Московский государственный университет имени М.В. Ломоносова
Московская школа экономики
Институт социально-экономического развития территорий
Российской академии наук

Е.А. Ивин, Н.В. Артамонов, А.Н. Курбацкий

# МЕТОДИЧЕСКОЕ ПОСОБИЕ ПО ЭКОНОМЕТРИКЕ (сокращённый вариант)

для социально-экономических специальностей

Вологда
2016

УДК 330.43
ББК 65в631
И25

**Ивин, Е. А.**
**И25** Методическое пособие по эконометрике: для социально-экономических специальностей [Текст] / Е. А. Ивин, Н. В. Артамонов, А. Н. Курбацкий. – Вологда: ИСЭРТ РАН, 2016. – 184 с.

ISBN 978-5-93299-331-6

Настоящее пособие предназначено для самостоятельного знакомства и освоения основных методик эконометрических исследований в области анализа экономических и социально-экономических данных. Это пособие написано на основе курса, прочтенного авторами для аспирантов и сотрудников ИСЭРТ РАН, а также обсуждения изложенного материала слушателями. Каждый раздел снабжен кратким сводом основных понятий и теорем, необходимых для решения задач, а также решениями типовых задач по ключевым темам.

Книга адресована студентам и аспирантам социально-экономических специальностей, а также преподавателям.

УДК 330.43
ББК 65в631

*Рецензент:*

**В.А. Ильин**
доктор экономических наук, профессор, научный руководитель
Института социально-экономического развития территорий
Российской академии наук, заслуженный деятель науки РФ

ISBN 978-5-93299-331-6

© Ивин Е.А., Артамонов Н.В., Курбацкий А.Н., 2016
© ИСЭРТ РАН, 2016

# Оглавление


**1 Линейная регрессия в условиях Гаусса-Маркова** **4**

1.1. Статистические данные в эконометрике . . . . . 4

1.2. О корреляционном и регрессионном анализе . . . . . 5

1.3. Парный коэффициент корреляции . . . . . 6

1.3.1. Коэффициент корреляции . . . . . 6

1.3.2. Выборочный коэффициент корреляции . . . . . 8

1.4. Линейная аппроксимация. Метод наименьших квадратов . . . . . 11

1.5. Парная линейная модель регрессии . . . . . 13

1.5.1. Теорема Гаусса – Маркова . . . . . 15

1.5.2. Статистические свойства OLS-оценок коэффициентов . . . . 19

1.5.3. Доверительные интервалы. Проверка гипотез . . . . . 21

1.5.4. Коэффициент $R^2$ и «качество подгонки» . . . . . 24

1.6. Прогнозирование в модели парной регрессии . . . . . 26

1.7. Парная регрессия без константы . . . . . 29

1.8. Нелинейные модели . . . . . 30

1.9. Стохастические регрессоры . . . . . 33

1.10. Задачи . . . . . 34

**2 Многофакторная регрессия** **39**

2.1. Многофакторная регрессия . . . . . 39

2.2. Метод наименьших квадратов . . . . . 40

2.3. Основные предположения.
Теорема Гаусса – Маркова . . . . . 41

2.4. Статистические свойства OLS-оценок . . . . . 44

2.5. Коэффициент $R^2$. Проверка сложных гипотез о коэффициентах регрессии . . . . . 47

2.6. Прогнозирование в линейной модели регрессии . . . . . 53

2.7. Нелинейные модели . . . . . 54

2.8. Стохастические регрессоры . . . . . 58

2.9. Задачи . . . . . 59

**3 Разные аспекты линейной регрессии** **69**

3.1. Фиктивные переменные . . . . . 69

3.2. Тест Чоу . . . . . 74

3.3. Мультиколлинеарность . . . . . 75

3.4. Спецификация модели регрессии . . . . . 78
3.4.1. Невключение в модель значимого фактора . . . . . 78
3.4.2. Включение в модель незначимого фактора . . . . . 80
3.4.3. Сравнение вложенных моделей . . . . . 81
3.4.4. Информационные критерии . . . . . 81
3.4.5. Сравнение невложенных моделей . . . . . 82
3.5. Выбор функциональной формы зависимости . . . . . 83
3.5.1. Тест на функциональную форму . . . . . 84
3.6. Задачи . . . . . 87

**4 Гетероскедастичность и автокоррелированность ошибок 91**
4.1. Гетероскедастичность ошибок регрессии . . . . . 91
4.1.1. Тесты на гетероскедастичность . . . . . 92
4.1.2. Корректировка на гетероскедастичность . . . . . 98
4.2. Корреляция во времени ошибок регрессии . . . . . 103
4.2.1. Автокорреляция первого порядка . . . . . 104
4.2.2. Автокорреляция произвольного порядка . . . . . 108
4.3. Корректировка модели на гетероскедастичность и автокорреляцию 110
4.4. Задачи . . . . . 113

**5 Модели с дискретными зависимыми переменными 121**
5.1. Дискретная зависимая переменная . . . . . 121
5.2. LPM-модель, спецификация и свойства . . . . . 122
5.2.1. LPM-модель: прогноз . . . . . 122
5.2.2. LPM-модель: интерпретация коэффициентов . . . . . 123
5.2.3. LPM-модель: гетероскедастичность ошибки . . . . . 123
5.2.4. Особенности модели . . . . . 123
5.3. Нелинейные модели . . . . . 124
5.3.1. Латентная переменная . . . . . 124
5.3.2. Интерпретация коэффициентов . . . . . 125
5.3.3. Logit- и Probit-модель . . . . . 125
5.3.4. Проблема оценивания. Метод максимального правдоподобия 126
5.3.5. Общее качество модели . . . . . 127
5.3.6. Проверка гипотез . . . . . 128
5.3.7. Прогнозирование . . . . . 129
5.3.8. Информационные критерии. Выбор модели . . . . . 129
5.4. Задачи . . . . . 130

**A Статистические таблицы 136**

**B Список обозначений 149**

**Литература 150**

# Глава 1

# Линейная регрессия в условиях Гаусса-Маркова

## 1.1. Статистические данные в Эконометрике

В современной эконометрике различаются следующие основные классы выборочных статистических данных:

- пространственные выборки (cross-sectional data);
- временные ряды (time series);
- панельные данные (panel data).

Пространственные выборки характеризуются тем, что выборочные данные получены в один (или очень близкие) период времени и их следует рассматривать как случайную выборку из некоторой генеральной совокупности (population). Примеры таких выборок дают опросы людей, домашних хозяйств, статистические данные по фирмам, городам, странам. Преимущественно такие выборки мы будем рассматривать в данном пособии.

Временны́е ряда (по одному или нескольким факторам) представляют собой статистические данные, полученные в результате наблюдения в течение некоторого промежутка времени. Как правило, эти данные получены через равные промежутки (кванты) времени. Основное отличие от пространственных выборок состоит в следующем. Фактор времени естественным образом упорядочивает данные временного ряда (в хронологическом порядке), в то время как в пространственных выборках такой естественный порядок отсутствует. Во многих случаях временные ряды уже нельзя рассматривать как реализацию случайной выборки, так как естественно полагать, что на текущие значения могут оказывать влияние прошлые значения временного ряда («эффект памяти»). Примеры временных рядов дают, например, финансовые данные (котировки и биржевые индексы), индексы цен, макроэкономические данные (ВВП, уровень инфляции и безработицы) и др.

Панельные данные являются обобщением первых двух классов данных: эти данные состоят из временных рядов по **каждому** члену пространственной выборки. Другими словами, мы имеем набор пространственных выборок для **одних и тех же** объектов, полученные в разные моменты времени. Пример панельных данных дают полученные в течение нескольких лет данные об **одних и тех же** домашних хозяйствах или индивидуумах.

Иногда для увеличения выборочной информации используются pooled cross sectional data, представляющие собой объединение нескольких пространственных выборок, полученных в разные периоды времени. В отличие от панельных данных, pooled data формируется из пространственных выборок, для разных объектов генеральной совокупности в **разные** моменты времени.

## 1.2. О корреляционном и регрессионном анализе

Основным аппаратом эконометрики является раздел математической статистики: корреляционно-регрессионный анализ. Задача корреляционного анализа – выявление характера и степени взаимосвязи между экономическими показателями, являющимися случайными величинами. Задача регрессионного анализа – выявление того, насколько изменение одной экономической переменной (фактора) в среднем влияет на изменение другой экономической переменной (результативного признака). В корреляционном анализе определяется один показатель, характеризующий степень тесноты взаимосвязи экономических показателей. В регрессионном анализе строится модель регрессии в виде математической функции, которая показывает влияние факторов на некоторый экономический показатель.

Теоретически корреляция и регрессия связаны между собой. Рассмотрим виды зависимости между случайными величинами. Пусть имеется двумерная (многомерная) случайная величина, например $(X, Y)$. Зависимость между случайными величинами может быть следующих видов:

- функциональная: если значению случайной величины $x$ по определенному закону ставится в соответствие значение случайной величины $y$;

- статистическая (вероятностная): если значению случайной величины $x$ ставится в соответствие определенное распределение случайной величины $y$.

Математическое ожидание $\mathsf{E}y$, определенное для каждого значения $X = x$ в вероятностной зависимости, называется условным математическим ожиданием. Статистическая зависимость, в которой при изменении случайной величины $x$ изменяется условное математическое ожидание случайной величины $Y$, называется корреляционной зависимостью.

Если условное математическое ожидание случайной величины $Y$ не изменяется при изменении значений $X$, то корреляционной зависимости нет (т.е., любая корреляционная зависимость является статистической, но не всякая статистическая зависимость является корреляционной.)

Функция, которая описывает закон изменения условного математического ожидания случайной величины $Y$ при изменении другой случайной величины $X$, называется функцией регрессии $Y$ на $X$. Если двумерная случайная величина $(X, Y)$ распределена по нормальному закону, то функция регрессии линейная, корреляционная зависимость тоже линейная. Из вышесказанного понятна практическая интерпретация корреляционного и регрессионного анализа - выводы исследования можно интерпретировать лишь как некоторые усредненные свойства, обобщенные характеристики экономических процессов. Это необходимо иметь в виду тем, кто применяет корреляционный и регрессионный анализ.

Любое конкретное исследование имеет дело с некоторыми выборочными данными из некоторой генеральной совокупности. Задача исследования состоит в том, чтобы на основе анализа выборки сделать выводы о свойствах всей совокупности. Эта задача решается с помощью специальных приемов математической статистики - проверки статистических гипотез. Принимается с определенной вероятностью гипотеза о наличии (отсутствии) данного свойства в генеральной совокупности. Эта гипотеза проверяется с помощью специального показателя, определяемого по выборке, - статистического критерия, который позволяет с заданной вероятностью ошибки сделать вывод о том, значимы ли значения выборочных показателей для выводов о наличии определенных свойств в генеральной совокупности, или их значения случайно отличны от нуля и они не значимы.

Для проверки статистической значимости различных видов выборочных статистических показателей существуют тесты, использующие свои критерии, каждый из которых является случайной величиной, имеющей то или иное известное вероятностное распределение, зависящее от степеней свободы и выбранного уровня значимости (доверительной вероятности).

