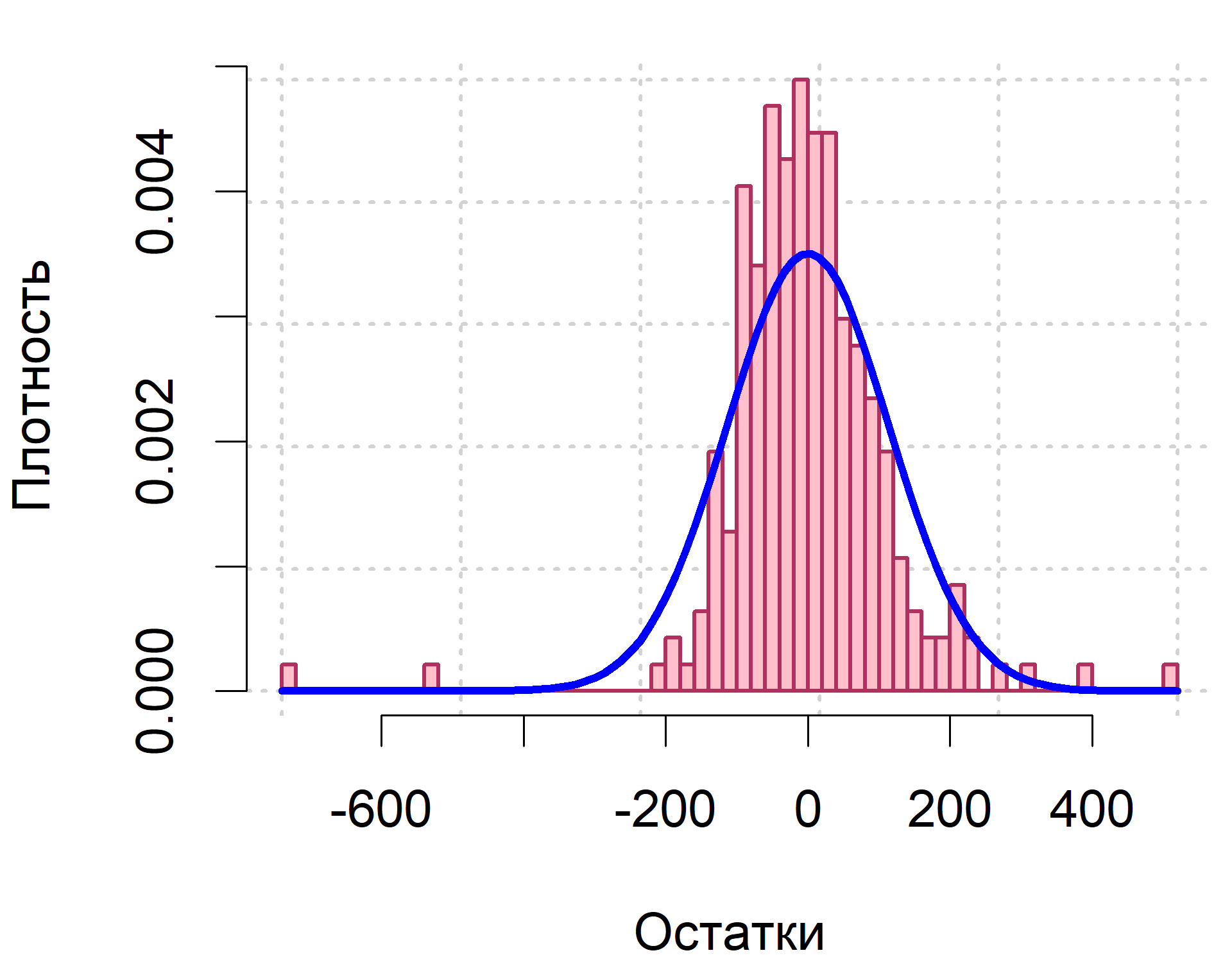


Рисунок 65. Оценки плотности для остатков регрессии по сравнению с нормальной плотностью а) гистограмма, б) ядерная оценка

Можно рассчитать по остаткам выборочные коэффициенты асимметрии и куртозиса:  
и

Асимметрия у симметричного распределения ноль. Куртозис нормального распределения равен 3. Если выборочные показатели сильно отличаются от теоретических аналогов, то это может быть признаком отсутствия нормальности. В частности, большой куртозис (существенно больше 3) указывает на распределение с толстыми хвостами.

На этих двух коэффициентах основан формальный метод диагностики – **тест Харке—Беры** (*Jarque-Bera test*). Используется статистика  
При выполнении нулевой гипотезы о нормальности ошибок регрессии статистика имеет примерно распределение (хи-квадрат с 2 степенями свободы). Если статистика больше критического значения, что случается когда асимметрия или куртозис достаточно сильно отличаются от 0 и 3 соответственно, то нулевую гипотезу следует отклонить.

Однако, как всегда бывает при использовании диагностических тестов, статистическая значимость здесь не тождественна сильному отклонению от нормальности. Выражение в скобках измеряет собственно отличие от нормальности, но затем оно умножается на . Если в регрессии много наблюдений, то гипотеза о нормальности отклоняется даже при небольших отклонениях асимметрии и куртозиса от 0 и 3.

## Асимптотическая нормальность, проверка гипотез и следствия отсутствия нормальности

При выводе распределения *t-* и *F*-статистик для коэффициентов мы использовали полный набор предположений классической модели линейной регрессии, в том числе предположение (A5) о нормальности распределения ошибок регрессии. На самом деле, на практике регрессии с ошибками, похожими на нормальные, встречаются не так уж часто. Поэтому требуется дополнительные основания для использования этих статистик. (Заметим попутно, что также следует как-то обосновать использование *t-* и *F*-статистик в тех случаях, когда они рассчитываются по робастным ковариационным матрицам класса «сэндвич».)

Такие основания дает **асимптотическая теория**. При выполнении ряда предположений в пределе при большом количестве наблюдений оценка МНК имеет многомерное нормальное распределение с соответствующей ковариационной матрицей. Из этого факта и некоторых других теоретических фактов выводится, что при выполнении нулевой гипотезы *t-*статистика будет иметь стандартное нормальное распределение , а *F-*статистика, умноженная на , – распределение , где – число ограничений. Поскольку распределение при сходится к стандартному нормальному, а распределение , масштабированное на , при сходится к , то использование квантилей *t-* и *F*-распределений вполне оправданно в качестве асимптотического приближения. Данный аргумент годится и тогда, когда при вычислении статистик используются робастные ковариационные матрицы.

В асимптотической теории предположение о нормальности заменяется другими предположениями. В частности, может использоваться предположение о конечности четвертых моментов ошибок. Это чисто теоретическое предположение, а с практической точки зрения можно сделать вывод, что нужно следить за тем, чтобы хвосты распределения ошибок были не сильно толстые.

Таким образом, мы не можем гарантировать, что *t-* и *F*-статистики имеют в точности распределения и , но это могут быть достаточно точные аппроксимации.

Кроме того, несмещенность оценки МНК не теряется при отсутствии нормальности ошибок. Более того, свойство BLUE тоже остается в силе.

В то же время, свойство BLUE не означает, что у оценки МНК нет какой-либо лучшей альтернативы. Свойство BLUE означает только, что оценка МНК *самая точная в классе линейных несмещенных оценок*. В то же время, *среди смещенных и/или нелинейных оценок могут найтись более точные оценки, чем оценка МНК*. В качестве примера, можно привести метод наименьших модулей, о котором речь ниже.

## Что делать при сильном отклонении ошибок он нормальности?

Как следует из предыдущего, отсутствие нормальности ошибок не приводит к таким серьезным проблемам, как ошибки в функциональной форме или гетероскедастичность. В то же время существуют определенные ситуации, когда *отклонение ошибок он нормальности может быть симптомом неудачной спецификации модели или каких-то других проблем*.

Во-первых, отсутствие нормальности ошибок может быть следствием гетероскедастичности. При гетероскедастичности часть ошибок имеют маленькую дисперсию, а часть – большую. Как следствие, распределение ошибок может иметь острую вершину и толстые хвосты, т. е. большой куртозис. Если в этом случае использовать взвешенную регрессию, то в ней взвешенная (или нормированная) ошибка уже, как правило, будет ближе к нормальному распределению.

Рисунок иллюстрирует случай, когда дисперсия ошибок зависит от некоторой переменной . По гистограмме остатков видно, что распределение ошибок существенно отличается от нормального.

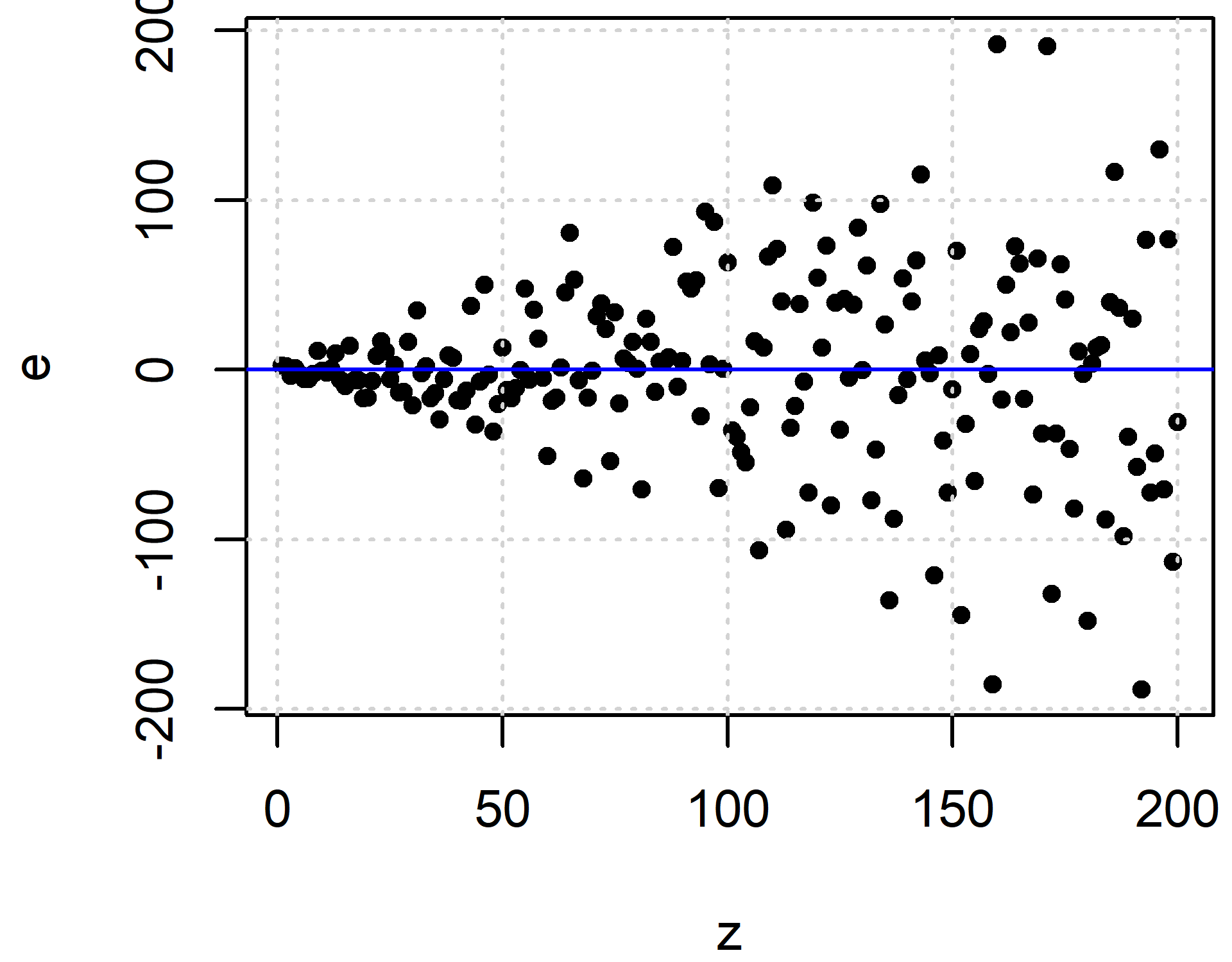
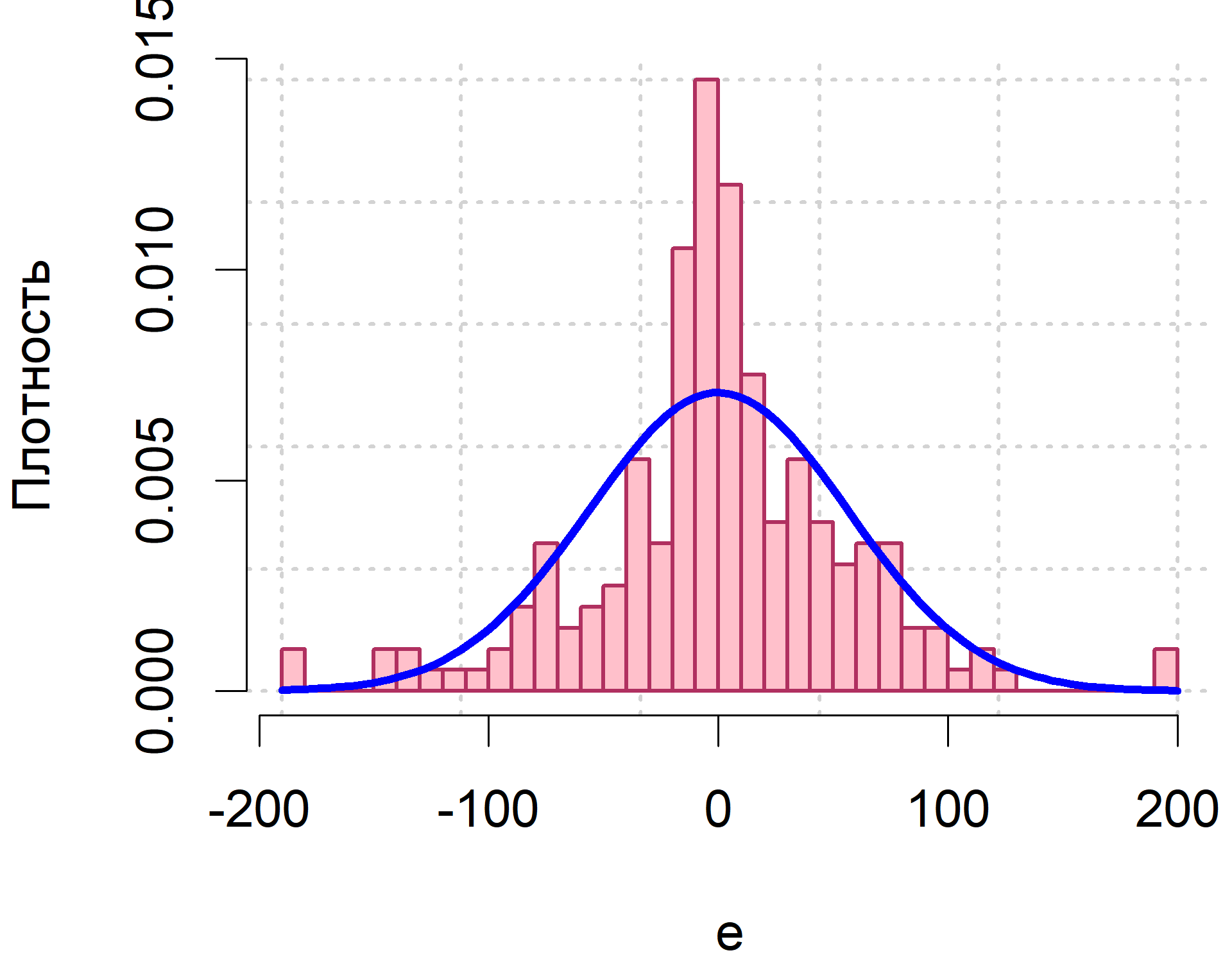
 