Далее эти вопросы будут рассмотрены более подробно и на конкретных примерах.

## 1.3. Парный коэффициент корреляции

### 1.3.1. Коэффициент корреляции

Пусть $(X, Y)$ – двумерная **нормально распределенная** случайная величина. Тогда «степень зависимости» случайных величин $X$ и $Y$ характеризуется парным коэффициентом корреляции

$$\rho = \operatorname{corr}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{Var}(X) \cdot \operatorname{Var}(Y)}} = \frac{\mathsf{E}(XY) - \mathsf{E}X \cdot \mathsf{E}Y}{\sqrt{\operatorname{Var}(X) \cdot \operatorname{Var}(Y)}}.$$

Из определения коэффициента корреляции следует, что

1. всегда $-1 \leqslant \rho \leqslant 1$;

2. не меняется при линейных преобразованиях величин, т.е.

$$\operatorname{corr}(X, Y) = \operatorname{corr}(a_0 + a_1 X, b_0 + b_1 Y), \quad a_1, b_1 \neq 0.$$

Коэффициент корреляции принимает крайние значения $\pm 1$ в том и только том случае, когда между случайными величинами $X$ и $Y$ существует **линейная функциональная** зависимость, т.е.

$$\rho = \pm 1 \Leftrightarrow Y = \beta_0 + \beta_1 X, \quad \beta_1 \neq 0,$$

причем

$$\beta_1 = \rho \sqrt{\frac{\mathrm{Var}(Y)}{\mathrm{Var}(X)}},$$

т.е. знак коэффициента $\beta_1$ совпадает по знаком коэффициента корреляции.

В общем случае коэффициент корреляции возникает при решении следующей экстремальной задачи: подобрать линейную функцию $l(x) = \beta_0 + \beta_1 x$ так, чтобы случайная величина $l(X)$ меньше всего отклонялась от $Y$ в среднеквадратичном, т.е.

$$\mathsf{E}\left(Y - \beta_0 - \beta_1 X\right)^2 \underset{\beta_0, \beta_1}{\longrightarrow} \min.$$

Решение этой задачи задается равенствами

$$\beta_1^* = \frac{\mathrm{cov}(X, Y)}{\mathrm{Var}(X)} = \rho \sqrt{\frac{\mathrm{Var}(Y)}{\mathrm{Var}(X)}}, \quad \beta_0^* = \mathsf{E}Y - \beta_1^* \cdot \mathsf{E}X$$

и наименьшее среднеквадратичное отклонение равно

$$\mathsf{E}\left(Y - \beta_0^* - \beta_1^* X\right)^2 = \left(1 - \rho^2\right) \mathrm{Var}(Y).$$

Кроме того, для всех $x \in \mathbb{R}$ верно

$$\mathsf{E}(Y|X = x) = \beta_0^* + \beta_1^* x,$$

т.е. наилучший прогноз случайной величины $Y$, при условии, что известно значение случайной величины $X = x$, равен $\widehat{Y} = \beta_0^* + \beta_1^* x$. Рассмотрим три случая:

1. $\rho > 0$. Тогда $\beta_1^* > 0$ и при увеличении $x$ ожидаемое (среднее) значение $\mathsf{E}(Y|X = x)$ случайной величины $Y$ также увеличивается. В этом случае говорят о **прямой линейной зависимости** между величинами.

2. $\rho < 0$. Тогда $\beta_1^* < 0$ и при увеличении $x$ ожидаемое (среднее) значение $\mathsf{E}(Y|X = x)$ случайной величины $y$ уменьшается. В этом случае говорят об **обратной линейной зависимости** между величинами.

3. $\rho = 0$. Тогда $\beta_1^* = 0$, $\mathsf{E}(Y|X = x) = \beta_0^*$ и знание значения случайной величины $X$ не улучшает прогноз $Y$.

Важное значение коэффициента корреляции обусловлено следующей теоремой

**Теорема.** *Пусть $(X, Y)$ – двумерная нормально распределенная случайная величина. Тогда случайные величины $X$ и $Y$ независимы тогда и только тогда, когда* $\mathrm{corr}(X, Y) = 0$.

Таким образом, парный коэффициент корреляции можно рассматривать как *меру зависимости* двух случайных величин (факторов), имеющих совместное нормальное распределение, причем:

- $\rho = 0 \Leftrightarrow$ величины независимы;
- $\rho = \pm 1 \Leftrightarrow$ между величинами линейная функциональная зависимость: $y = \beta_0^* + \beta_1^* x$.

### 1.3.2. Выборочный коэффициент корреляции

Пусть $(x_i, y_i)_{i=1}^n$ – выборка из двумерной нормально распределенной случайной величины, $n$ – объем выборки.

Напомним, что выборочные (неисправленные) оценки дисперсий случайных величин $X$ и $Y$ определяются как

$$\widehat{\mathrm{Var}}(X) = \widehat{\sigma}_x^2 = \sum_{i=1}^n (x_i - \bar{x})^2 = \overline{(x^2)} - (\bar{x})^2$$
$$\widehat{\mathrm{Var}}(Y) = \widehat{\sigma}_y^2 = \sum_{i=1}^n (y_i - \bar{y})^2 = \overline{(y^2)} - (\bar{y})^2,$$

где

$$\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i, \quad \overline{x^2} = \frac{1}{n}\sum_{i=1}^n x_i^2.$$

Напомним также, что $\widehat{\mathrm{Var}}(X)$ и $\widehat{\mathrm{Var}}(Y)$ – состоятельные, но смещенные оценки дисперсий $\mathrm{Var}(X)$ и $\mathrm{Var}(Y)$ соответственно.

*Выборочный коэффициент ковариации* определяется как[1]

$$\widehat{\mathrm{cov}}(X, Y) = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \overline{xy} - \bar{x} \cdot \bar{y},$$

а *выборочный коэффициент корреляции* определяется равенством[2]

$$r = \widehat{\mathrm{corr}}(X, Y) = \frac{\widehat{\mathrm{cov}}(X, Y)}{\sqrt{\widehat{\mathrm{Var}}(X) \cdot \widehat{\mathrm{Var}}(Y)}}, \quad -1 \leqslant r \leqslant 1$$

Выборочные коэффициенты ковариации и корреляции являются состоятельными оценками коэффициентов ковариации и корреляции в генеральной совокупности. Выборочный коэффициент корреляции может рассматриваться как выборочная «мера линейной зависимости» между случайными величинами.

[1] В MS Excel функция КОВАР$(\cdot, \cdot)$
[2] В MS Excel функция КОРРЕЛ$(\cdot, \cdot)$

**Проверка значимости коэффициента корреляции**

Проверка значимости подразумевает проверку статистической гипотезы

$$H_0 \ : \ \rho = 0$$

против двусторонней альтернативы

$$H_0 \ : \ \rho \neq 0.$$

Другими словами, проверяется статистическая гипотеза, что в генеральной совокупности случайные величины (факторы) $X$ и $Y$ **не коррелируют**. Так как двумерная случайная величина $(X, Y)$ по предположению имеет совместное нормальное распределение, то некоррелируемость означает независимость факторов. Проверка гипотезы о независимости факторов основана на следующем результате: при справедливости нулевой гипотезы $t$-статистика

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \underset{H_0}{\sim} t_{n-2}$$

имеет распределение Стьюдента с $(n-2)$ степенями свободы. Следовательно, получаем следующий статистический критерий проверки нулевой гипотезы:

*при заданном уровне значимости $\alpha$ гипотеза $H_0$ отвергается в пользу альтернативы $H_1$ при $|t| > t_{\text{кр}}$,*

где $t_{\text{кр}} = t(\alpha; n-2)$ есть **двустороннее** критическое значение распределения Стьюдента $t_{n-2}$. Напомним, что двустороннее критическое значение определяется как решение уравнения

$$\mathsf{P}(|t_{n-2}| > t_{\text{кр}}) = \alpha.$$

При $|t| < t_{\text{кр}}$ говорят, что данные *согласуются* или *не противоречат* нулевой гипотезой, $H_0$ *не отвергается.*

*Пример.* Был рассчитан выборочный коэффициент корреляции $r = 0.68$ между дневными логарифмическими доходностями[3] биржевых индексов NASDAQ и FTSE на основе $n = 62$ выборочных данных. Проверим значимость коэффициента корреляции, т.е. проверим статистическую гипотезу $H_0$ о **независимости** доходностей обоих биржевых индексов (в предположении их **нормальной распределенности!**). Вычислим значение $t$-статистики:

$$t = \frac{0.68 \cdot \sqrt{62-2}}{\sqrt{1-0.68^2}} \approx 7.1838.$$

Критическое значение распределения Стьюдента при уровне значимости $\alpha = 5\%$ равно: $t_{\text{кр}} = t(5\%; 62-2) \approx 2.003$. Так как $|t| > t_{\text{кр}}$, то гипотеза $H_0$ о независимости доходностей **отвергается**, коэффициент корреляции значим.

[3]Логарифмическая доходность рассчитывается как $h_t = \ln(S_t/S_{t-1})$.

*Пример.* Пусть в результате наблюдений над некоторой двумерной случайной величиной $(X, Y)$ получены методом случайной выборки некоторые совокупности пар значений

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $x_i$ | 1,5 | 3,5 | 0,5 | 1,5 | 2 | 1,5 | 4,5 | 3 | 3,5 | 3 |
| $y_i$ | 6,5 | 6 | 6,5 | 4,5 | 8 | 5,5 | 6,5 | 6 | 8 | 4,5 |

Вычислим парный коэффициент корреляции. Имеем

$$\bar{x} = 2.45; \quad (\bar{x})^2 = 6.0025; \quad \overline{x^2} = 7,375$$
$$\bar{y} = 6.2; \quad (\bar{y})^2 = 38.44; \quad \overline{y^2} = 39.75$$
$$\overline{xy} = 15.375; \quad r \approx 0.161.$$

Проверим значимость выборочного коэффициента корреляции. Находим значение $t$-статистики

$$t = \frac{0.161 \cdot \sqrt{10-2}}{\sqrt{1-0.161^2}} \approx 0.46.$$

При уровне значимости $\alpha = 5\%$ критическое значение распределения Стьюдента равно $t_{\text{кр}} = t(5\%; 10-2) \approx 2.31$. Так как $|t| < t_{\text{кр}}$, то коэффициент незначим, данные согласуются с нулевой гипотезой о независимости факторов при уровне значимости 5%.

*Пример.* Пусть в результате наблюдений над некоторой двумерной случайной величиной $(X, Y)$ получены методом случайной выборки некоторые совокупности пар значений

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $x_i$ | $1,5$ | $3,5$ | $0,5$ | $1,5$ | 2 | $1,5$ | $4,5$ | 3 | $3,5$ | 3 |
| $y_i$ | $6,5$ | 6 | $6,5$ | $4,5$ | 8 | $5,5$ | $6,5$ | 6 | 8 | $4,5$ |

Проверить значимость выборочного коэффициента корреляции.

Вычислим парный коэффициент корреляции. Имеем

$$\bar{x} = 2.45; \quad (\bar{x})^2 = 6.0025; \quad \overline{x^2} = 7,375$$
$$\bar{y} = 6.2; \quad (\bar{y})^2 = 38.44; \quad \overline{y^2} = 39.75$$
$$\overline{xy} = 15.375;$$

$$r = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sqrt{\overline{x^2} - (\bar{x})^2}\sqrt{\overline{y^2} - (\bar{y})^2}} \approx 0.161.$$

Проверим значимость выборочного коэффициента корреляции

$$t = \frac{0.161 \cdot \sqrt{10-2}}{\sqrt{1-0.161^2}} \approx 0.46.$$

При уровне значимости $\alpha = 0.05$ критическое значение распределения Стьюдента равно $t_{\text{кр}} = t(0.05; 10-2) \approx 2.31$. Так как $|t| < t_{\text{кр}}$, то коэффициент незначим, данные согласуются с нулевой гипотезой о независимости факторов при уровне значимости 5

## 1.4. Линейная аппроксимация. Метод наименьших квадратов

Пусть на координатной плоскости заданы $n$ точек с координатами $(x_i, y_i)_{i=1}^n$. Требуется найти прямую, «меньше всего отклоняющуюся от заданных точек». Так как прямая задается уравнением

$$y = f(x) = \beta_0 + \beta_1 x,$$

зависящим от двух параметров $\beta_0$ и $\beta_1$, то необходимо по заданным значениям $\{x_i\}$ и $\{y_i\}$ найти значения этих параметров «оптимальной» прямой. Основной вопрос: что понимать под «наименьшим отклонением прямой от точек» и, более общо, как определить «меру отклонения прямой от точек»?

С вероятностной точки зрения, в случае нормального распределения выборочных данных «наилучшими вероятностными и статистическими свойствами» обладают оценки параметров прямой, полученным минимизацией суммы квадратов отклонений (второй случай). Этот метод получения оценок параметров оптимальной прямой называется *Методом Наименьших Квадратов* (сокращенно МНК) или *Ordinary Least Squares* (сокращенно OLS), а полученные оценки параметров называются МНК- или OLS-оценками.

Итак, в качестве меры отклонений прямой от заданных на плоскости точек $(x_i, y_i)_1^n$ возьмем сумму квадратов отклонений в каждой точке[4]:

$$S = S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2.$$

Тогда параметры прямой, для которой эта мера отклонения минимальна, находятся как решение экстремальной задачи без ограничений:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2 \longrightarrow \min.$$

Согласно необходимым условиям существования экстремума, параметры оптимальной прямой находятся как решение системы уравнений

$$\begin{cases} \frac{\partial S}{\partial \beta_0} = \sum\limits_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i) \cdot (-1) = 0 \\ \frac{\partial S}{\partial \beta_1} = \sum\limits_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i) \cdot (-x_i) = 0 \end{cases}$$

[4]Очевидно, $S(\beta_0, \beta_1)$ есть многочлен второго порядка от параметров $\beta_0$ и $\beta_1$.

После простых преобразований приходим к системе линейных уравнений

$$\begin{cases} n\beta_0 + \beta_1 \sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} y_i \\ \beta_0 \sum\limits_{i=1}^{n} x_i + \beta_1 \sum\limits_{i=1}^{n} x_i^2 = \sum\limits_{i=1}^{n} x_i y_i \end{cases} \tag{1.1}$$

называемой системой *нормальных уравнений*. Найдем явные формулы для решения этой системы. Для удобства разделим каждое из уравнение в системе (1.1) на $n$:

$$\begin{cases} \beta_0 + \beta_1 \bar{x} = \bar{y} \\ \beta_0 \bar{x} + \beta_1 \overline{x^2} = \overline{xy} \end{cases}$$

Выразим $\beta_0$ из первого уравнения

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

и подставим во второе уравнение. После преобразования получаем

$$\hat{\beta}_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\left(\overline{x^2}\right) - (\bar{x})^2} = \frac{\widehat{\mathrm{cov}}(x, y)}{\widehat{\mathrm{Var}}(x)} = \widehat{\mathrm{corr}}(x, y) \sqrt{\frac{\widehat{\mathrm{Var}}(y)}{\widehat{\mathrm{Var}}(x)}} = \widehat{\mathrm{corr}}(x, y) \frac{\widehat{\sigma}_y}{\widehat{\sigma}_x}$$

и

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

где $\widehat{\sigma}_x = \sqrt{\widehat{\mathrm{Var}}(x)}$ и $\widehat{\sigma}_y = \sqrt{\widehat{\mathrm{Var}}(y)}$ – выборочные стандартные отклонения $x$ и $y$ соответственно.

Несложно показать, что функция $S(\beta_0, \beta_1)$ выпукла. Следовательно, решение системы нормальных уравнений (1.1) будет глобальным минимумом функции $S(\beta_0, \beta_1)$. Таким образом, оптимальная прямая задается уравнением

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

*Замечание.* Из первого уравнения системы (1.1) следует, что

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x},$$

т.е. оптимальная прямая проходит через точку с координатами $(\bar{x}, \bar{y})$.

Отметим, что метод наименьших квадратов может быть применен для нахождения параметров любой функции, меньше всего отклоняющейся от заданных точек. Эта задача корректно разрешима в случае, когда неизвестные **параметры входят в функцию линейно**. В этом случае система нормальных уравнений будет **системой линейных уравнений** и в общем случае будет иметь единственное решение.

*Пример.* Методом наименьших квадратов построить линейную аппроксимацию функции $y(x)$:

| $x$ | $-1$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|
| $y$ | 1 | 0 | 2 | 1 | 4 |

Считаем суммы

$$\sum_{i=1}^{n} x_i = -1 + 0 + 1 + 2 + 3 = 5, \quad \sum_{i=1}^{n} x_i^2 = (-1)^2 + 0^2 + 1^2 + 2^2 + 3^2 = 15,$$

$$\sum_{i=1}^{n} y_i = 1 + 0 + 2 + 1 + 4 = 8, \quad \sum_{i=1}^{n} x_i y_i = (-1) \cdot 1 + 0 \cdot 0 + 1 \cdot 2 + 2 \cdot 1 + 3 \cdot 4 = 15.$$

Откуда находим средние

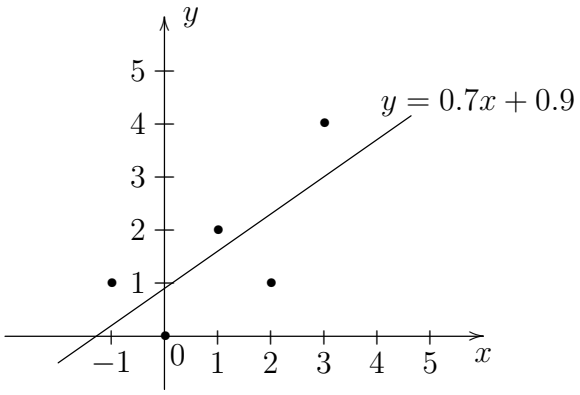
$$\bar{x} = \frac{5}{5} = 1, \quad \bar{y} = \frac{8}{5} = 1.6,$$

$$\bar{xy} = \frac{15}{5} = 3, \quad \bar{x^2} = \frac{15}{5} = 3,$$

Вычисляем коэффициенты оптимальной прямой $\beta_1$ и $\beta_0$:

$$\beta_1 = \frac{\bar{xy} - \bar{x}\bar{y}}{\bar{x^2} - \bar{x}^2} = \frac{3 - 1.6}{3 - 1} = \frac{1.4}{2} = 0.7, \quad \beta_0 = \bar{y} - \beta_1 \bar{x} = 1.6 - 0.7 \cdot 1 = 0.9.$$

Ответ: Искомая прямая $f(x) = 0.7x + 0.9$.




## 1.5. Парная линейная модель регрессии

Перейдем теперь к задаче количественного описания зависимости между двумя экономическими факторами $y$ и $x$, например $y$ – уровень зарплаты индивидуума, а $x$ – уровень образования (в годах). Естественно ожидать, что значение фактора $y$ не всегда однозначно определяется значением фактора $x$. Так, уровень зарплаты зависит не только от уровня образования, но и от множества других факторов (стажа работы, возраста, индивидуальных способностей, места работы и проч.). Кроме того, учесть **все** факторы, влияющие на $y$ помимо $x$ просто не представляется возможным в силу недостаточного количества информации или невозможности ее получения (например, как оценить или измерить индивидуальные способности индивидуума, несомненно влияющие на уровень зарплаты?).

Также для одного значения фактора $x$ могут наблюдаться различные значения фактора $y$.

Обычно для описания ситуаций с недостаточной информацией используют различные вероятностные математические модели. Рассмотрим подробно модель зависимости между факторами, описываемую уравнением

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n \tag{1.2}$$

где $y_i$ и $\varepsilon_i$ суть случайные величины, а $x_i$ – **неслучайная** (детерминированная) величина, $i$ – номер наблюдения. Фактор $y$ называется *зависимой переменной* (dependent variable), а фактор $x$ называется *регрессором* или *объясняющей переменной* (explanatory variable). Параметр $\beta_1$ называется параметром *наклона прямой* (slope), а $\beta_0$ – *константой*, *свободным членом* или *параметром сдвига* (intercept).

Уравнение (1.2) называется *уравнение регрессии* или *регрессионным уравнением*, а случайные величины $\varepsilon_i$ называются *ошибками* регрессии. Ошибки регрессии удобно представлять себе как «неучтенные факторы», влияющие на $y$ помимо фактора $x$. Таким образом, уравнение (1.2) отражает наши представления о характере зависимости между факторами.

Относительно ошибок регрессии будем предполагать выполнения следующих условий, называемых иногда условиями Гаусса – Маркова:

1. $\mathsf{E}\varepsilon_i = 0$, $i = 1, \ldots, n$ (ошибки регрессии несистематические);
2. $\mathrm{Var}(\varepsilon_i) = \sigma^2$ не зависит от $i$;
3. $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$ при $i \neq j$ (некоррелируемость ошибок для разных наблюдений);
4. $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, $i = 1, \ldots, n$ (нормальная распределенность ошибок регрессии).

Из условия $\mathsf{E}\varepsilon_i = 0$ следует, что

$$\mathsf{E}y_i = \beta_0 + \beta_1 x_i,$$

т.е. среднее значение фактора $y$ при заданном значении $x_i$ равно $\beta_0 + \beta_1 x_i$ и не зависит от ошибок регрессии. Отсюда термин: несистематические ошибки.