Рисунок 66. Гетероскедастичность может приводить к распределению ошибок с толстыми хвостами

Во-вторых, многие положительные экономические переменные, такие как население регионов страны, доходы домохозяйств, цены квартир на вторичном рынке и т. п. могут колебаться во много раз. Распределение таких переменных может быть сильно скошено вправо из-за отдельных больших значений (регионов с большим населением, домохозяйств с большими доходами). Если такие переменные используются в регрессии, то в остатках могут наблюдаться очень большие выбросы. Кроме того, в таких моделях регрессии очень часто наблюдается гетероскедастичность.

В данной ситуации есть два стандартных подхода. Один из подходов – использовать в регрессии относительные величины, деля на что-то, что отвечает за размер данной экономической единицы: ВВП на душу населения, стоимость одного квадратного метра, долю населения с определенной характеристикой в общей численности населения, долю расходов в общей величине расходов, и т. п. Другой подход – использовать переменные в логарифмах. Можно сочетать оба подхода.

На рисунках показан эффект логарифмирования. Распределение дохода в выборке домохозяйств имеет сильную положительную асимметрию с длинным правым хвостом. После логарифмирования распределение становится более похожим на нормальное, выбросов почти нет.

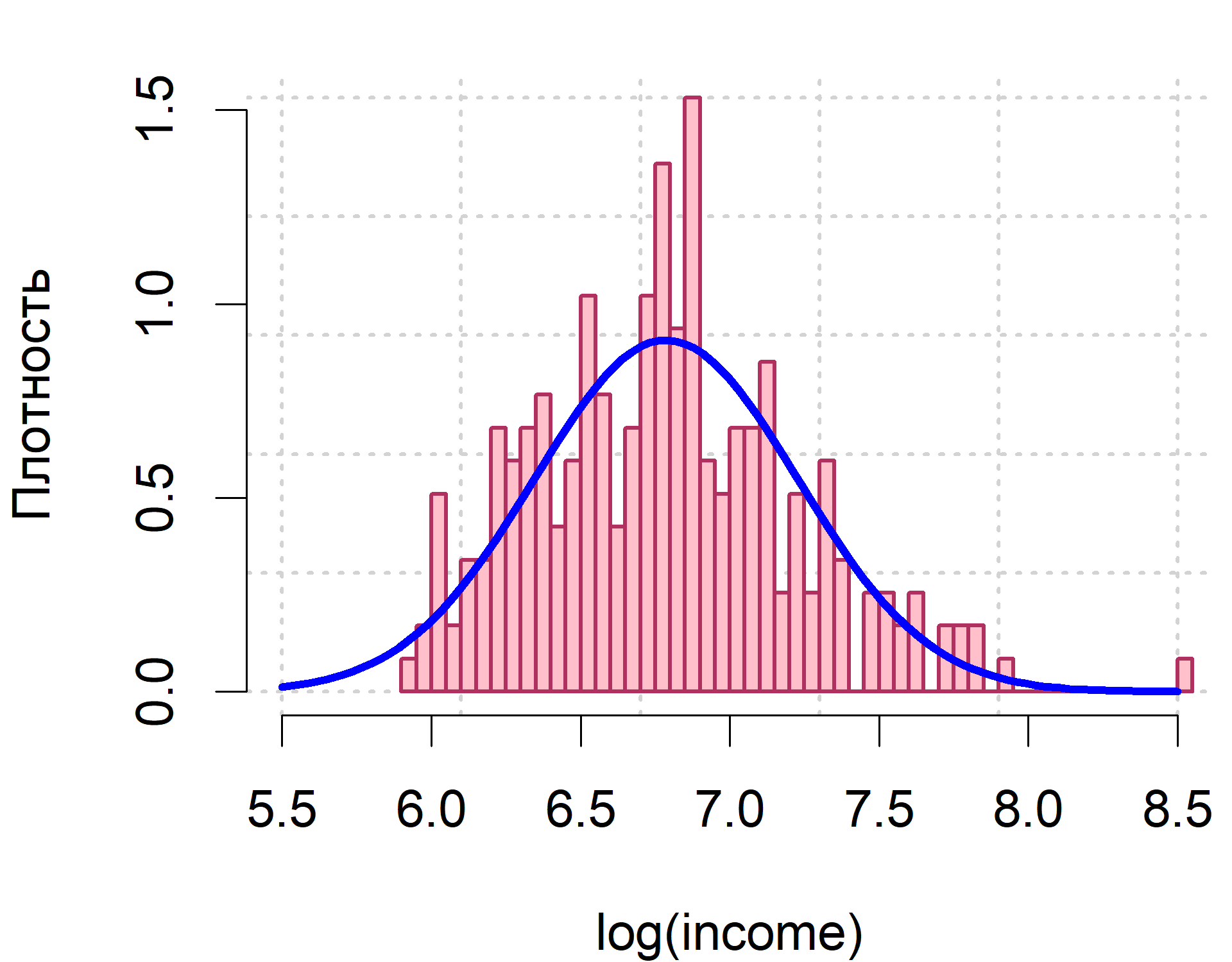
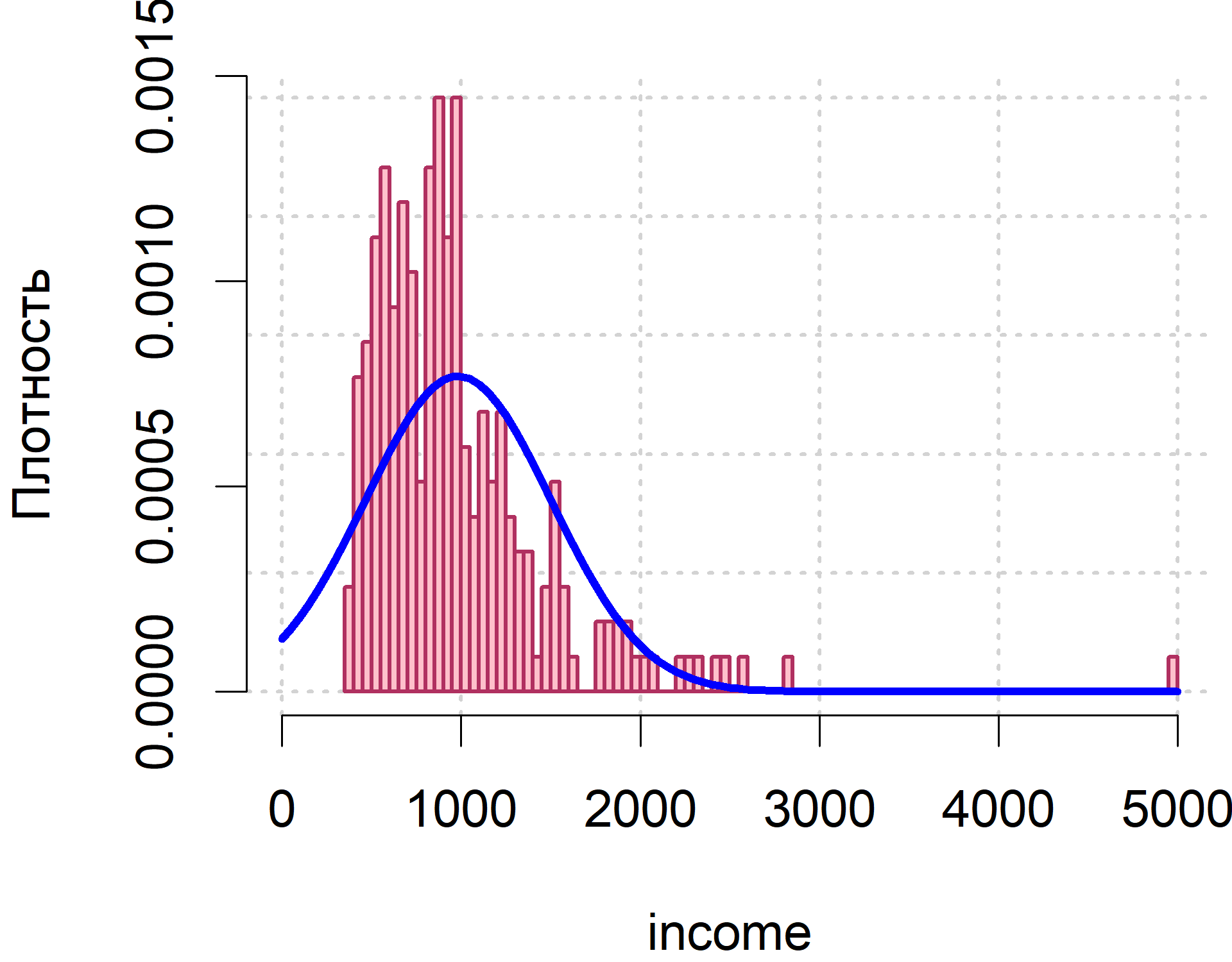


Рисунок 67. Гистограммы дохода в выборке домохозяйств и логарифма дохода

Гистограмма остатков в регрессии расходов на питание от дохода домохозяйств демонстрирует сильное отклонение от нормального распределения. Если же используется логлинейная форма модели, распределение остатков становится более похожим на нормальное.