Очевидно $\mathrm{Var}(y_i) = \mathrm{Var}(\varepsilon_i)$ (т.к. $x_i$ детерминированны). Следовательно, условие постоянства дисперсий ошибок регрессии влечет за собой постоянство дисперсий случайных величин $y_i$. Следует напомнить, что дисперсию $\mathrm{Var}(y_i)$ можно рассматривать как «меру разброса» значений случайной $y_i$ величины относительно своего среднего значения (математического ожидания) $\mathsf{E}y_i = \beta_0 + \beta_1 x_i$. Если смотреть на ошибки регрессии как на «неучтенные факторы», условие постоянства дисперсий можно описательно трактовать следующим образом: «степень влияния» невключенных в модель факторы в разных наблюдениях постоянна. Условие постоянства дисперсий ошибок называется *гомоскедастичностью* (homoskedasticity), и говорят, что ошибки модели регрессии гомоскедастичны, или

однородны. При нарушении условия постоянства дисперсий ошибок регрессии говорят, что ошибки *гетероскедастичны*, или *неоднородны*.

Условие некоррелируемости (независимости в случае нормального распределения) ошибок для разных наблюдений можно трактовать как «локальность» их влияния: невключенные в модель факторы, которые моделируются ошибками регрессии, влияют только на «свое» наблюдение и не влияют на другие. В случае пространственных выборок (cross-sectional data) это условие обычно считается выполненным. Оно, как правило, нарушается в случае построения регрессионных моделей для временных рядов.

### 1.5.1. Теорема Гаусса – Маркова

Итак, мы предполагаем, что зависимость между факторами $y$ и $x$ описывается уравнением регрессии (1.2), но параметры уравнения $\beta_0$, $\beta_1$ и $\sigma^2$ нам неизвестны.

Основная задача – получить «наилучшие» оценки параметров регрессии на основе выборочных данных. Ограничимся рассмотрением только оценок параметров, линейных относительно $y_i$. Под «наилучшими» будем подразумевать несмещенные оценки с минимальной дисперсией[5]. Такие оценки называются BLUE–*оценками* (BLUE = Best Linear Unbiased Estimators), или *эффективными оценками*.

Основным результатом является следующая теорема:

**Теорема** (Гаусс – Марков)**.** *Пусть для линейной модели парной регрессии*

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n$$

*выполнены условия 1. – 3. на ошибки регрессии $\varepsilon_i$. Тогда OLS-оценки $\hat{\beta}_0$ и $\hat{\beta}_1$ параметров $\beta_0$ и $\beta_1$ являются BLUE–оценками, т.е. среди несмещенных линейных (относительно $y_i$) оценок имеют наименьшую дисперсию.*

*Доказательство.* Докажем несмещенность OLS-оценок. Рассмотрим оценку параметра $\beta_1$. Для нее имеем следующее выражение

$$\hat{\beta}_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n\left(\overline{x^2} - (\bar{x})^2\right)} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Так как величины $x_i$ неслучайны и $\mathsf{E} y_i = \beta_0 + \beta_1 x_i$ (условие 1. на ошибки регрессии), то

$$\mathsf{E}\hat{\beta}_1 = \frac{\sum (x_i - \bar{x}) \mathsf{E} y_i}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{\sum (x_i - \bar{x})^2}$$
$$= \beta_0 \frac{\sum (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} + \beta_1 \frac{\sum (x_i - \bar{x}) x_i}{\sum (x_i - \bar{x}) x_i} = \beta_1.$$

[5]Напомним, что оценка параметров вероятностной модели в математической статистике рассматривается как случайная величина.

При выводе мы воспользовались равенствами

$$\sum(x_i - \bar{x}) = 0, \quad \sum (x_i - \bar{x})^2 = \sum(x_i - \bar{x})x_i.$$

Далее, так как

$$\mathsf{E}\,(\bar{y}) = \mathsf{E}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}y_i = \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i) = \beta_0 + \beta_1\bar{x},$$

то для оценки константы $\hat{\beta}_0$ в уравнении регрессии получаем

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \Longrightarrow \mathsf{E}\left(\hat{\beta}_0\right) = \mathsf{E}\,(\bar{y}) - \mathsf{E}\left(\hat{\beta}_1\bar{x}\right) =$$
$$\beta_0 + \beta_1\bar{x} - \bar{x}\cdot\mathsf{E}\left(\hat{\beta}_1\right) = \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} = \beta_0$$

Итак, $\hat{\beta}_0$ и $\hat{\beta}_1$ – несмещенные (unbiased) оценки параметров $\beta_0$ и $\beta_1$ уравнения регрессии.

Вычислим теперь дисперсии оценок $\hat{\beta}_0$ и $\hat{\beta}_1$. Для этого воспользуемся тем фактом, что из условий 2. и 3. на ошибки регрессии следует, что $\mathrm{Var}(y_i) = \sigma^2$ и $\mathrm{cov}(y_i, y_j) = 0$ при $i \neq j$. Следовательно, используя свойства дисперсии, для оценки $\hat{\beta}_1$ получаем:

$$\mathrm{Var}\left(\hat{\beta}_1\right) = \mathrm{Var}\left(\frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2}\right) =$$
$$\frac{\mathrm{Var}\left(\sum_{i=1}^{n}(x_i - \bar{x})y_i\right)}{\left(\sum(x_i - \bar{x})^2\right)^2} = \frac{\sum(x_i - \bar{x})^2\,\mathrm{Var}(y_i)}{\left(\sum(x_i - \bar{x})^2\right)^2} =$$
$$\frac{\sum(x_i - \bar{x})^2\sigma^2}{\left(\sum(x_i - \bar{x})^2\right)^2} = \sigma^2\frac{\sum(x_i - \bar{x})^2}{\left(\sum(x_i - \bar{x})^2\right)^2} = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}.$$

Для нахождения дисперсии оценки $\hat{\beta}_0$ сначала перепишем ее в виде

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = \sum_{i=1}^{n}\frac{1}{n}y_i - \bar{x}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sum(x_i - \bar{x})^2}\right)y_i$$
$$= \sum_{i=1}^{n}\left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}\right)y_i.$$

Следовательно,

$$\operatorname{Var}\left(\hat\beta_0\right) = \operatorname{Var}\left(\sum_{i=1}^n \left(\frac{1}{n} - \frac{\bar x(x_i-\bar x)}{\sum(x_i-\bar x)^2}\right) y_i\right) =$$

$$\sum_{i=1}^n \left(\frac{1}{n} - \frac{\bar x(x_i-\bar x)}{\sum(x_i-\bar x)^2}\right)^2 \operatorname{Var}(y_i) =$$

$$\sigma^2 \sum_{i=1}^n \left(\frac{1}{n^2} - 2\frac{\bar x(x_i-\bar x)}{n\sum(x_i-\bar x)^2} + \frac{(\bar x)^2\,(x_i-\bar x)^2}{\left(\sum(x_i-\bar x)^2\right)^2}\right) =$$

$$\sigma^2 \left(\sum_{i=1}^n \frac{1}{n^2} - \frac{2\bar x\sum_{i=1}^n(x_i-\bar x)}{n\sum(x_i-\bar x)^2} + \frac{(\bar x)^2\sum_{i=1}^n(x_i-\bar x)^2}{\left(\sum(x_i-\bar x)^2\right)^2}\right) =$$

$$\sigma^2\left(\frac{1}{n} + \frac{(\bar x)^2}{\sum(x_i-\bar x)^2}\right) = \frac{\sigma^2\sum x_i^2}{n\sum(x_i-\bar x)^2} = \frac{\sigma^2\cdot\overline{x^2}}{\sum(x_i-\bar x)^2}.$$

Покажем теперь, что любая другая линейна несмещенная оценка имеет бóльшую дисперсию. Пусть $\widetilde\beta_1 = \sum c_i y_i$ – произвольная линейная (по $y_i$) несмещенная оценка параметра наклона $\beta_1$. Представим ее коэффициенты $c_i$ как $c_i = \omega_i + \theta_i$, где $\hat\beta_1 = \sum \omega_i y_i$ ($\omega_i = (x_i-\bar x)/\sum(x_i-\bar x)^2$). Так как $\mathsf{E}\widetilde\beta_1 = \mathsf{E}\hat\beta_1 = \beta_1$, то

$$0 = \mathsf{E}\widetilde\beta_1 - \mathsf{E}\hat\beta_1 = \mathsf{E}\left(\widetilde\beta_1 - \hat\beta_1\right)$$

$$= \mathsf{E}\left(\sum\theta_i y_i\right) = \sum\theta_i\mathsf{E}y_i$$

$$= \sum\theta_i(\beta_0+\beta_1 x_i) = \beta_0\sum\theta_i + \beta_1\sum\theta_i x_i.$$

Так как это равенство должно быть выполнено для произвольных значений $\beta_0$ и $\beta_1$, то получаем, что

$$\sum\theta_i = 0 \quad \sum\theta_i x_i = 0.$$

Далее,

$$\operatorname{Var}\left(\widetilde\beta_1\right) = \operatorname{Var}\left(\sum c_i y_i\right) = \sum c_i^2\operatorname{Var}(y_i)$$

$$= \sigma^2\sum_i(\omega_i+\theta_i)^2 = \sigma^2\left(\sum\omega_i^2 + 2\sum\omega_i\theta_i + \sum\theta_i^2\right).$$

По условию $\omega_i = (x_i-\bar x)/(\sum(x_i-\bar x)^2)$, откуда

$$\sum\omega_i\theta_i = \sum\frac{(x_i\theta_i - \bar x\theta_i)}{\sum(x_i-\bar x)^2} = \frac{\sum x_i\theta_i - \bar x\sum\theta_i}{\sum(x_i-\bar x)^2} = 0.$$

Так как $\operatorname{Var}\left(\hat\beta_1\right) = \sigma^2\sum\omega_i^2$, то окончательно получаем

$$\operatorname{Var}\left(\widetilde\beta_1\right) = \sigma^2\left(\sum\omega^2 + \sum\theta_i^2\right) = \sigma^2\sum\omega_i^2 + \sigma^2\sum\theta_i^2$$

$$= \operatorname{Var}\left(\hat\beta_1\right) + \sigma^2\sum\theta_i^2 \geqslant \operatorname{Var}\left(\hat\beta_1\right).$$

Таким образом, $\mathrm{Var}\left(\widetilde{\beta}_1\right) \geqslant \mathrm{Var}\left(\hat{\beta}_1\right)$.

Аналогично можно показать, что для произвольной несмещенной оценки $\widetilde{\beta}_0$ параметра $\beta_0$ всегда $\mathrm{Var}\left(\widetilde{\beta}_0\right) \geqslant \mathrm{Var}\left(\hat{\beta}_0\right)$. Теорема доказана. $\square$

*Замечание.* Из доказательства видно, что для несмещенности OLS-оценок достаточно **только** условия 1. на ошибки регрессии.

*Замечание.* Из теоремы Гаусса – Маркова следует, что среди линейных по $y$ несмещенных оценок параметров $\beta_0$ и $\beta_1$ наилучшими (т.е. с минимальной дисперсией) будут OLS-оценки. Однако могут существовать и нелинейные оценки параметров $\beta_0$ и $\beta_1$ с дисперсией меньшей, чем у OLS-оценок.

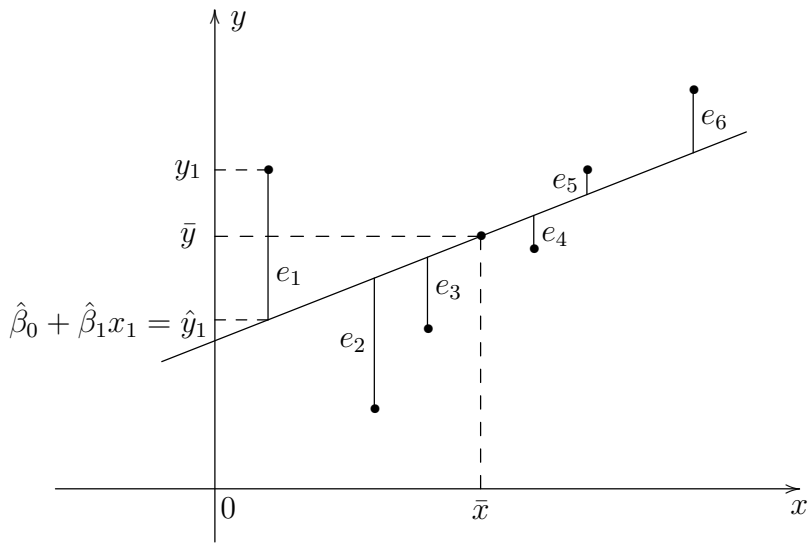
Найдем теперь оценку третьего параметра уравнения регрессии – дисперсии ошибок $\sigma^2$. Обозначим через

$$\widehat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

прогноз фактора $y$ при заданном значении $x_i$. Значения $\widehat{y}_i$ также называются *подогнанными* (fitted value) или *предсказанными* значениями зависимой переменной.

**Определение.** *Остатки* (residual) модели регрессии определяются равенством $e_i = y_i - \widehat{y}_i$.

Важно различать ошибки $\varepsilon_i$ и остатки $e_i$. Остатки также являются случайными величинами, но в отличие от ошибок (имеющих теоретический характер), они наблюдаемы.




Кроме того, для остатков всегда выполнено соотношение $\sum\limits_{i=1}^{n} e_i = 0$, следующее из первого уравнения системы (1.1), т.е. остатки **всегда зависимы**, в отличие от ошибок регрессии $\varepsilon_i$. Но, тем не менее, можно считать, что остатки в некотором

смысле «моделируют» ошибки регрессии и «наследуют» их свойства. На этом основаны методы исследования отклонений выборочных данных от предположений теоремы Гаусса – Маркова.

Введем следующее обозначение:

$$\mathrm{RSS} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \widehat{y}_i\right)^2 .$$

Величина RSS называется *остаточной суммой квадратов* (residual sum of squares) в модели регрессии. Можно показать, что

$$\mathsf{E}(\mathrm{RSS}) = (n-2)\sigma^2.$$

Следовательно, статистика

$$s^2 = \frac{\mathrm{RSS}}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$$

является несмещенной оценкой дисперсии ошибок регрессии. Выборочная *стандартная ошибка* регрессии SER (Standard Error of Regression) определяется как

$$\mathrm{SER} = s = \sqrt{s^2} = \sqrt{\frac{\mathrm{RSS}}{n-2}}.$$

### 1.5.2. Статистические свойства OLS-оценок коэффициентов

При доказательстве теоремы Гаусса – Маркова мы нашли дисперсии оценок параметров регрессии $\hat{\beta}_0$ и $\hat{\beta}_1$. В выражениях для дисперсий участвует дисперсия ошибок $\sigma^2$, значение которой в большинстве прикладных задач неизвестно. Поэтому в прикладных вычислениях используют **оценки** дисперсий величин $\hat{\beta}_0$ и $\hat{\beta}_1$:

$$\widehat{\mathrm{Var}}\left(\hat{\beta}_0\right) = \frac{s^2 \cdot \overline{x^2}}{\sum \left(x_i - \bar{x}\right)^2},$$
$$\widehat{\mathrm{Var}}\left(\hat{\beta}_1\right) = \frac{s^2}{\sum \left(x_i - \bar{x}\right)^2},$$
$$\widehat{\mathrm{cov}}\left(\hat{\beta}_0, \hat{\beta}_1\right) = -\frac{s^2 \cdot \bar{x}}{\sum \left(x_i - \bar{x}\right)^2}.$$

получаемые формальной заменой неизвестного параметра $\sigma^2$ в выражении для дисперсии и ковариации оценок коэффициентов на его несмещенную оценку $s^2$. Стандартные ошибки оценок коэффициентов регрессии определяются как

$$s_0 = \sqrt{\widehat{\mathrm{Var}}\left(\hat{\beta}_0\right)} = \sqrt{\frac{s^2 \cdot \overline{x^2}}{\sum \left(x_i - \bar{x}\right)^2}}$$
$$s_1 = \sqrt{\widehat{\mathrm{Var}}\left(\hat{\beta}_1\right)} = \sqrt{\frac{s^2}{\sum \left(x_i - \bar{x}\right)^2}}$$

*Замечание.* Часто при записи оцененной модели регрессии выборочные стандартные ошибки коэффициентов указываются в круглых скобках под коэффициентами.

В теореме Гаусса –Маркова была доказана несмещенность и оптимальность OLS-оценок параметров линейной регрессии и были вычислены дисперсии этих оценок. Для этого было достаточно условий 1. – 3. на ошибки модели регрессии.

Для получения статистических свойств оценок (доверительных интервалов и тестирования статистических гипотез) необходимо знать закон распределения оценок $\hat{\beta}_0$ и $\hat{\beta}_1$ параметров регрессии.

Из условия 4. на ошибки регрессии очевидно следует, что

$$y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2) \quad \text{и} \quad \operatorname{cov}(y_i, y_j) = 0,\ i \neq j.$$

Следовательно, случайные величины $\hat{\beta}_0$ и $\hat{\beta}_1$ также *нормально распределены*, так как они линейно выражаются через $y_i$. Таким образом, мы получили следующее

**Предложение.** *Если выполнено условие 4. на ошибки регрессии, то случайные величины $\hat{\beta}_0$ и $\hat{\beta}_1$ имеют нормальное распределение с параметрами*

$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \frac{\sigma^2 \overline{x^2}}{\sum (x_i - \bar{x})^2}\right)$$
$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum (x_i - \bar{x})^2}\right)$$

Однако в прикладных задачах дисперсия ошибок регрессии $\sigma^2$, как правило, неизвестна, и следовательно, для получения статистических свойств оценок параметров регрессии полученного результата недостаточно.

Нам понадобится следующая теорема:

**Теорема.** *Если выполнены условия 1. – 4. на ошибки регрессии, то*

1. *оценка $s^2$ (как случайная величина)* ***не зависит*** *от $\hat{\beta}_0$ и $\hat{\beta}_1$;*
2. *случайная величина $(n-2)s^2/\sigma^2$ имеет распределения $\chi^2_{n-2}$.*

Далее

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \operatorname{Var}(\hat{\beta}_1)\right) \Longrightarrow$$
$$\hat{\beta}_1 - \beta_1 \sim \mathcal{N}\left(0, \operatorname{Var}(\hat{\beta}_1)\right) \Longrightarrow$$
$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\operatorname{Var}(\hat{\beta}_1)}} \sim \mathcal{N}(0,1).$$

Напомним, что выборочная стандартная ошибка оценки $\hat{\beta}_1$ равна

$$s_1 = \sqrt{\widehat{\operatorname{Var}}\left(\hat{\beta}_1\right)} = \sqrt{\frac{s^2}{\sum (x_i - \bar{x})^2}}.$$

Тогда статистика

$$t_1 = \frac{\hat{\beta}_1 - \beta_1}{s_1} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\mathrm{Var}(\hat{\beta}_1)}} \cdot \frac{1}{s/\sigma}$$

имеет распределение Стьюдента $t_{n-2}$, так как

$$\frac{s}{\sigma} \sim \sqrt{\frac{1}{n-2}\chi^2_{n-2}}.$$

Аналогично доказывается, что статистика

$$t_0 = \frac{\hat{\beta}_0 - \beta_0}{s_0} \sim t_{n-2}$$

Итак, нами доказана

**Теорема.** *Если выполнены условия 1. – 4. на ошибки регрессии, то $t$-статистики*

$$t_0 = \frac{\hat{\beta}_0 - \beta_0}{s_0}, \quad t_1 = \frac{\hat{\beta}_1 - \beta_1}{s_1}$$

*имеют распределение Стьюдента $t_{n-2}$.*

### 1.5.3. Доверительные интервалы. Проверка гипотез

**1.** Выведем формулы для доверительных интервалов для коэффициентов регрессии с доверительной вероятностью $\gamma$. Для определенности рассмотрим коэффициент наклона $\beta_1$. Согласно теореме, при выполнении условий 1. – 4. на ошибки регрессии статистика $t_1$ имеет распределение Стьюдента $t_{n-2}$. Пусть $t_{\text{кр}} = t_{\text{кр}}(\alpha; n-2)$ – **двустороннее** критическое значение распределения $t_{n-2}$ при уровне значимости $\alpha = 1 - \gamma$. Тогда

$$\mathsf{P}(|t_1| > t_{\text{кр}}) = \alpha \Longrightarrow \mathsf{P}(|t_1| < t_{\text{кр}}) = 1 - \alpha = \gamma.$$

Далее

$$|t_1| < t_{\text{кр}} \Leftrightarrow -t_{\text{кр}} < t_1 < t_{\text{кр}} \Leftrightarrow -t_{\text{кр}} < \frac{\hat{\beta}_1 - \beta_1}{s_1} < t_{\text{кр}} \Leftrightarrow$$
$$-t_{\text{кр}} \cdot s_1 < \hat{\beta}_1 - \beta_1 < t_{\text{кр}} \cdot s_1 \Leftrightarrow \hat{\beta}_1 - t_{\text{кр}} \cdot s_1 < \beta_1 < \hat{\beta}_1 + t_{\text{кр}} \cdot s_1.$$

Итак, доверительный интервал для коэффициента $\beta_1$ с доверительной вероятностью $\gamma$ равен

$$\mathsf{P}\left(\hat{\beta}_1 - t_{\text{кр}} \cdot s_1 < \beta_1 < \hat{\beta}_1 + t_{\text{кр}} \cdot s_1\right) = \gamma.$$

Аналогично, доверительный интервал для коэффициента $\beta_0$ с доверительной вероятностью $\gamma$ равен

$$\mathsf{P}\left(\hat{\beta}_0 - t_{\text{кр}} \cdot s_0 < \beta_0 < \hat{\beta}_0 + t_{\text{кр}} \cdot s_0\right) = \gamma.$$

**2.** Приведем статистический критерий для тестирования гипотезы

$$H_0 \,:\, \beta_1 = \theta_0$$

($\theta_0$ – заданное значение) против **двусторонней** альтернативы

$$H_1 \,:\, \beta_1 \neq \theta_0.$$

Предположим, что верна нулевая гипотеза. Тогда статистика

$$t = \frac{\hat{\beta}_1 - \theta_0}{s_1}$$

имеет распределение Стьюдента $t_{n-2}$. Пусть $t_{\text{кр}} = t(\alpha; n-2)$ – **двустороннее** критическое значение распределения $t_{n-2}$ при заданном уровне значимости $\alpha$. Если верна гипотеза $H_0$, то вероятность $\mathsf{P}(|t| > t_{\text{кр}}) = \alpha$ мала. Для проверки гипотезы получаем следующий статистический критерий:

- если $|t| > t_{\text{кр}}$, то гипотеза $H_0$ отвергается в пользу альтернативы $H_1$ при заданном уровне значимости (произошло маловероятное, с точки зрения нулевой гипотезы, событие); также говорят, что коэффициент **значимо** отличается от числа $\theta_0$;

- если $|t| < t_{\text{кр}}$, то данные согласуются с нулевой гипотезой при заданном уровне значимости; также говорят, что коэффициент **незначимо** отличается от числа $\theta_0$.