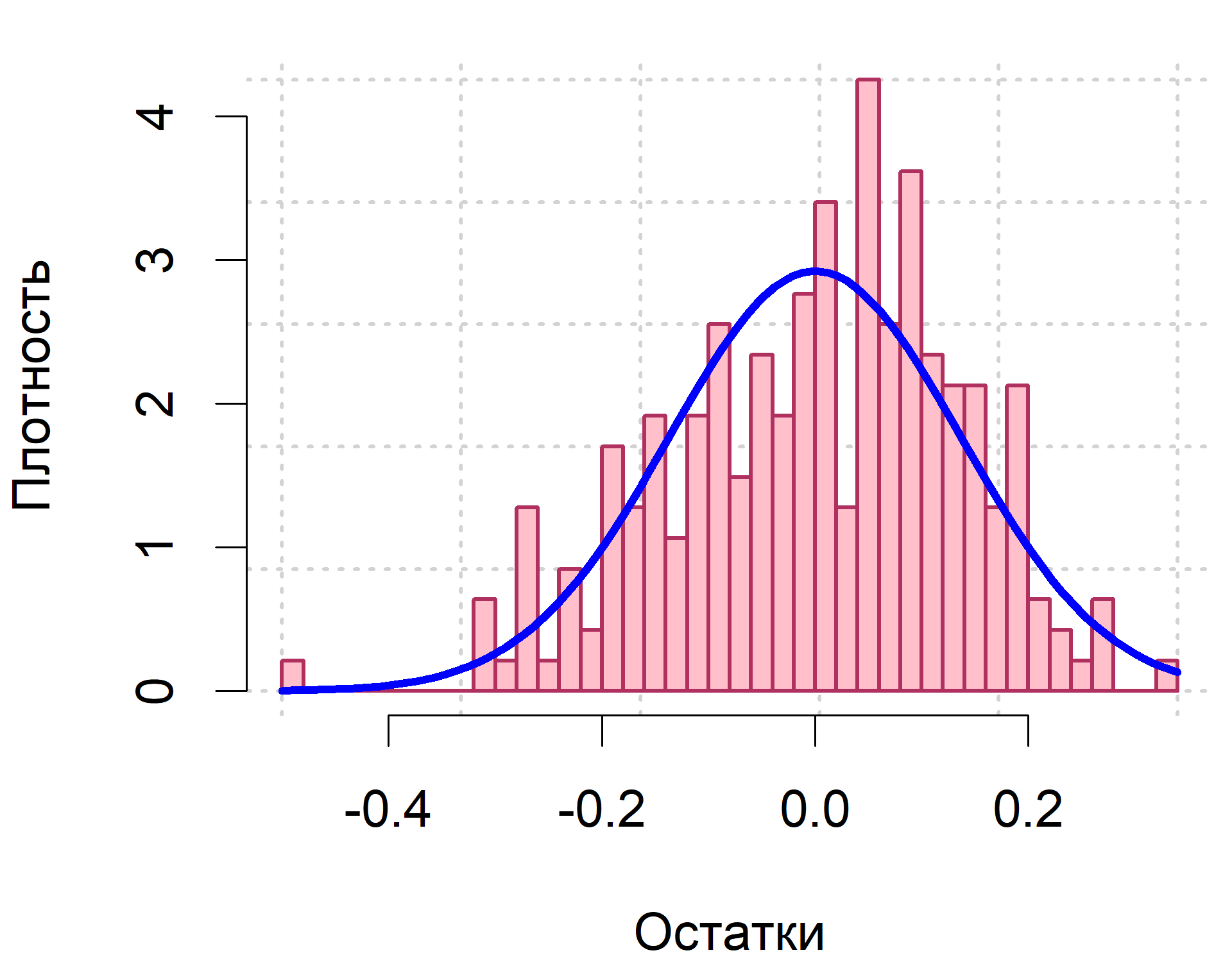
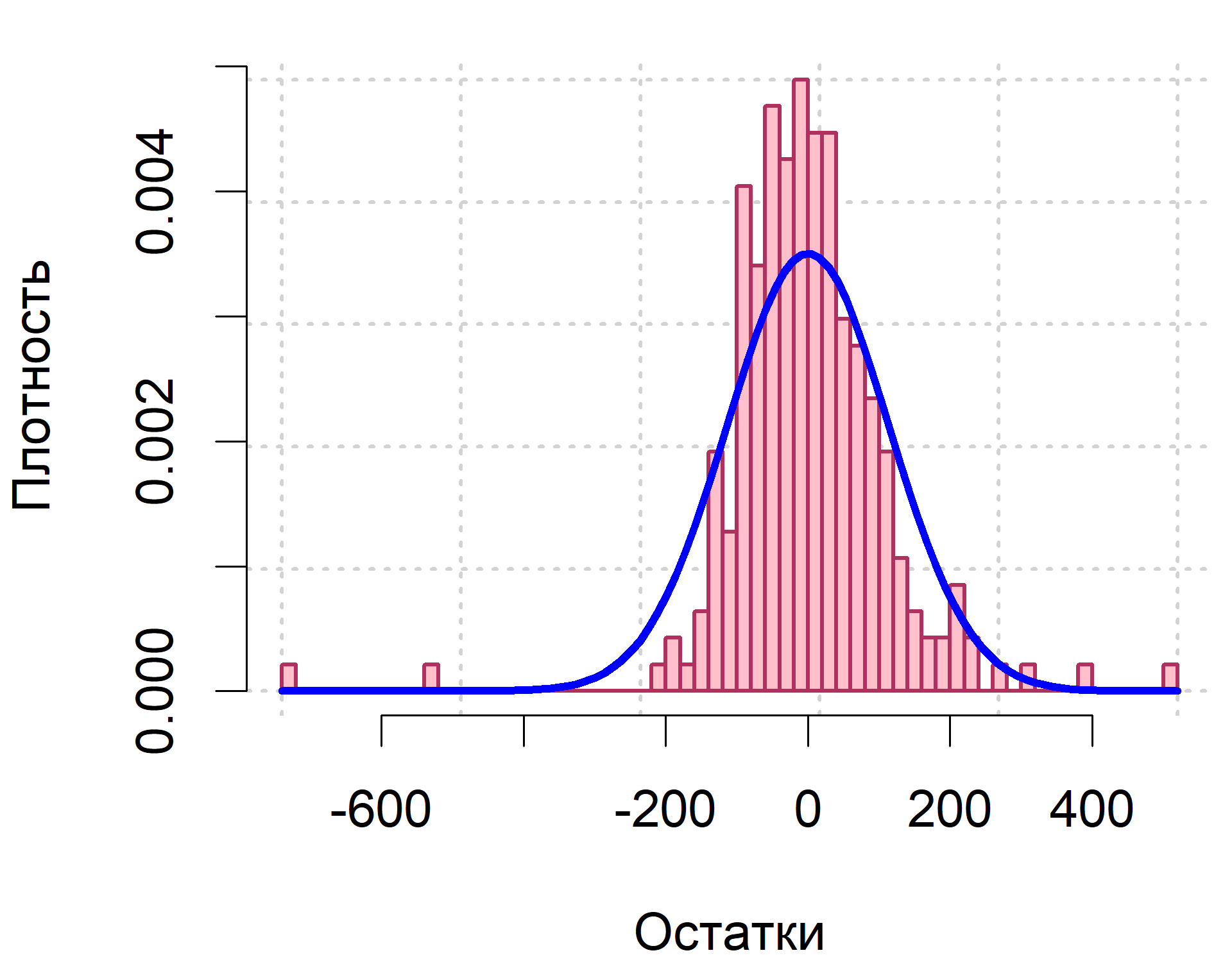


Рисунок 68. Гистограммы остатков в линейной и логлинейной регрессии расходов на питание от доходов

Основной причиной ненормальности остатков в данном случае является гетероскедастичность, как можно увидеть из точечных графиков остатков по расчетным значениям в линейной и логлинейной модели (см. рисунок).

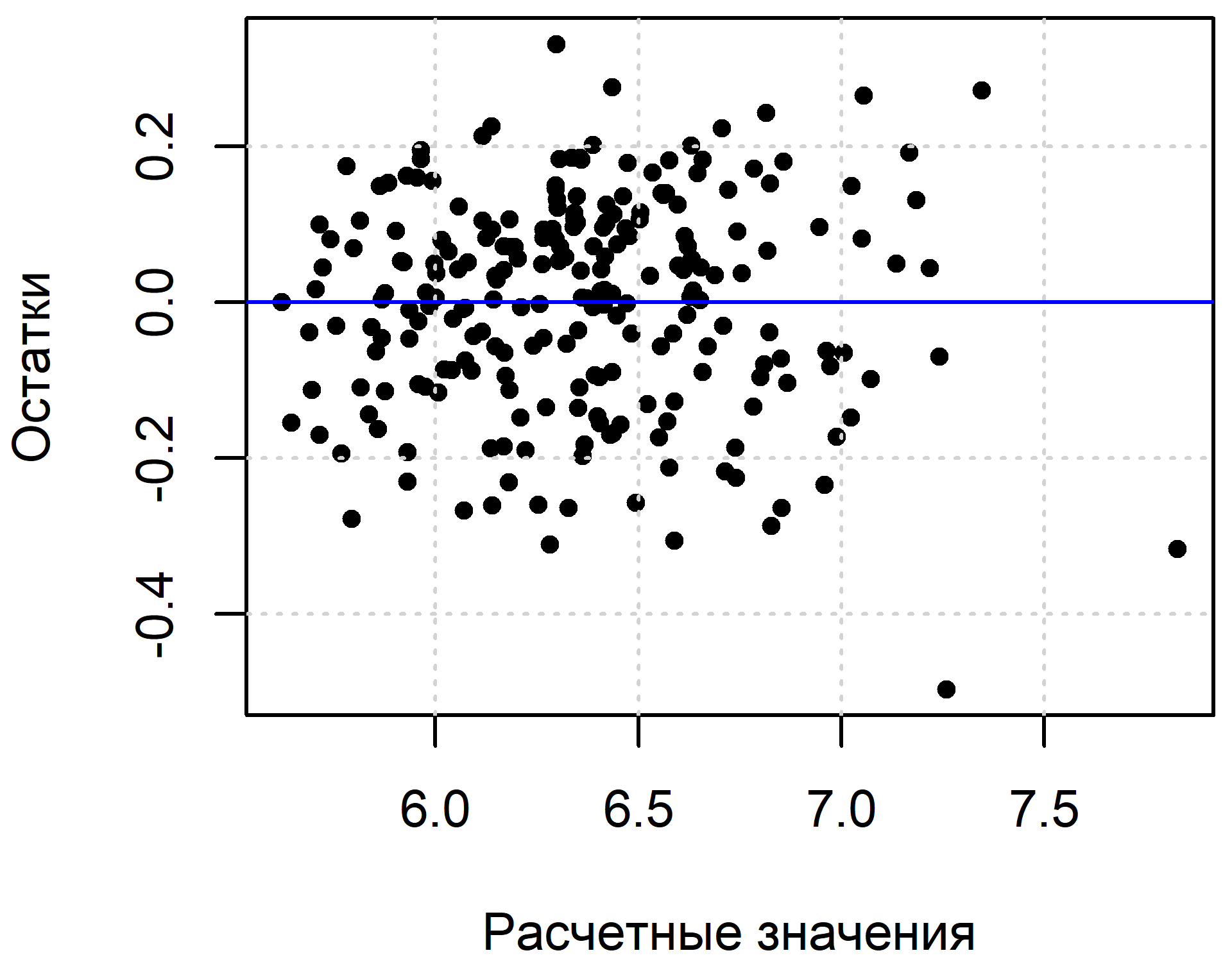
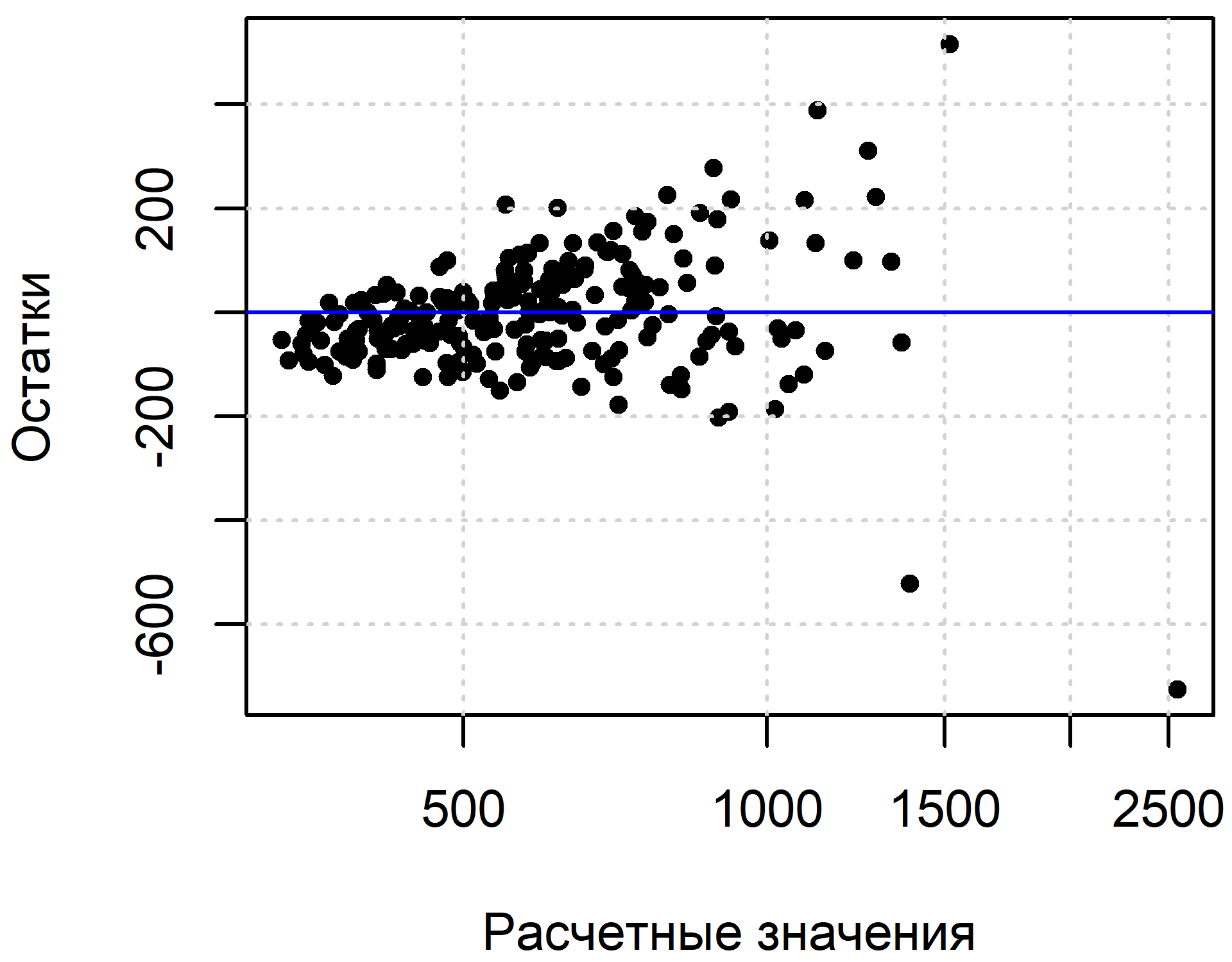


Рисунок 69. Графики остатков по расчетным значениям в линейной и логлинейной моделях (для линейной модели шкала горизонтальной оси логарифмическая)

В-третьих, выбросы в остатках регрессии могут быть следствием каких-то ошибок в исходных данных. Такие ошибки возникают, например, при ручной набивке данных (случайный пропуск запятой и т. п.). Поэтому при обнаружении больших выбросов можно порекомендовать тщательно проверить исходные данные. Если в данных обнаруживаются неправдоподобные значения, то такие наблюдения лучше удалить из регрессии.

Если отсутствие нормальности не связано с проблемами спецификации регрессии и используемых данных, то нет серьезных оснований для беспокойства, кроме случаев, когда есть очень большие выбросы.

Существуют методы оценивания, которые являются робастными к выбросам. Самый простой метод из этого класса – это **метод наименьших модулей** (метод наименьших абсолютных отклонений, англ. *least absolute deviations*, *LAD*):

Соответствующая оценка будет нелинейной по . В то же время она может быть более предпочтительной, чем оценка МНК в том случае, когда ошибки регрессии имеют очень толстые хвосты. Заметим попутно, что данную регрессию называют **медианной регрессией**. Дело в том, что при использовании обычного МНК является по смыслу оценкой математического ожидания , равного . При использовании метода наименьших модулей является по смыслу *оценкой медианы* . (Это частный случай так называемой квантильной регрессии.)

## Приложение. Ядерная оценка плотности

В основе ядерной оценки плотности лежит **ядерная функция** (или просто «ядро») . Обычно это плотность некоторого распределения. Самые широко используемые ядерные функции — это ядро Епанечникова  
и **Гауссово ядро,** соответствующее плотности стандартного нормального распределения,

Интеграл ядерной функции должен быть равен 1:

Для получения ядерной оценки плотности по выборке надо сопоставить каждой точке небольшой «холмик» с интегралом . Форма «холмика» задается ядерной функцией, в качестве центра его берется точка , а растянутость по горизонтали задается масштабирующим параметром , называемым **шириной полосы** (или шириной окна). Общая оценка плотности получается суммированием «холмиков» по всем точкам (см. рис. 70):

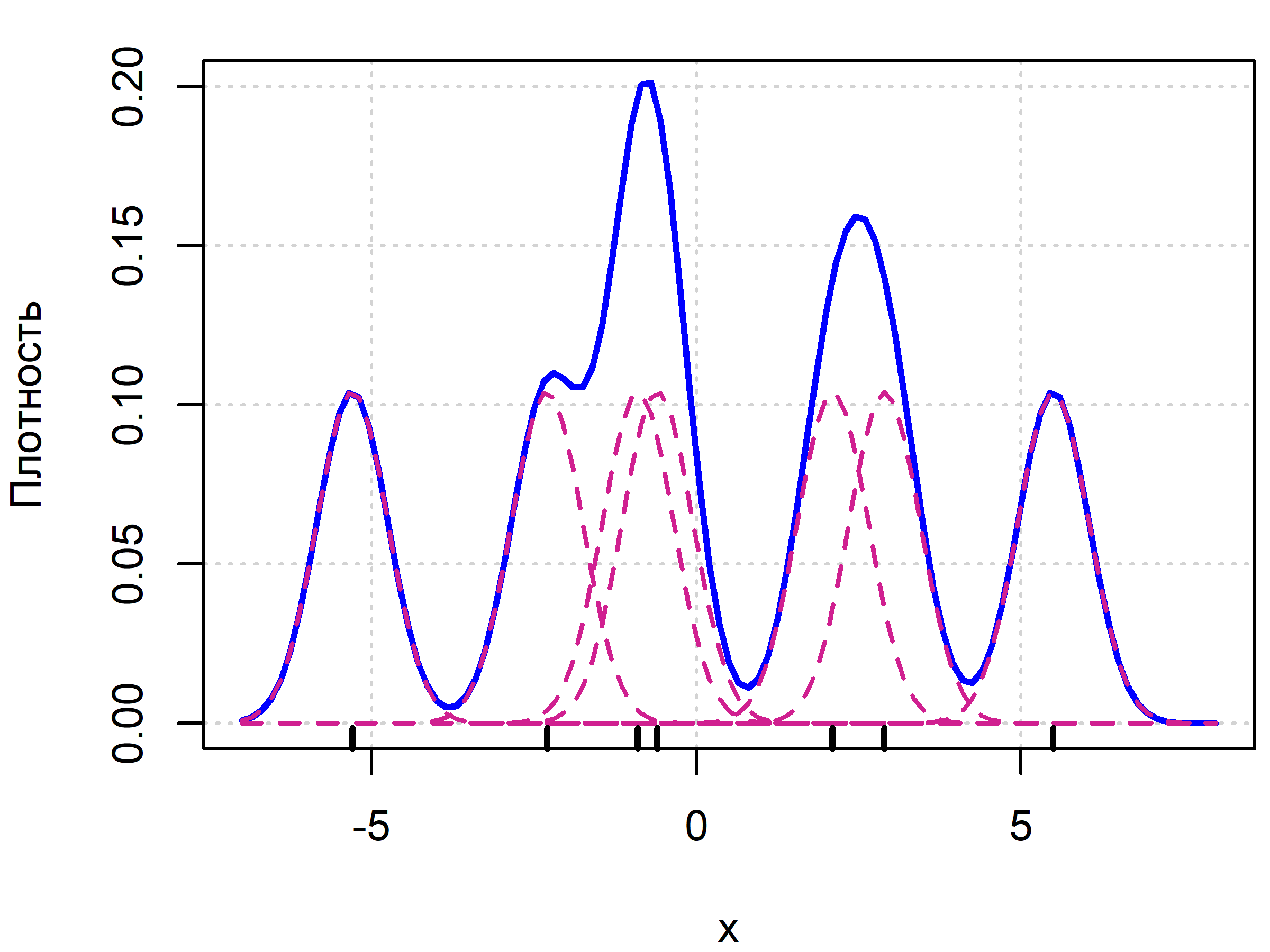


Рисунок 70. Суммирование «холмиков» для получения ядерной оценки плотности

При большом значении ширины полосы оценка плотности получается сильно сглаженной, поскольку «холмики» сильно растянуты по горизонтали. При малом же значении оценка плотности получается сильно неровной, поскольку «холмики» тогда узкие. Рис. 71 иллюстрирует различие оценок при различных величинах .

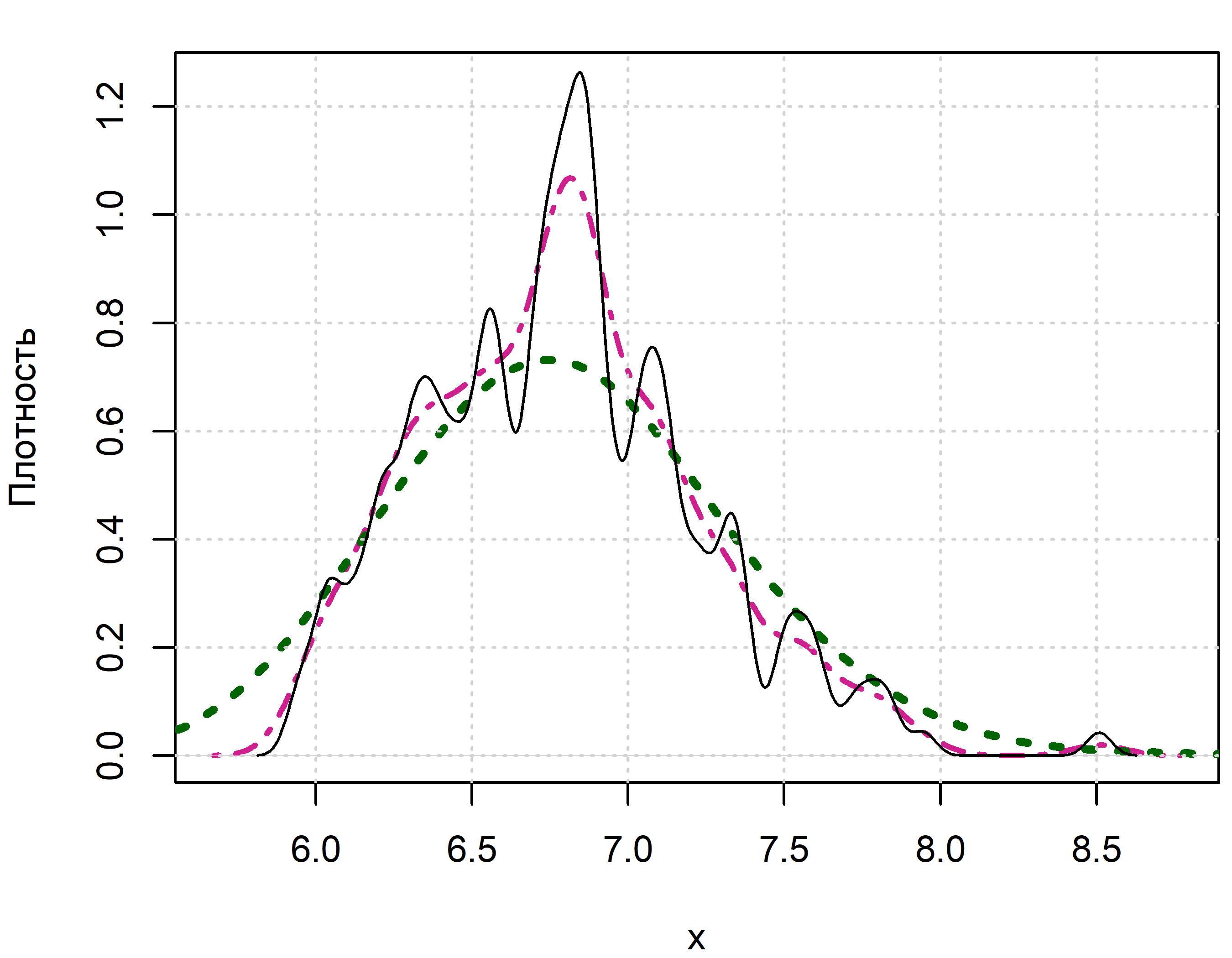


Рисунок 71. Ядерные оценки плотности при различных значениях ширины полосы

Ширину полосы следует подбирать таким образом, чтобы оценка не была слишком гладкой, скрывающей детали, и в то же время, чтобы она не была слишком неровной, подверженной случайным колебаниям.

## Контрольные вопросы

1. Как с помощью графиков диагностировать отсутствие нормальности?
2. Записать формулу выборочных коэффициентов асимметрии и куртозиса. Чему равны коэффициенты асимметрии и куртозиса нормального распределения?
3. Для чего и как используется тест Харке—Беры ?
4. Основные следствия отсутствия нормальности ?
5. Объясните, что означает асимптотическая нормальность.
6. Что такое медианная регрессия?
7. В чем суть ядерной оценки плотности?

## Экзаменационные вопросы

1. Диагностика отсутствия нормальности и наличия выбросов.
2. Асимптотическая нормальность, проверка гипотез и следствия отсутствия нормальности.
3. Основные последствия отсутствия нормальности. Что делать при сильном отклонении ошибок он нормальности?

## Литература

1. Вербик М. Путеводитель по современной эконометрике. – М.: Научная книга,. 2008. [стр.73, 292-295 ]

# Лекция: – Проблема эндогенности

## Классическая модель линейной регрессии – повторение

Для удобства изложения мы внесем переобозначения в нашу модель. А именно, пусть теперь – вектор-строка объясняющих переменных, не включающая 1 для константы, а – вектор-столбец соответствующих коэффициентов. Константу по-прежнему будем обозначать . Таким образом, будем в этом разделе курса работать с уравнением регрессии  
или в матричном виде

Кратко перечислим предположения классической модели линейной регрессии.

(A0) Модель имеет вид

(A1)

(A2) Объясняющие переменные экзогенны (детерминированы).

(A3), (A4) .

(A5)

Также кратко перечислим основные свойства оценки МНК для коэффициентов , рассчитываемой по формуле

(На оценку константы здесь не будем обращать внимание как на менее интересную.) Оценку МНК можно выразить через вектор ошибок:

(В формуле можно центрированные ошибки заменить нецентрированными – она остается верной.) Пользуясь этой формулой, выводим следующие свойства:

* (несмещенность).
* .
* .
* Оценка является BLUE.

## Случайные регрессоры и экзогенность

В этой главе мы рассматриваем проблему эндогенности. Для упрощения рассуждений ранее мы считали что объясняющие переменные не являются случайными. При этом мы смогли рассмотреть много важных аспектов модели регрессии – проверку гипотез, прогнозирование, нарушение предположений о функциональной форме, об отсутствии гетероскедастичности и автокорреляции, о нормальности ошибок. Но с предположением об экзогенности (A2) такой подход уже не годится, поскольку если не являются случайными, то они экзогенны и не могут быть эндогенными.

Таким образом, прежде чем рассматривать проблему эндогенности мы разберем, как поменяется наш анализ, если объясняющие переменные все же являются случайными. Для многих ситуаций использования регрессионного анализа это более оправданное предположение. Мы наметим два основных сравнительно простых пути работы со случайными объясняющими переменными.

### Случайные регрессоры – точная теория для конечного числа наблюдений

Один из способов учесть случайность регрессоров – это использовать практически ту же теорию, что и в случае неслучайных регрессоров, но везде рассматривать **условные распределения** и **условные моменты**. Распределения и моменты, которые рассматривались ранее, нужно считать условными относительно переменных . В записи распределений и моментов при этом используется обозначение , где вместо стоит какая-то переменная.

В частности, мы можем принять следующие предположения об условных моментах вектора ошибок:  
и  
Это означает, что условное по математическое ожидание ошибок равно нулю, и что условная по ковариационная матрица является диагональной с одинаковыми диагональными элементами .

Предположение о нормальности распределения вектора ошибок можно заменить на предположение об условной нормальности:

Это означает, что условное по матрице регрессоров распределение вектора является многомерным нормальным с соответствующими параметрами.

С практической точки зрения использование условных распределений и моментов сводится к тому, что при вычислениях с матрицей и ее различными функциями можно обращаться как с константами – выносить из-под оператора условного математического ожидания , выносить множители с квадратом из-под оператора условной дисперсии и т. д.

При соответствующих предположениях можно вывести различные свойства оценки МНК, где в формулировках будут фигурировать условные распределения и моменты. В частности, мы можем вывести следующие свойства:

В предположении (условной) нормальности распределения ошибок оценка МНК тоже распределена нормально условно по :

### Случайные регрессоры – асимптотическая теория

Альтернативный подход к теории регрессии использует асимптотические соображения. Пусть для каждого наблюдения имеются следующие постоянные (т. е. не зависящие от ) вторые центральные моменты переменных:

Если ошибка не коррелирована с регрессорами, то есть  
то, получим  
или

При невырожденности матрицы это означает, что

(Невырожденность обеспечивает **асимптотическую идентифицируемость** коэффициентов – то есть, что их потенциально, при достаточно большом числе наблюдений, можно отличить от других коэффициентов .)

Соответствующие выборочные моменты аппроксимируют теоретические моменты (т. е. служат их приближениями):  
где  
Для обоснования здесь можно сослаться на один из вариантов закона больших чисел, который гласит, что (при определенных предположениях) выборочное среднее сходится по вероятности к математическому ожиданию.

Далее, величина аппроксимирует величину . Говоря формально сходится по вероятности к при :  
Здесь можно сослаться на другой асимптотический результат – теорему Манна — Вальда (непрерывные функции сохраняют предел по вероятности).

Как можно увидеть, здесь – это не что иное, как оценка МНК:  
Таким образом, имеет место **состоятельность оценки МНК**  – она сходится по вероятности к истинным коэффициентам при стремлении количества наблюдений к бесконечности:

При некотором другом наборе предположений, который мы не будем обсуждать, имеется также **асимптотическая несмещенность**, то есть свойство, что смещение оценки МНК стремиться к нулю при стремлении количества наблюдений к бесконечности:

Можно также при дополнительных предположениях доказать **асимптотическую нормальность** оценки МНК и обосновать использование стандартных *t*-статистик и *F*-статистик асимптотическими соображениями.

### Предположение об экзогенности регрессоров

Ранее в предположении (A2) под экзогенностью регрессоров мы понимали то, что они не случайны. В случае случайных регрессоров надо каким-то образом модифицировать это предположение. Из предыдущего обсуждения теории регрессии при случайной природе можно догадаться, что «экзогенность» – это не вполне однозначное понятие и что существует несколько различных вариантов экзогенности.

Неформально экзогенность означает, что регрессоры имеют внешнее по отношению к модели регрессии происхождение.

Мы рассмотрим два основных варианта экзогенности.