*Замечание.* Не сложно проверить, что $|t| < t_{\text{кр}}$ тогда и только тогда, когда число $\theta_0$ принадлежит доверительному интервалу для коэффициента $\beta_1$ с доверительной вероятностью $1-\alpha$. Таким образом, мы получаем альтернативный способ проверки нулевой гипотеза:

*гипотеза $H_0$ отвергается при заданном уровне значимости $\alpha \iff$ значение $\theta_0$ не принадлежит доверительному интервалу, построенному для доверительной вероятности $1-\alpha$.*

Этот критерий полезен в прикладных задачах, т.к. некоторые эконометрические пакеты вычисляют доверительные интервалы с заданной доверительной вероятностью автоматически.

В случае **проверки значимости** коэффициента регрессии, т.е. проверки нулевой гипотезы

$$H_0 \,:\, \beta_1 = 0$$

при двусторонней альтернативе $t$-статистика вычисляется как

$$t = \frac{\hat{\beta}_1}{s_1}$$

и именно это значение выводится в эконометрических программах. Коэффициент $\beta_1$ значим (нулевая гипотеза отвергается) при $|t| > t_{\text{кр}}$.

**3.** Приведем статистический критерий для тестирования гипотезы

$$H_0 \,:\, \beta_1 = \theta_0$$

($\theta_0$ – заданное значение) против **односторонней** альтернативы

$$H_1 \,:\, \beta_1 > \theta_0.$$

Пусть $t'_{\text{кр}} = t(\alpha; n-2)$ – **одностороннее**[6] критическое значение распределения $t_{n-2}$ при заданном уровне значимости $\alpha$. Если $H_0$ верна, то $t$-статистика

$$t = \frac{\hat{\beta}_1 - \theta_0}{s_1} \underset{H_0}{\sim} t_{n-2}$$

и вероятность события $\mathsf{P}(t > t'_{\text{кр}}) = \alpha$ мала. Для проверки гипотезы против односторонней альтернативы получаем следующий статистический критерий:

- если $t > t'_{\text{кр}}$, то гипотеза $H_0$ отвергается в пользу альтернативы $H_1$ при заданном уровне значимости (произошло маловероятное, с точки зрения нулевой гипотезы, событие);
- если $t < t'_{\text{кр}}$, то данные согласуются с нулевой гипотезой при заданном уровне значимости.

Аналогично проверяется гипотеза $H_0$ против односторонней альтернативы $H_1 : \beta_1 < \theta_0$.

*Пример.* Для изучения влияния образования на величину почасовой оплаты труда на основе опроса 40 человек была оценена модель регрессии (в скобках указаны стандартные ошибки коэффициентов)

$$\ln(\widehat{Wage}) = \underset{(0.02)}{2.2} + \underset{(0.04)}{0.1}\, Edu,$$

где $Edu$ - уровень образования (в годах), $Wage$ - уровень почасовой оплаты труда. Постройте 99% доверительный интервал для $\beta_1$.

*Решение.* Доверительный интервал для коэффициента $\beta_1$ с доверительной вероятностью 99

$$(\hat{\beta}_1 - t_{\text{кр}} \cdot s_1; \hat{\beta}_1 + t_{\text{кр}} \cdot s_1).$$

Так как $\hat{\beta}_1 = 0.1$, $s_1 = 0.04$, а из таблицы Стьюдента $t_{\text{кр}} = t(0.01; 40-2) = 2.704$, то доверительный интервал будет иметь вид:

$$(0.1 - 2.704 \cdot 0.04; 0.1 + 2.704 \cdot 0.04) = (-0.08; 0.208).$$

*Пример.* По данным с 1960 по 2005 гг. была оценена (статическая) кривая Филлипса, связывающая уровень инфляции $inf_t$ и уровень безработицы $unem_t$:

$$\widehat{inf}_t = 2.34 - 0.23 unem_t, \quad s_1 = 0.04 \quad R^2 = 0.12$$

Протестируйте гипотезу $H_0 : \theta_1 = 0$ против альтернативы $H_1 : \theta_1 < 0$ при уровне значимости 1%. Дайте объяснение проверяемой нулевой гипотезе и альтернативе.

[6]Напомним, что одностороннее критическое значение находится из условия $\mathsf{P}(t_{n-2} > t'_{\text{кр}}) = \alpha$.

*Решение.* Гипотеза $H_0 : \theta_1 = 0$ говорит о том, что безработица является незначимым фактором для инфляции. Для проверки гипотезы требуется вычислить значение статистики распределения Стюдента: $t = \frac{\widehat{\theta_1}}{s(\widehat{\theta_1})} = \frac{-0.23}{0.04} = -5.75$. Теперь сравним данное значение с критическим значением: $t(\alpha, n-k-1) = t(0.01, 45-1-1) = t(0.01; 43) = -2.423$. Так как значение статистики $t$ лежит в критической области, то гипотеза о том, что безработица не является значимым фактором, отвергается.

### 1.5.4. Коэффициент $R^2$ и «качество подгонки»

**1.** Рассмотрим величину $\sum(y_i - \bar{y})^2$ разброса (вариации) фактора $y$ относительно своего среднего значения $\bar{y}$. Несложно показать, что общая вариация $y$ может быть представлена в виде

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum e_i^2.$$

Введем следующие обозначения:

- $\mathrm{TSS} = \sum(y_i - \bar{y})^2$ – общая вариация зависимой переменной $y$ (общая сумма квадратов, Total Sum of Squares);
- $\mathrm{ESS} = \sum (\hat{y}_i - \bar{y})^2$ – вариация зависимой переменной, объясненная регрессией (объясненная сумма квадратов, Explained Sum of Squares);
- $\mathrm{RSS} = \sum (y_i - \hat{y}_i)^2 = \sum e_i^2$ – остаточная часть вариации $y$ (остаточная сумма квадратов, Residual Sum of Squares).

Тогда в новых обозначениях получаем разложение общей вариации фактора $y$ на объясненную регрессий и остаточную суммы квадратов

$$\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS}\,.$$

**Определение.** *Коэффициент*[7] $R^2$ определяется как доля объясненной регрессией суммы квадратов в общей сумме квадратов зависимой переменной

$$R^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}.$$

Из определения непосредственно следует, что всегда $0 \leqslant R^2 \leqslant 1$. Кроме того, для крайних значений коэффициента $R^2$ получаем.

- $R^2 = 0 \Leftrightarrow \mathrm{ESS} = 0 \Leftrightarrow \hat{y}_i = \bar{y} \Leftrightarrow \hat{\beta}_1 = 0$, т.е. значения фактора $x$ не улучшают прогноз фактора $y$ по сравнению с тривиальным прогнозом $\hat{y}_i = \bar{y}$.
- $R^2 = 1 \Leftrightarrow \mathrm{RSS} = 0 \Leftrightarrow e_i = 0 \Leftrightarrow y_i = \hat{y}_i$, т.е. получаем «идеальную подгонку» прямой: все выборочные лежат на одной прямой и значение фактора $x$ позволяет точно предсказать значение фактора $y$.

[7] В отечественной литературе коэффициент $R^2$ иногда называется коэффициентом детерминации.

Таким образом, чем ближе значение $R^2$ к 1, тем «лучше» качество подгонки прямой и модели регрессии и коэффициент $R^2$ можно рассматривать как меру «качества подгонки» («goodness-of-fit») однофакторной модели регрессии на выборочных данных.

Можно показать, что для линейной модели регрессии

$$R^2 = \left(\widehat{\operatorname{corr}}(x, y)\right)^2,$$

что показывает связь между выборочным коэффициентом корреляции и качеством подгонки прямой.

**2.** Рассмотрим статистические свойства коэффициента $R^2$. Так как

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right), \quad \frac{(n-2)s^2}{\sigma^2} \sim \chi^2_{n-2},$$

то

$$\frac{(\hat{\beta}_1 - \beta_1)\sqrt{\sum(x_i - \bar{x})^2}}{\sigma} \sim \mathcal{N}(0, 1), \quad \frac{\mathrm{RSS}}{\sigma^2} \sim \chi^2_{n-2}$$

Следовательно, $\mathcal{F}$-статистика

$$\mathcal{F} = \frac{\frac{(\hat{\beta}_1 - \beta_1)^2 \sum(x_i - \bar{x})^2}{\sigma^2}}{\frac{\mathrm{RSS}}{\sigma^2} \cdot \frac{1}{n-2}} = \frac{(\hat{\beta}_1 - \beta_1)^2 \sum(x_i - \bar{x})^2}{\mathrm{RSS}} \cdot \frac{n-2}{1} \sim F_{1,n-2}$$

имеет распределение Фишера со степенями свободы $(1, n-2)$. Далее,

$$\hat{\beta}_1^2 \sum(x_i - \bar{x})^2 = \sum\left(\hat{\beta}_1(x_i - \bar{x})\right)^2 =$$

$$\sum\left(\underbrace{\bar{y} - \hat{\beta}_1\bar{x}}_{\hat{\beta}_0} + \hat{\beta}_1 x_i - \bar{y}\right)^2 = \sum(\hat{y}_i - \bar{y})^2 = \mathrm{ESS}\,.$$

Следовательно, если $\beta_1 = 0$, то $\mathcal{F}$-статистика принимает следующий вид:

$$F = \frac{\hat{\beta}_1^2 \sum(x_i - \bar{x})^2}{\mathrm{RSS}} \cdot \frac{n-2}{1} = \frac{\mathrm{ESS}}{\mathrm{RSS}} \cdot \frac{n-2}{1} = \frac{\mathrm{ESS}\,/\,\mathrm{TSS}}{\mathrm{RSS}\,/\,\mathrm{TSS}} \cdot \frac{n-2}{1} = \frac{R^2}{1 - R^2} \cdot \frac{n-2}{1}.$$

Таким образом, $F$-статистику можно использовать для проверки значимости коэффициента $\beta_1$, т.е. для проверки статистической гипотезы

$$H_0 \,:\, \beta_1 = 0.$$

Так как при справедливости нулевой гипотезы эта статистика имеет распределение Фишера

$$F \underset{H_0}{\sim} F_{1,n-2},$$

то гипотеза $H_0$ отвергается при больших значения $F$-статистики. А именно, при заданном уровне значимости $\alpha$ гипотеза $H_0$ отвергается в пользу альтернативы

$$H_1 \,:\, \beta_1 \neq 0$$

если $F > F_{\text{кр}} = F(\alpha; 1, n-2)$.

*Замечание.* Для $F$-статистики в однофакторной модели регрессии можно показать, что

$$F = t^2 = \frac{\hat{\beta}_1^2}{s_1^2}$$

## 1.6. Прогнозирование в модели парной регрессии

Выше мы рассматривали задачу нахождения наилучших (BLUE) оценок параметров парной модели регрессии на основе выборочных данных. Рассмотрим теперь задачу прогнозирования: как оценить значение зависимой переменной $y$ для некоторого значения объясняющей переменной (регрессора). Будем рассматривать *точечный* и *интервальный* прогноз. Точечный прогноз – случайная величина, являющаяся оценкой зависимой переменной. Интервальный прогноз – это выборочный доверительный интервал для случайной величины $y$.

Более точно, пусть кроме выборочных данных $(x_i, y_i)_{i=1}^n$ задано также еще одно значение объясняющей переменной $x_{n+1}$, и известно, что соответствующее значение зависимой переменной удовлетворяет той же парной модели регрессии

$$y_{n+1} = \beta_0 + \beta_1 x_{n+1} + \varepsilon_{n+1}$$

и ошибка регрессии удовлетворяет условия 1. – 4. Задача состоит в нахождении оценки величины $y_{n+1}$ через известные $(x_i, y_i)_{i=1}^n$ и $x_{n+1}$.

**1.** Вначале рассмотрим простой случай, когда значения параметров регрессии $\beta_0, \beta_1$ и $\sigma^2$ известны точно. Тогда в качестве прогноза $\hat{y}_{n+1}$ случайной величины $y_{n+1}$ естественно взять ее математическое ожидание

$$\hat{y}_{n+1} = \mathsf{E}(y_{n+1}) = \beta_0 + \beta_1 x_{n+1}.$$

Среднеквадратическая ошибка прогноза равна

$$\mathsf{E}\left(y_{n+1} - \hat{y}_{n+1}\right)^2 = \mathsf{E}\varepsilon_{n+1}^2 = \sigma^2.$$

Так как по предположению в этом случае

$$y_{n+1} - \hat{y}_{n+1} = \varepsilon_{n+1} \sim \mathcal{N}(0, \sigma^2),$$

то для доверительного интервала с заданной доверительной вероятность $\gamma$ будем иметь следующее выражение

$$\mathsf{P}\left(\hat{y}_{n+1} - \sigma \cdot z_\gamma < y_{n+1} < \hat{y}_{n+1} + \sigma \cdot z_\gamma\right) = \gamma,$$

где $z_\gamma$ есть **двустороннее** критическое значение стандартного нормального распределения с уровнем значимости $(1 - \gamma)$ и находится из уравнения

$$\Phi(z_\gamma) = \frac{1 + \gamma}{2}.$$

**2.** Однако во многих прикладных задачах точные значения параметров регрессии неизвестны и оцениваются по выборочным данным. Тогда естественно в формуле для прогноза заменить неизвестные значения коэффициентов их OLS-оценками:

$$\hat{y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}.$$

Так как OLS-оценки коэффициентов регрессии являются несмещенными, то

$$\mathsf{E}\hat{y}_{n+1} = \mathsf{E}\hat{\beta}_0 + \mathsf{E}(\hat{\beta}_1 x_{n+1}) = \beta_0 + \beta_1 x_{n+1} = \mathsf{E}y_{n+1}.$$

Кроме того, так как OLS-оценки коэффициентов регрессии линейны по $y_i$, то и прогноз $\hat{y}_{n+1}$ также линеен относительно $y_1, \ldots, y_n$. Покажем теперь, что определенное таким образом прогнозное значение является «наилучшим» в смысле среднеквадратичного отклонения.

**Теорема.** *Пусть $\tilde{y}$ – линейный (по $y_1, \ldots, y_n$) прогноз с условием $\mathsf{E}\tilde{y}_{n+1} = \mathsf{E}y_{n+1} = \beta_0 + \beta_1 x_{n+1}$. Тогда*

$$\mathsf{E}\left(\tilde{y}_{n+1} - y_{n+1}\right)^2 \geqslant \mathsf{E}\left(\hat{y}_{n+1} - y_{n+1}\right)^2.$$

Приведем теперь формулу для доверительного интервала в случае неизвестных параметров модели регрессии. Можно показать, что среднеквадратичная ошибка прогноза $\hat{y}_{n+1}$ равна

$$\mathsf{E}\left(\hat{y}_{n+1} - y_{n+1}\right)^2 = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right).$$

Заменим $\sigma^2$ на ее OLS-оценку $s^2$ и обозначим

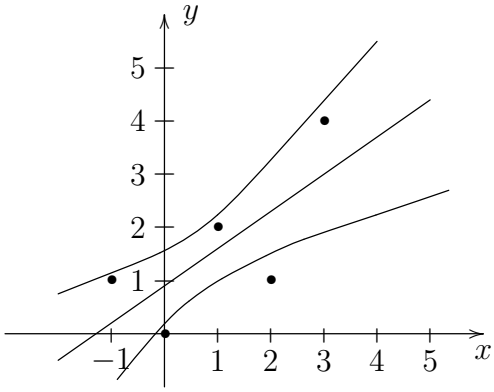
$$\delta = \sqrt{s^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)}.$$

Можно показать, что случайная величина $(\hat{y}_{n+1} - y_{n+1})/\delta$ имеет распределение Стьюдента $t_{n-2}$. Следовательно, доверительный интервал для зависимой переменной с заданной доверительной вероятностью $\gamma$ определяется как

$$\mathsf{P}\left(\hat{y}_{n+1} - \delta \cdot t_\gamma < y_{n+1} < \hat{y}_{n+1} + \delta \cdot t_\gamma\right) = \gamma,$$

где $t_\gamma$ есть **двустороннее** критическое значение распределения Стьюдента $t_{n-2}$ с уровнем значимости $(1 - \gamma)$.

Важно отметить, что значение регрессора $x_{n+1}$ входит как в выражение для точечного прогноза $\hat{y}_{n+1}$, так и в выражение для $\delta$. Т.к. длина доверительного интервала равна $2\delta t_\gamma$, то чем больше значение $(x_{n+1} - \bar{x})^2$, тем больше длина доверительного интервала. Другими словами, чем дальше $x_{n+1}$ от среднего значения $\bar{x}$, тем шире доверительный интервал.




*Пример.* Методом наименьших квадратов построить линейную аппроксимацию функции $y(x)$ и 95%-интервальный прогноз в точке $x_0 = 10$.

| $x$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| $y$ | $-3$ | $-3$ | $-3$ | $-5$ | $-6$ | $-7$ | $-7$ | $-9$ | $-11$ |

*Решение.* Методом наименьших квадратов находим $y = -x - 1$. Для построения интервального прогноза нам потребуется вычислить стандартную ошибку оценки $s = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$. Здесь $\hat{y}_1 = -1 - 1 = -2$, $\hat{y}_2 = -2 - 1 = -3$, $\hat{y}_3 = -3 - 1 = -5$, $\hat{y}_4 = -4 - 1 = -5$, $\hat{y}_5 = -5 - 1 = -6$, $\hat{y}_6 = -6 - 1 = -7$, $\hat{y}_7 = -7 - 1 = -8$, $\hat{y}_8 = -8 - 1 = -9$, $\hat{y}_9 = -9 - 1 = -10$. Подставив значения, получим $s = 0.756$.

Точность $\Delta$ найдем по формуле

$$\Delta = t_\alpha s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}},$$

где $t_\alpha = 2.4$ - это квантиль распределения Стьюдента уровня $1 - \frac{\alpha}{2} = 0.975$ с $n-2 = 7$ степенями свободы, а $x_0 = 10$ - точка, в которой строится прогноз. Вычисляем среднее $\bar{x} = 5$ и $\sum(x_i - \bar{x})^2 = 60$. Откуда $\Delta = 2.4 \cdot 0.756 \cdot \sqrt{1 + \frac{1}{9} + \frac{(10-5)^2}{60}} \approx 2.4 \cdot 0.756 \cdot 1.53 \approx 2.77$.

Итак, точечный прогноз равен $y_0 = -x_0 - 1 = -11$. 95%-доверительный интервал для прогнозного значения равен

$$(-11 - 2.77; -11 + 2.77) = (-13.77; -8.23).$$

Ответ: $(-13.77; -8.23)$.

*Пример.* Методом наименьших квадратов построить линейную аппроксимацию функции $y(x)$ и 90%-интервальный прогноз в точке 5.

| $x$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $y$ | 6 | 5 | 2 | 2 |

*Решение.* Методом наименьших квадратов находим $y = -1.5x + 7.5$.

Для построения интервального прогноза нам потребуется вычислить стандартную ошибку оценки $s = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$. Здесь $\hat{y}_1 = -1.5 + 7.5 = 6$, $\hat{y}_2 = -3 + 7.5 = 4.5$, $\hat{y}_3 = -4.5 + 7.5 = 3$, $\hat{y}_4 = -6 + 7.5 = 1.5$. Подставив значения, получим $s \approx 0.866$.

Точность $\Delta$ найдем по формуле

$$\Delta = t_\alpha s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}},$$

где $t_\alpha = 2.92$ - это квантиль распределения Стьюдента уровня $1 - \frac{\alpha}{2} = 0.95$ с $n-2 = 2$ степенями свободы, а $x_0 = 5$ - точка, в которой строится прогноз. Вычисляем $\bar{x} = 2.5$ и $\sum(x_i - \bar{x})^2 = 5$. Откуда $\Delta = 2.92 \cdot 0.866 \cdot \sqrt{1 + \frac{1}{4} + \frac{(5-2.5)^2}{5}} \approx 3.997$.

Вычисляем точечный прогноз равен $y_0 = -1.5x + 7.5 = 0$. 90%-доверительный интервал для прогнозного значения равен

$$(0 - 3.997; 0 + 3.997) = (-3.997; 3.997).$$

Ответ: $(-3.997; 3.997)$.

## 1.7. Парная регрессия без константы

Рассмотрим задачу о подгонке прямой без свободного члена, т.е. прямой, проходящей через начало координат и задаваемой уравнением

$$y = \beta x.$$

Найдем оценку параметра $\beta$ по методу наименьших квадратов. Оценка находится как решение экстремальной задачи

$$S = \sum_{i=1}^{n} (y_i - \beta x_i)^2 \longrightarrow \min.$$

Критическая точка функции $S$ находится из условия

$$\frac{dS}{d\beta} = 0 \Leftrightarrow \sum_{i=1}^{n} 2(y_i - \beta x_i)(-x_i) = 0 \Leftrightarrow -\sum_{i=1}^{n} y_i x_i + \beta \sum_{i=1}^{n} x_i^2 = 0.$$

Следовательно, OLS-оценка коэффициента $\beta$ находится по формуле

$$\hat{\beta} = \frac{\sum\limits_{i=1}^{n} y_i x_i}{\sum\limits_{i=1}^{n} x_i^2}.$$

Отметим, что в отличие от случая прямой с константой подогнанная прямая $y = \hat{\beta} x$ не обязана проходить через точку $(\bar{x}, \bar{y})$, но она всегда проходит через начало координат.