**Слабая экзогенность** означает, что регрессоры не коррелируют с ошибкой, т. е.

С учетом предположения (A1) о том, что ошибка в среднем равна нулю, т. е. , можем слабую экзогенность записать в виде **условия ортогональности** между регрессорами и ошибкой

В условиях слабой экзогенности при обсуждении теории МНК приходится опираться на асимптотические утверждения (асимптотическая несмещенность или состоятельность вместо обычной несмещенности, и т. п.).

**Строгая экзогенность** означает, что условное по математическое ожидание вектора ошибок совпадает с безусловным:

С учетом предположения (A1) о том, что ошибка в среднем равна нулю, строгую экзогенность можно записать в виде

При строгой экзогенности вектор ошибок не коррелирует с любыми функциями от регрессоров . Таким образом, если – это некая функция от , то имеем  
и

Фактически, это свойство – что вектор ошибок не коррелирует с любыми функциями от регрессоров – является альтернативным эквивалентным определением строгой экзогенности.

Понятно, что если слабая экзогенность нарушена, а именно, существует наблюдение , для которого некоторый регрессор , коррелирует с ошибкой , то строгая экзогенность тоже нарушена, поскольку тогда мы знаем функцию , для которой .

### Последствия отсутствия экзогенности

В дальнейшем мы для упрощения рассуждений будем использовать слабую экзогенность. При этом дополнительно предполагается, что наблюдения представляют собой выборку, то есть независимы между собой и имеют одинаковые распределения. Это упрощение позволяет проанализировать ключевые аспекты, связанные с экзогенностью. Анализ более тонких вопросов выходит за рамки данного курса. В частности, мы не будем рассматривать, эффекты, связанные с нелинейными зависимостями или с зависимостями между переменными из разных наблюдений ‍. Линейность взаимосвязей позволяет говорить о простых ковариациях между рассматриваемыми переменными.

Проанализируем при сделанных предположениях последствия отсутствия экзогенности. (Такое явление принято называть **эндогенность**.) Для этого рассмотрим связи между и в виде следующей **теоретической регрессии**:  
где . Такое представление всегда существует. Как мы видели ранее,  
где , .

Если мы с помощью МНК оцениваем модель регрессии  
где имеет место эндогенность, т. е. ошибка коррелирована с регрессорами :  
то вместо требуемых коэффициентов будем по факту оценивать , где  
Данный факт следует из того, что

В условиях эндогенности, т. е. при , имеем и, как следствие, . Тем самым, в оценках МНК будет наблюдаться **смещение из-за эндогенности**, равное

Поскольку в рамках обсуждения слабой экзогенности мы в общем случае не ожидаем несмещенности оценки МНК (а только асимптотической несмещенности или состоятельности), то в дальнейшем будем иметь в виду именно это смещение в теоретических коэффициентах, а не величину смещения в оценке МНК .

(Следует пояснить, что эндогенность может привести также к смещению в константе, т. ‍е. , но это смещение мы не будем здесь рассматривать, поскольку оно менее важное.)

### Экзогенность и причинность

Наиболее содержательно интересной с точки зрения экономического анализа является причинная интерпретация экзогенности. Дело в том, что, как мы видели, любой случайной выборке мы можем сопоставить теоретическую регрессию

Соответствующая эмпирическая МНК-регрессия даст асимптотически несмещенную и состоятельную оценку коэффициентов . Однако далеко не всегда такая регрессия имеет осмысленную интерпретацию.

Здесь можно вспомнить часто повторяемое высказывание:

«*Корреляция не означает причинность*».

Например, можно построить зависимость дохода семьи от потребления картофеля. Однако вряд ли доход российской семьи зависит от потребления картофеля. Скорее, наоборот, потребление картофеля зависит от дохода семьи. Поэтому мы получим осмысленную экономическую интерпретацию наших данных, только если поменяем объясняемую и объясняющую переменную местами.

При обсуждении причинности и направления зависимости между переменными удобно пользоваться диаграммами со стрелками. Однонаправленная стрелка на диаграмме показывает потенциальную одностороннюю зависимость одной переменной от другой. При этом подразумевается, что влияние в противоположном направлении отсутствует Двунаправленная стрелка будет означать потенциальную зависимость (корреляцию) между переменными, в источниках и направлении которой мы не уверены. Отсутствие стрелки означает отсутствие зависимости (корреляции).

Рисунок 72. Причинные влияния в «правильной» регрессии

С точки зрения причинных влияний в «правильной» регрессии влияние должно идти от регрессоров к зависимой переменной и от ошибки к зависимой переменной . Другие взаимосвязи в рассматриваемой простой ситуации предполагаются отсутствующими. (См. рис. 72.)

Почему для регрессии важна причинность? Дело в том, что очень часто требуется дать на основе регрессии какие-то конкретные рекомендации по выбору политики или способа действий. Правительству может быть важно знать, как те или иные мероприятия экономической политики скажутся на состоянии экономики. Менеджеру фирмы может быть важно знать, как те или управленческие решения скажутся на будущей прибыли. Примеров можно привести много.

Проблема состоит в том, что такая однозначная схема, как на приведенном рисунке, редко встречается в реальной жизни, и поэтому коэффициенты регрессии могут быть обманчивыми.

Пусть, например, построена межстрановая регрессия ВВП от государственных расходов на образование. В этой регрессии получен коэффициент при государственных расходах . Означает ли это, что если в некоторой стране увеличить расходы на образование на , то ВВП страны возрастет примерно на ? Скорее всего нет, поскольку в рассматриваемой регрессии может иметь место эндогенность. То есть регрессию построить несложно, но давать политические рекомендации на ее основе было бы опрометчиво.

Или пусть, например, из собранной статистики известно, что у закончивших магистратуру в среднем заработная плата на 15 % выше, чем у других получивших высшее образование. Если выпускник бакалавриата может поступить в магистратуру и в принципе готов дополнительно учиться, чтобы получить 15-ти процентную прибавку к заработку, то стоит ли ему ориентироваться на указанную статистику при принятии решения о поступлении? Как мы обсудим далее, в этой цифре может присутствовать смещение.

В общем случае рассмотрим ситуацию, когда по регрессии  
получена оценка МНК и делается предсказание , где – некоторый базовый уровень регрессоров для нового наблюдения . Каким будет наше предсказание изменения при изменении значений регрессоров на ? Если использовать формулу прогноза , то это будет  
Фактически приросту соответствует прирост ожидаемого на . Поскольку , то смещение в предсказании среднего эффекта прироста равно примерно  
(Заметим, что если мы находимся в положении принимающего решение и сами выбираем приращение , то надо рассматривать как неслучайную величину.)

Таким образом, взгляд на экзогенность с позиций причинных влияний важен для практики. Когда же мы можем быть уверены, что имеет место экзогенность? Как правило, такая уверенность бывает при понимании природы процессов, которые определяют уровень регрессоров .

Во-первых, бывают ситуации, когда мы сами контролируем уровни регрессоров. Обычно это ситуация, когда проводится **эксперимент**. При этом для обеспечения экзогенности тот, кто проводит эксперимент, должен сам следить за тем, чтобы выбор им уровней регрессоров не оказывался под влиянием тех факторов, которые определяют ошибку .

Во-вторых, бывают ситуации, когда уровень регрессоров определяется некоторыми чисто случайными факторами и мы обладаем информацией, что случайная ошибка определяется какими-то другими, независимыми факторами. Когда ситуация, похожая на эксперимент, создается сама собой, без вмешательства исследователя, принято говорить о ‍**естественном эксперименте**.

В некоторых случаях нам может помочь в анализе временна́я последовательность событий. Например, на факт поступления выпускника бакалавриата в магистратуру могут повлиять события, происходившие с выпускником в бакалавриате, но обратного причинного влияния быть не может. В подобном анализе используется известный принцип «**после этого — значит по причине этого**» (лат. *post hoc ergo propter hoc*). Конечно, далеко не всегда такие рассуждения являются корректными, но этот принцип все равно бывает полезен с практической точки зрения.

## Основные причины эндогенности

### Пропущенные переменные

Пусть уравнение  
удовлетворяет предположениям, что и пусть переменные и не коррелируют с ошибкой регрессии , то есть имеет место слабая экзогенность. Предполагается, что – это вектор объясняющих переменных, а – это некоторая дополнительная объясняющая переменная. Пусть вместо этого оценивается другое уравнение, где отсутствует:

Причины, по которым не включается в регрессию, могут быть самыми разными. Например, это может быть ошибка моделирования, связанная с невнимательностью исследователя. Или исследователь может не подозревать о том, что может быть важной переменной. Кроме того, переменная может являться ненаблюдаемой, т. ‍е. данные о ней недоступны.

Что же произойдет, если регрессия оценивается без переменной ? Мы можем получить из первого уравнения второе, положив

Для экзогенности требуется, чтобы для выполнялось . В данном случае

Как видим, регрессоры в регрессии с пропущенной переменной не будут коррелировать с ошибкой только если , то есть переменная отсутствует в исходной регрессии, или если , т. е. переменные не коррелируют с .

Рисунок 73. Схема для случая пропущенной переменной

В общем случае и могут коррелировать друг с другом, а может являться важной переменной (), то есть мы находимся в ситуации, изображенной на схеме (рис. 73). При этом мы не можем быть уверены в экзогенности и отсутствии смещения коэффициентов.

По общей формуле смещения из-за эндогенности смещение из-за пропущенной переменной равно  
(Здесь, как несложно увидеть, – это коэффициенты наклона в теоретической регрессии от .)

В частном случае , когда в входит только одна переменная, т. е. когда уравнение регрессии имеет вид  
смещение из-за пропущенной переменной равно

В качестве условного примера рассмотрим регрессию, связывающую размер заработной платы выпускника университета с учебой в магистратуре. Пусть , если выпускник закончил магистратуру, и , если он получил только степень бакалавра. Объясняемая переменная – заработная плата выпускника через 20 лет. В регрессию она входит в виде логарифма. Таким образом, первоначальная модель имеет вид

По сути, такая регрессия сравнивает средний уровень логарифма заработной платы в двух группах выпускников – учившихся в магистратуре и не учившихся.

Проблема с этой регрессией заключается в том, что не учитывается то, что обобщенно можно назвать уровнем выпускника на момент подачи документов в магистратуру. Введем соответствующую обобщающую переменную . В качестве показателя уровня можно взять средний балл за бакалавриат или баллы за выпускной экзамен. Кроме того, здесь можно принять во внимание природные способности, трудолюбие и т. п.

Если не учитывать уровень , то мы получим смещенную оценку для коэффициента . Дело в том, что у выпускника с большим больше шансов оказаться в магистратуре. Во-первых, ему проще сдать вступительные экзамены в магистратуру. Во-вторых, в среднем такому выпускнику проще учиться в магистратуре, поэтому больше вероятность, что он попробует туда поступить. Могут быть и другие соображения, из-за которых вероятность того, что выше для выпускника с большим уровнем . Но у выпускника с большим уровнем при прочих равных условиях и зарплата в среднем выше. Таким образом, здесь будет коррелировать с , а , в свою очередь, будет одной из составляющих ошибки . Это приведет к тому, что и будут положительно коррелированы, то есть предположение об экзогенности будет нарушаться.

На схеме (рис. 74) изображены зависимости, которые подразумеваются в данной ситуации. Здесь – это ошибка нашей регрессии, которая включает все факторы, которые могут повлиять на заработную плату, и которые мы не учли. Это уровень выпускника и прочие неучтенные факторы . На поступление в магистратуру влияет как уровень выпускника , так и прочие неучтенные факторы. Мы предполагаем здесь, что эти факторы не связаны с и .

Рисунок 74. Схема взаимосвязей в примере с магистратурой

Если выпускник бакалавриата будет ориентироваться на оценку МНК для коэффициента , то он может из-за смещения получить в результате более низкую заработную плату, чем ожидал. Данное смещение равно  
где – коэффициент прямого влияния уровня на логарифм заработной платы . Поскольку мы ожидаем, что и , то смещение должно быть положительным. Даже если бы магистратура вообще не давала прироста зарплаты (истинный коэффициент при был бы нулевой), в регрессии от мы все равно бы получили положительную зависимость из-за смещения. В дальнейшем при обсуждении контрольных переменных мы более подробно проиллюстрируем это смещение на графике регрессии.

### Ошибки в переменных

Пусть имеется некоторые «исходные» значения регрессоров, но они нам неизвестны, и пусть для этих исходных регрессоров выполнено уравнение  
в котором нет проблемы эндогенности, то есть . Что будет, если нам известны значения , измеренные с ошибкой, и мы используем их в регрессии вместо ?

Введем наблюдаемые аналоги переменных :  
Здесь – ошибки измерения. Предположим, что оценивается модель регрессии  
Выражая исходные переменные через и ошибки измерения, получим  
Отсюда видно, что  
Поскольку ошибки измерения входят как в , так и в , следует ожидать, что в оцениваемой регрессии с ошибками в переменных может иметь место эндогенность.

Предположим более конкретно, что

Тогда

При сделанных предположениях эта величина упрощается до  
Здесь – ковариационная матрица ошибок измерения. Если она нулевая, то и проблемы эндогенности нет. Если же она не нулевая, то, кроме особых случаев, .

Смещение из-за ошибок в переменных равно  
Здесь (с учетом некоррелированности и )  
Таким образом,

В случае будем иметь  
и формула упрощается до  
откуда

Видим, что при нетривиальной ошибке измерения, т. е. когда , измеряемый коэффициент отличается от истинного на множитель . Таким образом, происходит смещение по сравнению с истинным коэффициентом в сторону нуля. По знаку оказывается таким же, как , но по модулю меньше. Линия регрессии, таким образом, оказывается более пологой, чем требуется.



Рисунок 75. Смещение из-за ошибок измерения

Рис. 75 иллюстрирует смещение из-за ошибок измерения. Исходная линия регрессии (сплошная) соответствует исходной переменной . Вместо переменной наблюдается переменная с ошибкой . Ошибки показаны стрелками. Они смещают точки наблюдений по горизонтали, из-за чего облако наблюдений «расползается» в горизонтальном направлении, а соответствующая линия регрессии (штриховая) оказывается более пологой, чем исходная.

### Одновременность. Системы одновременных уравнений

Пусть у нас есть наблюдения в разные моменты времени за рынком некоторого товара: , , где – объем продаж в натуральном выражении, а – средняя цена товара. Предположим, что мы захотели по этим данным оценить уравнение спроса – что объем продаж линейно зависит от цены :

Здесь мы ожидаем отрицательный знак у коэффициента наклона – чем больше цена, тем меньше спрос.

Однако с таким подходом возникает следующая проблема – в модели рынка есть как сторона спроса, так и сторона предложения. Со стороны предложения тоже есть зависимость продаж от цены – чем больше цена, тем больше товара будут предлагать продавцы. Мы можем предположить, что объемы и цены в каждом случае определяются как точка пересечения кривой спроса () и предложения (), где кривые спроса и предложения могут быть подвержены сдвигу из-за неучтенных случайных факторов:

Мы не можем по одним и тем же данным оценить как уравнение спроса, так и уравнение предложения. Прежде всего, проблема в том, что два уравнения выглядят одинаково. Чтобы их отличить друг от друга, нужно ввести в соответствующие уравнения дополнительные объясняющие переменные, которые влияют только на спрос и только на предложение. Пусть на спрос влияет переменная , а на предложение – переменные и . Модель примет вид

Тогда, оценивая первую регрессию, мы оцениваем уравнение спроса, а оценивая вторую регрессию, – уравнение предложения.

Мы имеем здесь дело с так называемой системой **одновременных регрессионных уравнений**. Переменные и здесь являются **эндогенными** (внутрисистемными), а переменные , и – **экзогенными** (внешними по отношению к рассматриваемой системе – рынку одного товара). Система одновременных уравнений представлена здесь в так называемой **структурной форме**, то есть в такой форме, что уравнения в ней имеют содержательную экономическую интерпретацию. В данном случае это кривые спроса и предложения, – коэффициент наклона для кривой спроса, – коэффициент наклона для кривой предложения.

Это неявные уравнения, совместно задающие эндогенные переменные и . Мы можем решить данную систему относительно и и выразить эндогенные переменные в явном виде через экзогенные переменные , и и ошибки и :

Это так называемая **приведенная форма** системы одновременных уравнений, в которой слева стоят эндогенные переменные, а справа – только экзогенные перемененные (и случайные ошибки).

Из приведенной формы видно, что в общем случае эндогенные переменные и коррелируют с обеими ошибками приведенной формы ( и ). Таким образом, если рассматривать уравнения структурной формы (спрос и предложение) как уравнения регрессии, то в них имеет место проблема эндогенности – регрессор коррелирует с ошибкой уравнения ( и соответственно). Из-за этого оценки, полученные обычным МНК будут смещенными и несостоятельными.

Это так называемое **смещение из-за одновременности** (англ. *simultaneity bias*), где под одновременностью имеют в виду такую ситуацию, когда зависимая переменная и регрессор совместно определяются внутри рассматриваемой системы.

На рис. 76 приведена схема причинности для рассматриваемого примера.

Рисунок 76

Обычно при рассмотрении одновременных уравнений каждому из уравнений сопоставляют ровно одну эндогенную переменную, которую ставят в правую часть. Пусть, например, для спроса это , а для предложения – :

Здесь второе уравнение системы – это обратная функция предложения: какой должна быть цена , чтобы продавцы выставили на рынок данный товар в количестве . В этом виде структурной формы схема причинности выглядит более наглядно (см. рис. 77).

Рисунок 77

В таком представлении видно, что две эндогенные переменные влияют друг на друга, то есть имеет место **двусторонняя причинность**.

Двусторонняя (и многосторонняя причинность) очень часто имеет место в экономических явлениях. Например, если мы строим зависимость уровня экономического развития региона (например, валового регионального продукта) от размера прямых иностранных инвестиций в регион, то можем столкнуться с проблемой эндогенности из-за двусторонней причинности. Дело в том, что не только прямые иностранные инвестиции могут влиять на уровень развития, но и, наоборот, уровень развития может влиять на размер прямых иностранных инвестиций, поскольку инвесторы могут более охотно вкладывать деньги в более развитый регион. Из-за подобных эффектов очень легко получить бессмысленное уравнение регрессии, в котором оценки коэффициентов не будут нести полезной информации.

Во второй половине XX века тема одновременных уравнений была очень популярной. Исследователи оценивали системы, состоящие из большого числа уравнений и большого числа переменных. Однако постепенно мода на такие системы сошла на нет, поскольку стало ясно, что обосновать вид каждого из уравнений и экзогенность каждой из экзогенных переменных не представляется возможным. Ошибка даже в одном из предположений может поставить под вопрос всю систему. Поэтому в настоящее время все большую популярность набирает более скромный подход, когда оценивается одно уравнение, а все остальные потенциальные зависимости используются только для обоснования эндогенности и экзогенности переменных в этом уравнении. Отдельное уравнение можно оценить одним из методов, о которых речь пойдет ниже.

Более того, некоторые исследователи придерживаются взгляда, что предпочтительнее исследовать причинную зависимость одной объясняемой переменной от одной объясняемой, а все остальные влияющие переменные рассматривать в качестве контрольных.

## Подходы к проблеме эндогенности

### Сопоставление схожих наблюдений

В примере с магистратурой, чтобы получить корректную оценку для коэффициента при (), можно было бы сравнить заработную плату у выпускников с похожими уровнями . Если у нас, например есть достаточно много наблюдений для выпускников с уровнем достаточно близким к некоторой величине ( для малого ), и среди этих наблюдений достаточно много как таких, для которых , так и таких, для которых , то мы могли бы построить регрессию, дающую корректную оценку. В этой регрессии мы бы так же, как и в исходной проблемной модели, сравнили средний уровень логарифма заработной платы в двух группах выпускников – учившихся в магистратуре и не учившихся, но уровень в этих двух группах был бы уже сопоставимым. Конечно, здесь мы должны предположить, что после учета все остальные факторы, которые могут повлиять на , не влияют существенно на заработную плату. Например, это могут быть какие-то семейные обстоятельства и т. п.

Понятно, что данный подход имеет свои ограничения. В частности, количество наблюдений в сопоставимых группах может быть слишком малым.

### Включение контрольных переменных

Самым важным способом, позволяющим решать проблемы эндогенности, является включение в модель контрольных переменных.

В примере с магистратурой можно использовать уровень в качестве контрольной переменной. В простейшем случае мы получим модель следующего вида:

Проиллюстрируем последствия включения в модель графически на условном примере (см. рис. 78). (С другой стороны, это же будет иллюстрация смещения из-за пропуска .) Наблюдения для выпускников, закончивших магистратуру (), показаны закрашенными треугольниками, а для выпускников, не закончивших магистратуру (), – кругами. Также показаны линии для двух регрессий. Первая регрессия не содержит , и поэтому соответствующие линии горизонтальны. Второй регрессии, в которую включена, соответствуют линии с положительным наклоном. Для каждой регрессии имеем здесь две параллельные линии. Сплошная линия соответствует , а пунктирная – . Расстояние по вертикали между параллельными линиями – это коэффициент при фиктивной переменной . Он измеряет разрыв в заработных платах закончивших магистратуру и не закончивших. По регрессии без в нашем условном примере разрыв примерно равен 0.15, т. е. около 15 %. По регрессии с этот разрыв гораздо скромнее – он примерно равен 0.05 (5 %). Таким образом, из-за не включенной контрольной переменной эффект магистратуры в исходной модели преувеличивается примерно на 10%.

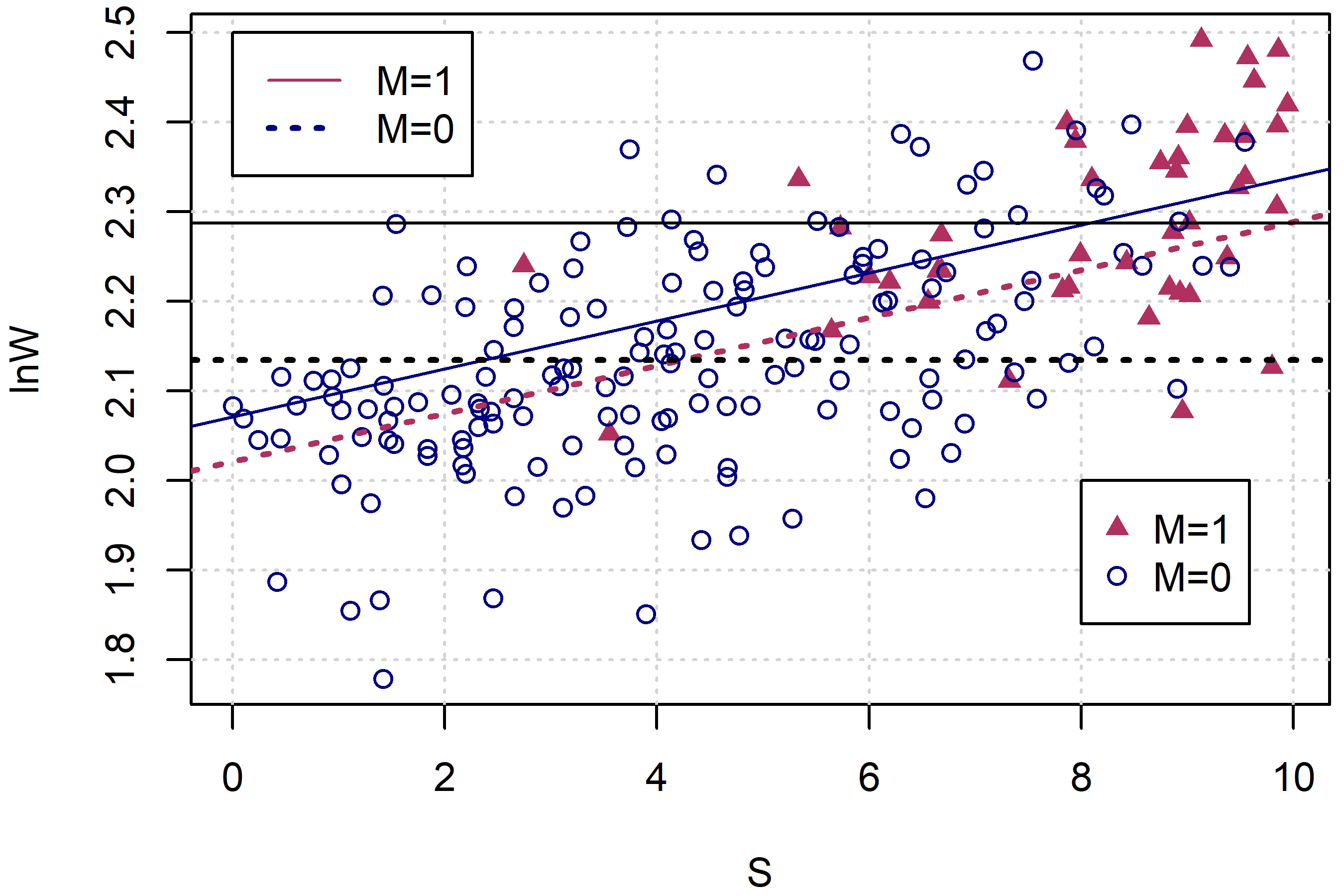


Рисунок 78. Условный пример смещения во влиянии магистратуры

Проблема часто состоит в том, что нужные переменные могут быть нам недоступны, поэтому их невозможно включить в регрессию в качестве контрольных. Но, по крайней мере, когда «подозрительные» переменные доступны, то не надо забывать их включать в качестве контрольных, чтобы не получить смещения.

### Частный случай контрольных переменных – нормировка

Один из простейших и стандартных приемов, служащих для контроля за теми факторами, которые могут влиять как на зависимую переменную, так и на объясняющие – это **нормировка**, то есть приведение показателей к сопоставимому виду.

В частности, когда рассматривается регрессия по странам или регионам одной страны, то принято делать нормировку на душу населения, чтобы учесть различия в размерах стран (регионов).

Рассмотрим характерный пример. Пусть  – инвестиции в основной капитал в регионах России в 2018 г. (млн руб.),  – число заражений COVID-19 по регионам России на 29 мая 2020 г. (человек).[[7]](#footnote-7) На рис. 79 приведена точечная диаграмма этих показателей в логарифмической шкале. Зависимость лучше строить в логарифмах, поскольку значения переменных для разных регионов различаются во много раз.

Если построить регрессию логарифма заражений от логарифма инвестиции, то регрессия покажет значимую положительную зависимость с . Такой результат несколько парадоксален (вряд ли инвестиции способствуют росту заболеваемости), но легко объясним. Дело в том, что есть регионы разного размера. В крупных регионах, например, в г. Москва, как объем инвестиций, так и число зараженных большие. Это пример того, что называют **ложной регрессией**.



Рисунок 79

Для того чтобы ложной зависимости не наблюдалось, в регрессию достаточно добавить переменную, которая отвечает за размер региона. Самый простой вариант такой переменной ‍– это численность населения в регионе. Пусть – это оценка численности постоянного населения на 1 января 2020 г. (человек). Логарифм населения довольно сильно коррелирует как с логарифмом заражений , так и с логарифмом инвестиций (коэффициенты корреляции 0.78 и 0.73 соответственно).

Добавим логарифм численности населения в регрессию. Получим следующий результат (в квадратных скобках *p*-значения):

Видим, что логарифм инвестиций в этой регрессии оказался незначимой переменной.

Можно предположить, что как , так и , просто приблизительно пропорциональны численности населения. Имеет смысл оценить последнюю модель, используя среднедушевые показатели:

Эта регрессия по смыслу не отличается от предыдущей, но, поскольку мы нормировали зависимую переменную – число заражений, коэффициент детерминации оказался очень низким. Здесь мы не нашли значимой зависимости числа заражений на душу населения как от объема инвестиций на душу населения, так и от логарифма численности населения.

Можно дать следующее объяснение полученному результату. Численность населения определяет как число заражений, так и объем инвестиций, а сами по себе, при сравнимой численности населения, заражения и инвестиции практически не связаны. Такую ситуацию можно представить в виде диаграммы со стрелками (см. рис. 80).

Рисунок 80

Переменную, которая может влиять как на объясняемую переменную, так и на объясняющую, создавая ложную взаимосвязь между ними, в статистике принято называть вмешивающимся фактором (англ. *confounding factor*).

Можно было с самого начала взять среднедушевые показатели. При этом оцененная регрессия имеет вид

Зависимость очевидно незначимая.

Правда, *не включать в регрессию нормирующую переменнуюне всегда безопасно* по тем же самым причинам, связанным с пропуском контрольной переменной. Чтобы проиллюстрировать возможную проблему, используем те же данные. Введем обозначения для наших среднедушевых показателей: и . Как мы только что видели, и коррелируют между собой. Если мы забудем, что пронормировали наши переменные и поделим их на численность населения еще раз, то получим следующую регрессию:

(Вообще говоря, можно таким же образом взять две невзаимосвязанные положительные переменные, «нормировать» их, поделив на другую не связанную положительные переменную, и получить ложную взаимосвязь.) Включение в приведенную регрессию логарифма нормирующей переменной в качестве контрольной решает указанную проблему:

Здесь переменная не значима.

Другая часто используемая в экономике процедура нормировки – это **дефлирование**, то есть переход к сопоставимым ценам. Если в регрессии используются временные ряды за много лет, то цены в различные периоды становятся несопоставимыми из-за инфляции. Если использовать показатели в текущих ценах (или, как еще говорят, **номинальные показатели**), то легко получить бессмысленную значимую регрессию просто из-за того, что, как в зависимой переменной, так и в объясняющей, есть компонента, связанная с ростом общего уровня цен. Пусть мы хотим построить регрессию некоторой положительной переменной от другой положительной переменной и пусть – некоторый подходящий индекс цен, который показывает отношение уровня цен в текущем периоде к уровню цен в некотором базовом периоде (индекс в данном контексте называют дефлятором). Тогда можно, во-первых, прологарифмировать наши переменные, а во-вторых, дефлировать их (получив **реальные показатели**), т. е. построить регрессию вида

### Метод инструментальных переменных с одной эндогенной переменной и одним инструментом

Один из классических методов оценивания при наличии проблемы эндогенности – это метод инструментальных переменных.

Пусть мы хотим оценить парную регрессию  
но состоятельное оценивание затруднено тем, что объясняющая переменная может коррелировать с ошибкой . Предположим, что существует некоторая переменная , про которую мы знаем, что она не коррелирует с ошибкой , но достаточно сильно коррелирует с . Подобную переменную принято называть **инструментальной переменной** или просто **инструментом**.

Ковариация инструмента с зависимой переменной равна

Если инструмент обладает свойством экзогенности , то  
откуда при получим

Оценку **метода инструментальных переменных** (англ. *instrumental variables estimator*) для коэффициента наклона можно найти, подставив вместо теоретических моментов соответствующие выборочные моменты:  
где в числителе и знаменателе стоят выборочные ковариации  
В векторном виде

Пользуясь асимптотической теорией можно обосновать сходимость выборочных величин к соответствующим теоретическим величинам:  
и  
Таким образом, при выполнении определенных предположений оценка инструментальных переменных является состоятельной.

На рис. 81 показана схема для причинных связей между переменными, которые требуются для корректности применения метода инструментальных переменных.

Рисунок 81. Схема связей для метода инструментальных переменных

Основная проблема с методом инструментальных переменных состоит в том, что на практике бывает сложно найти хороший инструмент, то есть такой, который достаточно сильно коррелировал с объясняющей переменной и не коррелировал с ошибкой . Для получения надежной оценки требуется, чтобы инструмент одновременно удовлетворял *обоим* указанным условиям. С одной стороны, правдоподобные кандидаты на инструментальные переменные слишком часто оказываются слабо коррелированными с объясняющей переменной (проблема **слабого инструмента**). С другой стороны, тут легко ошибиться и выбрать **негодный инструмент**, для которого предположение об экзогенность не выполняется.

### Метод инструментальных переменных в общем случае

В более общем случае множественной регрессии (здесь мы опять «спрячем» константу)  
можно использовать набор инструментальных переменных , таких что

Пусть – матрица, составленная из наблюдений за инструментальными переменными. Оценка метода инструментальных переменных в общем случае рассчитывается по формуле  
где

Метод инструментальных переменных можно представить в виде двухшагового метода наименьших квадратов.

Шаг 1. Обычным МНК оценивается регрессия каждого из регрессоров от инструментальных переменныхи вычисляются расчетные значения из этой регрессии ==.

Шаг 2. В исходную регрессию вместо исходных переменных подставляются и вычисляются соответствующие оценки МНК ( обозначим их ).

Оценки второго шага совпадут с по приведенной ранее формуле. Действительно, первый шаг можно представить в более компактном матричном виде:

На втором шаге, подставив матрицу вместо , мы получим

Если учтем свойства матрицы , а именно, что  
то увидим, что

Идея метода состоит в том, чтобы заменить эндогенные регрессоры, такими их приближениями , которые были бы основаны на экзогенных переменных . Это как бы «очищенные» от ошибки варианты исходных переменных .

Если врегрессорах есть экзогенные переменные, то они по смыслу метода входят также и в и для них поэтому будет выполнено =. В частности, это относится к вектору из единиц, соответствующему константе регрессии.

Для вычисления оценки требуется, чтобы матрица была обратимой, т. е. чтобы столбцы матрицы «очищенных» регрессоров были линейно независимы. Это **условие идентификации** в контексте метода инструментальных переменных. По меньшей мере требуется, чтобы в матрице было не меньше переменных, чем в матрице , т. е. чтобы . Другими словами, каждому эндогенному регрессору должен соответствовать по меньшей мере один внешний инструмент, т. е. такая инструментальная переменная, которая не входит в число регрессоров. («Каждому жителю – по огнетушителю!») Если , то уравнение точно идентифицировано.

Из-за того, что выбранные инструментальные переменные слабо связаны с исходными регрессорами матрица может быть близка к вырожденности. Это случай так называемой **слабой идентификации** и слабых инструментов. Использование слабых инструментов приводит к целому ряду проблем, поэтому следует следить, чтобы была сильная связь между регрессорами и инструментами. Об этом, например, можно судить по коэффициентам детерминации первого шага двухшагового МНК.

Ковариационная матрица оценок инструментальных переменных равна примерно  
где – остаточная дисперсия. Следует следить, чтобы была рассчитана по павильным остаткам:  
(Остатки второго шага двухшагового МНК здесь не годятся. Поэтому обычные *t*- и *F*-статистики со второго шага использовать не следует.)

### Использование разрывности в регрессии

Рассмотрим снова пример с магистратурой и сделаем следующие предположения:

* Все выпускники бакалавриата, для которых мы хотим узнать зависимость логарифма заработной платы от факта учебы в магистратуре , имеют желание и потенциальную возможность учиться в магистратуре и участвуют во вступительном экзамене.
* Каждый получил некоторые непрерывные баллы на экзамене, и в магистратуру попали те, у кого .
* Участники экзамена не могут гарантировать себе определенные баллы. На влияют разные неподконтрольные им случайные факторы.
* Все, кто набрал нужное количество баллов, в дальнейшем учатся в магистратуре. Таким образом, тогда и только тогда, когда и иначе.
* Выпускники бакалавриата, у которых баллы немного ниже , мало отличаются от выпускников, у которых баллы немного выше . В том числе, пусть это касается и их уровня , т. е. у всех выпускников с баллами, близкими к , примерно одинаковое распределение уровня .

Если указанные предположения выполнены, то мы можем взять всех выпускников с баллами из малой окрестности , т. е. всех с для малого , и сравнить только среди них среднюю зарплату учившихся и не учившихся в магистратуре.

Это особый вариант сопоставления схожих наблюдений, о котором говорилось ранее, основанный на разрыве, который имеется в зависимости индикаторной объясняющей переменной () от некоторой непрерывной переменной (). Большое преимущество этого варианта состоит в том, что тут не обязательно знать уровень или других влияющих факторов. (Хотя, если подобная переменная известна, то ее тоже можно и даже очень желательно включить в регрессию.). Идея в том, что описанная ситуация в чем-то похожа на ситуацию, когда с выпускниками проводится рандомизированный эксперимент. Локально в районе границы для выпускников со схожими характеристиками действуют случайные факторы, влияющие на баллы и определяющие попадание в одну из двух групп: и .

Конечно, в рассмотренной схеме, использующей разрыв в регрессии, есть много допущений (см. список выше), не все из которых правдоподобны. Например, в реальности не все выпускники бакалавриата могут захотеть участвовать в конкурсе на поступление, не все поступившие сумеют доучиться до конца и т. д., а это уже приведет к эндогенности. Кроме того, количество наблюдений в интервалах и может оказаться слишком малым для надежной оценки различий в зарплате.

## Контрольные вопросы

1. Как с помощью графиков диагностировать отсутствие нормальности?
2. Записать формулу выборочных коэффициентов асимметрии и куртозиса. Чему равны коэффициенты асимметрии и куртозиса нормального распределения?
3. Для чего и как используется тест Харке—Беры ?
4. Основные следствия отсутствия нормальности ?
5. Объясните, что означает асимптотическая нормальность.
6. Что такое медианная регрессия?
7. В чем суть ядерной оценки плотности?

## Экзаменационные вопросы

1. Диагностика отсутствия нормальности и наличия выбросов.
2. Асимптотическая нормальность, проверка гипотез и следствия отсутствия нормальности.
3. Основные последствия отсутствия нормальности. Что делать при сильном отклонении ошибок он нормальности?

## Литература

1. Вербик М. Путеводитель по современной эконометрике. – М.: Научная книга,. 2008. [стр.73, 292-295 ]

# Приложения (базовые понятия из курсов теории вероятностей, математической статистике и высшей математики)

Для успешного изучения учебного курса «Эконометрия» студенты должны иметь базовые знания по теории вероятностей и математической статистике, линейной алгебре и математическому анализу. Рассмотрим некоторые основные понятия из этих курсов.

## Основные понятия теории вероятностей

• Функцией распределения случайной величины называется функция , сопоставляющая числу вероятность того, что не превышает . Функция распределения полностью характеризует отдельную случайную величину.

• Если случайная величина непрерывна, то она имеет плотность , которая связана с функцией распределения соотношением .

Функция распределения имеет следующие свойства: это неубывающая, непрерывная справа функция, , причем и .

; .

• Вероятность того, что , равна .

• Квантилью уровня , где , (-квантилью) непрерывной случайной величины называется число , такое что

• Медианой называется -квантиль.

• Модой непрерывной случайной величины называется величина, при которой плотность распределения достигает максимума, т. е.

• Если распределение непрерывной случайной величины симметрично относительно нуля, т. е. и , то двусторонней *p*-квантилью называется число , такое что

• Математическим ожиданием непрерывной случайной величины называется .

Математическое ожидание является начальным моментом первого порядка. Начальным моментом *q*-го порядка называется .

• По случайной величине может быть построена соответствующая ей центрированная величина , имеющая аналогичные законы распределения и нулевое математическое ожидание.

• Центральным моментом *q*-го порядка случайной величины называется начальный момент -го порядка для соответствующей центрированной величины, т. е. . Для непрерывной случайной величины центральный момент -го порядка равен

• Дисперсией случайной величины называется центральный момент второго порядка:

Для непрерывной случайной величины дисперсия равна

• Среднеквадратическим отклонением называется квадратный корень из дисперсии  
• Нормированной (стандартизованной) случайной величиной называется  
=0 и =1

• Коэффициентом асимметрии называется начальный момент третьего порядка нормированной случайной величины, т. е.  
Если , то распределение величины симметрично, если , то имеет место правая асимметрия, если , то левая асимметрия.

• Куртозисом называется начальный момент четвертого порядка нормированной случайной величины, т. е.

Если , то распределение близко к нормальному, если то распределение высоко вершинное, если – низко вершинное. Коэффициентом эксцесса называется .

• Для *k*-мерного случайного вектора (многомерной случайной величины) функцией распределения называется

Если распределение случайного вектора непрерывно, то он имеет плотность (называемую совместной плотностью случайных величин ), которая связана с функцией распределения соотношениями

• Для многомерной случайной величины

Случайные величины называются независимыми (в совокупности), если .

Случайные величины называются независимыми (в совокупности), если .

• Ковариацией случайных величин и называется

• Корреляцией случайных величин и называется

.

• Ковариационной матрицей *k*-мерной случайной величины называется матрица размерности на

• Корреляционной матрицей *k*-мерной случайной величины называется

Свойства математическое ожидание:

• Если – константа, то .

• Если и – любые две случайные величины, то

• Если – константа, то .

• В общем случае . Если , то

• Для симметричного распределения выполено .

Свойства дисперсия:

• .

• Для любой случайной величины выполнено .

• Если – константа, то выполнено:

; ; .

• Если и – любые две случайные величины,

Если и независимы, то .

Свойства ковариации :

• .

• .

• .

• .

• .

• Если и независимы, то . Обратное, вообще говоря, неверно.

### Проверка гипотез

Пусть – случайная выборка из распределения , заданного параметром . (Обычно это распределение хи-квадрат**,** Стьюдента и Фишера)**.**

Общаязадача – это получении выводов о параметре на основании наблюдений .

Нулевая гипотеза относительно параметра состоит в том, что он принадлежит некоторому более узкому множеству: , где .

Альтернативная гипотеза состоит в том, что параметр принадлежит другому множеству: , где . Рассматривается некоторая *статистика* , которая является функцией от выборки: . Процедуру (правило) проверки гипотезы называют *статистическим критерием* или статистическим тестом. Суть проверки гипотезы против альтернативной гипотезы состоит в том, что задаются две непересекающиеся области, и , такие что – вся область значений статистики . Если , то нулевая гипотеза () принимается, а если , то нулевая гипотеза отвергается.

Обычно и , где – критическая граница. Такой критерий называется односторонним. При этом критерий состоит в следующем:  
, то принимается,  
если , то отвергается.

и выбираются так, чтобы в случае, когда верна, вероятность того, что , была бы равна некоторой заданной малой вероятности . Как правило, на практике используют вероятность (хотя это не имеет под собой каких-либо теоретических оснований).

*Ошибкой первого рода* называется ошибка, состоящая в том, что отвергается *верная* нулевая гипотеза (). Вероятность ошибки первого рода равна . Вероятность ошибки первого рода называется уровнем значимости или размером. Вероятность называют уровнем доверия.

Ошибкой второго рода называется ошибка, состоящая в том, что принимается неверная нулевая гипотеза. Вероятность ошибки второго рода обозначают .

Мощностью критерия называют величину . Мощность характеризует, насколько хорошо работает критерий. Мощность должна быть как можно большей при данном . Требуется, по крайней мере, чтобы . Критерий, не удовлетворяющий этому условию, называют смещенным.

Альтернативный способ проверки гипотез использует вероятность ошибки первого рода, если принять равной , т. е. вероятность того, что . Эту вероятность называют уровнем значимости или P-значением. Обозначим ее . При заданной вероятности критерий состоит в следующем:  
, то принимается,  
если , то отвергается.

Еще один способ проверки гипотез основан на доверительных областях для параметра . Пусть – доверительная область для параметра , такая что с некоторой веротностью , и пусть проверяется гипотеза : против альтернативной гипотезы : . Критерий состоит в следующем:  
если , то принимается,  
если , то отвергается.

Отметим, что в этом случае не случайная величина; случайной является доверительная область .

### Свойства оценок:

Пусть некоторая оценка параметра θ. Качество оценки характеризуется наличием или отсутствием некоторых важных свойств – несмещенности, состоятельности и эффективности.

Оценка параметра θ называется *несмещенной*, если Требование несмещенности означает отсутствие некоторой системной, постоянно присутствующей ошибки, которая бы завышала оценку или занижала ее . Требование несмещенности особо важно при малом количестве наблюдений. Если при , то оценка называется *асимптотически несмещенной*.

Оценка параметра θ называется *состоятельной*, если для любого сколь угодно малого положительного числа имеем Требование состоятельности означает, что при увеличении объема выборки мы все ближе приближаемся к истинному значению параметра. Такое стремление называется сходимостью по вероятности – вероятность больших отличий между и θ стремится к нулю. Если оценка является несмещенной и ее дисперсия стремится к нулю ), то оценка является и *состоятельной*.

Оценка параметра называется эффективной, если ее дисперсия является наименьшей из всех возможных оценок параметра θ по выборкам объема n. Требование эффективности означает наименьший разброс вокруг своего среднего. Это требование важно для несмещенных оценок, когда их среднее (то есть математическое ожидание) совпадает с истинным значением параметра. Тогда и наименьший разброс оказывается по отношению к истинному значению параметра.

## Распределения, используемые в эконометрии

### Равномерное распределение

Непрерывная случайная величина равномерно распределена на отрезке , если плотность задается формулой

### Нормальное распределение

Непрерывная случайная величинаимеет нормальное (или гауссовское) распределение с математическим ожиданием и дисперсией обозначается если плотность распределения задается формулой

Нормальное распределение симметрично относительно , и для него выполняется .

Моменты нормального распределения: и  
 при целых , в частности, .

Коэффициент асимметрии:.

Куртозис , коэффициент эксцесса равен нулю.

Если , то имеет стандартное нормальное распределение. Его плотность ; функция распределения . Для стандартного нормального распределения 95-процентная двусторонняя квантиль равна 1.96, а 99-процентная квантиль равна 2.57.

### Распределение хи-квадрат

Распределение хи-квадрат с степенями свободы имеет плотность:  
где – гамма-функция:

Если , и независимы в совокупности, то случайная величина имеет распределение хи-квадрат с степенями свободы.  
 и .

Коэффициент асимметрии: .

Куртозис: , коэффициент эксцесса .

При больших распределение хи-квадрат похоже на и

95-процентная (односторонняя) квантиль распределение хи-квадрат при k=1 равна 3.84 (квадрат 1.96), при k=5 – 11.1, при k=20 – 31.4, при k=100 – 124.3.

### Распределение Стьюдента

Распределение Стьюдента с степенями свободы обозначается через . Его также называют -распределением. Его плотность:

Если , и независимы в совокупности, то случайная величина

Распределение Стьюдента симметрично относительно нуля и .

Математическое ожидание существует при и .

При не существует *k*-го момента.

Дисперсия (существует при ).

Коэффициент асимметрии (существует при ).

Куртозис ; коэффициент эксцесса (существуют при ).

При больших распределение Стьюдента похоже на .

95-процентная двусторонняя квантиль . при k=1 равна 12.7, при k=5 — 2.57, при k=20 — 2.09, при k=100 — 1.98.

### Распределение Фишера

Распределение Фишера с и степенями свободы обозначается . Его также называют F-распределением или распределением Фишера---Снедекора. Его плотность:

Если , , , и независимы в совокупности, то

Если , то

При больших величина похожа на .

95-процентная (односторонняя) квантиль при k2=1 равна 161, при k2=5 – 6.61, при k2=20 – 4.35, при k2=100 – 3.94 (квадраты соответствующих ); квантиль при k2=1 равна 200, при k2=5 – 5.79, при k2=20 – 3.49, при k2=100 – 3.09; квантиль при k1=3 равна 3.10, при k1=4 – 2.87, при k1=5 – 2.71, при k1=6 – 2.60.

### Многомерное нормальное распределение

-мерное нормальное распределение с математическим ожиданием () и ковариационной матрицей () обозначается . Его плотность:

#### Свойства многомерного нормального распределения:

• Если , то

• Если , то .

• Если , где () – невырожденная матрица, то .

• Если , где () – матрица, имеющая полный ранг по столбцам, то .

• Если , где () – матрица, имеющая полный ранг по столбцам, то .

• Если , то независимы в совокупности и .

## Вспомогательные сведения из высшей математики.

### Матричная алгебра

называется вектор-столбцом размерности .

называется вектор-строкой размерности .

называется матрицей размерности .

• Сумма матриц и (): , ().

• Произведение матриц () и (): , .

• Скалярное произведение вектор-столбцов () и (): .

• Квадратичная форма вектор-столбца () и матрицы ():

• Произведение матрицы () на скаляр : , ().

• Транспонирование матрицы (): , ().

• След матрицы (): .

• Рангом () матрицы называется количество линейно независимых столбцов (равное количеству линейно независимых строк). Матрица () имеет полный ранг по столбцам, если . Матрица () имеет полный ранг по строкам, если .

• Матрица () называется невырожденной (неособенной), если . В противном случае она называется вырожденной.

• Матрица () называется диагональной, если при . Для диагональной матрицы используется обозначение .

• Матрица () называется единичной.

• Матрица () называется симметричной (симметрической), если  
.

•Матрица () называется верхней треугольной, если при . Матрица () называется нижней треугольной, если при .

• Матрица () называется обратной матрицей к матрице (), если .

• Матрица () называется идемпотентной, если .

• Вектор-столбцы () и () называются ортогональными, если их скалярное произведение равно нулю: .

• Матрица (), где , называется ортогональной, если ее столбцы ортогональны, т. е. .

• Матрица () называется положительно определенной, если для любого вектор-столбца () выполняется . Матрица () называется отрицательно определенной, если для любого вектор-столбца () выполняется .

• Матрица () называется положительно полуопределенной (неотрицательно определенной), если для любого вектора-столбца () выполняется . Матрица () называется отрицательно полуопределенной (неположительно определенной), если для любого вектора-столбца () выполняется .

• Определителем матрицы () называется  
где – номер любой строки, а матрицы () получены из матрицы путем вычеркивания *i*-й строки и -го столбца.

• Для матрицы () уравнение называется характеристическим уравнением. Решение этого уравнения называется собственным числом (собственным значением) матрицы . Вектор () называется собственным вектором матрицы , соответствующим собственному числу , если .

• Прямое произведение (произведение Кронекера) матриц () и () это матрица ():

### Свойства матриц

#### Сложение матриц

•  (коммутативность).

•  (ассоциативность).

#### Произведение матриц

• В общем случае (свойство коммутативности не выполнено).

•  (ассоциативность).

•  (дистрибутивность).

•  для матрицы ().

• .

#### Ранг

• Для матрицы () выполнено .

• .

• Если матрица () является невырожденной, то для матрицы () выполнено . Если матрица () является невырожденной, то для матрицы () выполнено .

• .

#### Cлед

• .

• .

• .

• .

• .

• .

• .

• , где матрица () имеет полный ранг по столбцам, т. е. .

• , где и – квадратные матрицы.

#### Транспонирование

• .

• .

#### Определитель

• Для матрицы (): .

• .

• .

•  для матрицы ().

• .

• .

• Если матрица () является треугольной (например, диагональной), то .

• .

• .

•  для матрицы () и вектор-столбцов , ().

• , где и – квадратные матрицы.

• , где и – квадратные невырожденные матрицы.

• Матрица () является невырожденной () тогда и только тогда, когда .

#### Обращение

• Если обратная матрица существует, то она единственна (в частности, левая и правая обратные матрицы совпадают).

• Матрица () имеет обратную тогда и только тогда, когда она является невырожденной, т. е. .

• Матрица () имеет обратную тогда и только тогда, когда .

• Обозначим через элементы обратной матрицы . Тогда

() получены из матрицы путем вычеркивания *j*-й строки и -го столбца.

Во всех приводимых ниже формулах предполагается, что существуют обратные матрицы там, где это требуется.

• Для матрицы ():

•  тогда и только тогда, когда .

• .

• .

• .

• Если () – ортогональная матрица, то .

• Для диагональной матрицы выполнено:

• .

• .

• .

• , где и – квадратные матрицы.

• ,  
где и – квадратные матрицы.

#### Положительно определенные матрицы

• Если матрица положительно определенная, то . Если матрица положительно полуопределенная, то .

• Если матрица положительно (полу-)определенная, то матрица отрицательно (полу-)определенная.

• Если матрица положительно определенная, то обратная матрица также положительно определенная.

• Если матрицы и положительно (полу-)определенные, то матрицы и также положительно (полу-)определенные.

• Если матрица положительно определенная, а положительно полуопределенная, то . Если положительно определенная, то .

• Матрицы и () являются симметричными положительно полуопределенными для любой матрицы () и симметричной положительно полуопределенной матрицы ().

• Если матрица () имеет полный ранг по столбцам, то матрица () симметричная положительно определенная. Если матрица () симметричная положительно определенная, то матрица () симметричная положительно определенная.

• Если матрица () положительно полуопределенная, то существует верхняя треугольная матрица (), такая что . Также существует нижняя треугольная матрица (), такая что . Такое представление матрицы называется разложением Холецкого (триангуляризацией).

#### Идемпотентные матрицы

• Если матрица идемпотентная, то матрица тоже идемпотентная, причем

• Если матрица симметричная и идемпотентная, то .

• Матрицы и являются симметричными и идемпотентными для любой матрицы (), имеющей полный ранг по столбцам. При этом и .

#### Собственные числа и векторы

• Для матрицы () является многочленом *m*-й степени (характеристическим многочленом) и имеет корней, , в общем случае комплексных, среди которых могут быть кратные. По определению, являются собственными числами матрицы .

• У матрицы () существует не больше различных собственных чисел.

• Если – собственный вектор матрицы , соответствующий собственному числу , то для любого скаляра , – тоже собственный вектор, соответствующий собственному числу .

• Если – собственные числа матрицы , то , .

• Если матрица идемпотентная, то все ее собственные числа равны или .

• Все собственные числа вещественной симметричной матрицы вещественны.

• Если и – собственные векторы вещественной симметричной матрицы, соответствующие двум различным собственным числам, то они ортогональны: .

• Если матрица () является вещественной и симметричной, то существуют матрицы и , где () – ортогональная матрица (), столбцы которой – собственные векторы матрицы , а () – диагональная матрица, состоящая из соответствующих собственных чисел матрицы , такие что выполнено .

• Если матрица () является вещественной, симметричной, невырожденной, то .

• Вещественная симметричная матрица является положительно полуопределенной (определенной) тогда и только тогда, когда все ее собственные числа неотрицательны (положительны). Вещественная симметричная матрица является отрицательно полуопределенной (определенной) тогда и только тогда, когда все ее собственные числа неположительны (отрицательны).

• Если матрица () является вещественной, симметричной и положительно полуопределенной, то , где () – вещественная, симметричная и положительно полуопределенная матрица; .

• Пусть – собственные числа вещественной симметричной матрицы (). Тогда собственый вектор , соответствующий наименьшему собственому числу , является решением задачи

• Пусть – собственные числа вещественной симметричной матрицы (). Тогда и .

#### Произведение Кронекера

•  и .

• .

• .

• .

• .

• .

•  для матриц () и ().

• .

• .

### Матричное дифференцирование

#### Определения

• Производной скалярной функции по вектор-столбцу () или, другими словами, градиентом является вектор-столбец ()

• Производной скалярной функции по вектор-строке () является вектор-строка ()

• Производной векторной функции () по вектору () или, другими словами, матрицей Якоби является матрица ()

• Производной векторной функции () по вектору () является матрица ()

• Производной скалярной функции по матрице () является матрица ()

• Производной матричной функции по скаляру является матрица ()

• Второй производной скалярной функции по вектору-столбцу () или, другими словами, матрицей Гессе является матрица ()

#### Свойства

•  и .

•  и .

• .

• .

•  и .

• .

Для симметричной матрицы : .

• .

• .

• .

•  и .

• .

• .

• .

• .

• .

• .

• .

• .

• .

• .

• .

• .

1. Строго говоря, абсолютно непрерывную. [↑](#footnote-ref-1)
2. Например, оценки двухшагового МНК можно рассчитать с помощью нескольких вспомогательных регрессий. В теме прогнозирования будет показано, как с помощью вспомогательной регрессии рассчитать точечный прогноз и его стандартную ошибку. [↑](#footnote-ref-2)
3. Использованы данные по домохозяйствам 27 раунда Российского мониторинга экономического положения и здоровья населения. См. <https://www.hse.ru/rlms/>. [↑](#footnote-ref-3)
4. Такая группировка вносит смещение в оценки, но смещение здесь в принципе неизбежно, поскольку мы используем не настоящий долгосрочный доход, а не очень совершенный его заменитель. [↑](#footnote-ref-4)
5. Использовался широко известный набор данных diamonds из пакета ggplot2 для R, в котором более 50 тыс. наблюдений. Алмазы отличаются качеством. Мы взяли только те наблюдения, для которых cut=Ideal, color=G и clarity=VS1. [↑](#footnote-ref-5)
6. См. в библиотеке library(quantreg) для R данные data(engel). [↑](#footnote-ref-6)
7. Данные с сайтов <https://gks.ru/bgd/regl/b19_14p/Main.htm>, <https://covid19.rosminzdrav.ru/> и <https://www.gks.ru/storage/mediabank/Popul2020.xls>. Дата доступа 2020-05-29. [↑](#footnote-ref-7)