Так же как и в случае парной регрессии (1.2) определяются полная (TSS), объясненная (ESS) и остаточная (RSS) суммы квадратов. Однако для модели регрессии без свободного члена в общем случае

$$\mathrm{TSS} \neq \mathrm{ESS} + \mathrm{RSS}$$

и коэффициент $R^2$ уже **не имеет смысла**.

В качестве меры «качества подгонки» прямой и модели регрессии без константы можно использовать *нецентрированный* коэффициент $R^2$, определяемый равенством

$$R^2_{\text{нецентр}} = \frac{\sum \widehat{y}_i^2}{\sum y_i^2} = 1 - \frac{\sum e_i^2}{\sum y_i^2}.$$

Ещё одним способом оценить в целом качество модели парной регрессии без свободного члена является вычисление коэффициента корреляции и проверка его значимости.

## 1.8. Нелинейные модели

Выше мы предполагали, что зависимость между факторами $y$ и $x$ описывается линейным уравнением регрессии

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

Из условия 1. на ошибки регрессии следует, что

$$\mathsf{E} y_i = \beta_0 + \beta_1 x_i,$$

т.е. среднее значение[8] $y$ линейно зависит от объясняющей переменной $x$. Согласно этой модели регрессии при изменении значения фактора $x$ на $\triangle_x$ среднее значение фактора $y$ изменяется на $\beta_1 \triangle_x$, и это изменение не зависит от первоначального значения $x$. В частности, средний эффект от увеличения значения фактора $x$ на единицу постоянен и равен $\beta_1$. Другими словами, маржинальное (предельное) значение $y$ по $x$ в линейной модели регрессии постоянно и равно $\beta_1$.

*Пример* (Wage equation [31]). Была оценена линейная модель зависимости уровня почасовой оплаты труда *wage* (в \$) от уровня школьного образования *educ* (в годах обучения):

$$\widehat{wage} = -0.62 + 0.45 educ.$$

Коэффициент наклона прямой 0.45 означает, что каждый дополнительный год обучения (в среднем) увеличивает уровень почасовой оплаты на \$0.45 и это увеличение **не зависит** от количества лет образования и, например, будет одинаково как для первого, так и для предпоследнего года обучения.

Константа $-0.62$ в этой модели формально означает, что человек, не имеющий никакого образования, в среднем получает $-$\$0.62 в час. Естественно ожидать, что в этом случаем константа будет незначима (на «разумном» уровне значимости).

[8] Под средним значением случайной величины будем подразумевать ее математическое ожидание.

Во многих экономических ситуациях наблюдается эффект убывания (или более общó непостоянства) маржинальных (предельных) значений при увеличении значения объясняющей переменной. Для моделирования таких ситуаций удобно использовать степенные и показательные функции

$$y = a_0 \cdot x^{a_1}; \quad y = a_0 \cdot a_1^x.$$

Прологарифмировав обе функции, получаем уравнения, линейные относительно параметров

$$\ln y = \beta_0 + \beta_1 \ln x; \quad \ln y = \beta_0 + \beta_1 x.$$

Со степенной и показательной функциями связаны две модели регрессии: полулогарифмическая и лог-линейная.

**Полулогафмическая модель регрессии** имеет вид

$$\ln y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad n = 1, \ldots, n.$$

Относительно ошибок регрессии предполагается, что они удовлетворяют стандартным требованиям парной линейной регрессии. Тогда наилучшими линейным оценками (BLUE-оценками) параметров регрессии являются OLS-оценки, и к ним применимы все выводы парной линейной модели регрессии (статистические свойства коэффициентов и проч.)

Далее, т.к. $\mathsf{E}\varepsilon_i = 0$, то

$$\mathsf{E}(\ln y_i) = \beta_0 + \beta_1 x_i.$$

Пусть значение фактора $x$ изменяется на $\triangle_x$:

$$\mathsf{E}(\ln y') = \beta_0 + \beta_1 x, \quad \mathsf{E}(\ln y'') = \beta_0 + \beta_1 (x + \triangle_x).$$

Тогда ожидаемое (т.е. среднее) изменение величины $\ln y$ равно

$$\triangle \mathsf{E} \ln y = \mathsf{E}(\ln y'') - \mathsf{E}(\ln y') = \beta_1 \triangle_x$$

Из свойств логарифма следует, что

$$\frac{y''}{y'} = \exp\left(\beta_1 \triangle_x\right),$$

т.е. значение фактора $y$ в среднем изменяется в $\exp\left(\beta_1 \triangle_x\right)$ раз.

Если величина $(\beta_1 \triangle_x)$ достаточно мала, то, воспользовавшись приближением $\exp\left(\beta_1 \triangle_x\right) \approx 1 + \beta_1 \triangle_x$, получаем

$$y'' = y' \exp\left(\beta_1 \triangle_x\right) \approx y'(1 + \beta_1 \triangle_x),$$

т.е. при изменении зависимой переменной на величину $\triangle_x$ значение фактора $y$ изменяется на $(\beta_1 \triangle_x) \cdot 100\%$ процентов.

*Пример* (log-Wage equation [31]). Была оценена полулогарифмическую модель зависимости почасовой оплаты труда $wage$ (в \$) от уровня школьного образования $educ$ (в годах)

$$\widehat{\ln wage} = 0.584 + 0.083 educ.$$

Согласно этой модели дополнительный год образования увеличивает почасовую оплату (в первом приближении!) на 8.3%. Более точно, дополнительный год образования увеличивает почасовую оплату в $\exp(0.083) \approx 1.08654$ раз, т.е. на 8.654%. Также отметим, что в этой модели средний уровень оплаты человека без образования равен $\exp(0.584) \approx 1.793$.

**Лог-линейная модель регрессии** имеет вид

$$\ln y_i = \beta_0 + \beta_1 \ln x_i + \varepsilon_i, \quad n = 1, \ldots, n.$$

Относительно ошибок регрессии предполагается, что они удовлетворяют стандартным требованиям парной линейной регрессии. Тогда наилучшими линейными оценками (BLUE-оценками) параметров регрессии являются OLS-оценки, и к ним применимы все выводы парной линейной модели регрессии (статистические свойства коэффициентов и проч.)

Далее, т.к. $\mathsf{E}\varepsilon_i = 0$, то

$$\mathsf{E}(\ln y_i) = \beta_0 + \beta_1 \ln x_i.$$

Пусть значение фактора $x$ изменяется в $p$ раз ($p > 0$):

$$\begin{aligned} \mathsf{E}(\ln y') &= \beta_0 + \beta_1 \ln x, \\ \mathsf{E}(\ln y'') &= \beta_0 + \beta_1 \ln(px) = \beta_0 + \beta_1 \ln x + \beta_1 \ln p \end{aligned}$$

Тогда ожидаемое (среднее) изменение величины $\ln y$ равно

$$\triangle \mathsf{E} \ln y = \mathsf{E}(\ln y'') - \mathsf{E}(\ln y') = \beta_1 \ln p$$

Из свойств логарифма следует, что

$$\frac{y''}{y'} = p^{\beta_1},$$

т.е. значение фактора $y$ в среднем изменяется в $p^{\beta_1}$ раз.

Если $p = 1 + r$ и $r$ достаточно мало, то $p^{\beta_1} \approx 1 + r\beta_1$, т.е. при изменении $x$ на $r \cdot 100\%$, зависимая переменная $y$ в среднем изменяется на $(r\beta_1) \cdot 100\%$.

В лог-линейной модели регрессии коэффициент $\beta_1$ есть не что иное, как коэффициент **эластичности** $y$ по $x$. В самом деле,

$$E_x = \frac{dy}{dx} \cdot \frac{x}{y} = \frac{d(\ln y)}{d(\ln x)} = \beta_1.$$

Таким образом, лог-линейная модель описывает зависимость с постоянной эластичностью.

*Пример* (Salary and Firm Sales [31]). Была оценена лог-линейная модель зависимости оклада CEO от объема продаж фирмы

$$\ln(\widehat{Salary}) = 5.267 + 0.136 \ln(Sales).$$

Таким образом, при увеличении объема продаж $Sales$ на 1% зависимая переменная $Salary$ увеличивается (в первом приближении!) на $(0.136 \cdot 1)\% = 0.136\%$.

## 1.9. Стохастические регрессоры

До сих пор мы использовали вероятностную модель, в которой значения фактора $x$ считались неслучайными (детерминированными). Однако в некоторых прикладных задачах значения фактора $x$ необходимо считать случайными. Например, в случае когда значения объясняющей переменной были измерены со случайной ошибкой. Также стохастические объясняющие переменные возникают в эконометрических моделях временных рядов. В этом случае статистические выводы должны быть скорректированы.

Итак, рассмотрим следующую вероятностную модель

$$y = \beta_0 + \beta_1 x + u, \tag{1.3}$$

где $y, x, u$ – случайные величины, причем $y$ и $x$ наблюдаемы, а $u$ ненаблюдаемо. Случайную величину $u$ (ошибку), как и раньше, мы представляем себе как «влияние факторов, не включенных в модель». Относительно ошибки $u$ будем предполагать выполнение следующих условий:

1) $\mathsf{E}(u|x) = 0$,

2) $\mathsf{E}(u^2|x) = \sigma^2$,

3) $u|x \sim \mathcal{N}(0, \sigma^2)$.

При выполнении условия 1) очевидно

$$\mathsf{E}(y|x) = \beta_0 + \beta_1 x.$$

Рассмотрим задачу оценивания параметров $\beta_0$, $\beta_1$ и $\sigma^2$ на основе выборочных данных $(x_i, y_i)$. Основной результат дается следующей теоремой.

**Теорема** (Гаусс – Марков)**.** *Пусть для линейной модели* (1.3) *выполнены условия 1) и 2) и* $(x_i, y_i)$ – ***случайная выборка****. Тогда OLS-оценки* $\hat{\beta}_0$ *и* $\hat{\beta}_1$ *параметров* $\beta_0$ *и* $\beta_1$ *будут линейными несмещенными оценками с минимальной дисперсией*[9]*, т.е. BLUE оценками. Кроме того, эти оценки состоятельны*[10]*, т.е.*

$$\hat{\beta}_0 \underset{\mathsf{P}}{\longrightarrow} \beta_0, \quad \hat{\beta}_1 \underset{\mathsf{P}}{\longrightarrow} \beta_1 \ (n \longrightarrow +\infty).$$

[9]Имеется ввиду условная дисперсия $\mathrm{Var}(\cdot|x_1, \ldots, x_n)$.

[10]Напомним, что состоятельность означает сходимость по вероятности: $\hat{\beta}_j \underset{\mathsf{P}}{\longrightarrow} \beta_j \Leftrightarrow$ для всех $c > 0$ вероятность $\mathsf{P}\left(\left|\hat{\beta}_j - \beta_j\right| > c\right) \to 0$ при $n \to +\infty$.

Так же как в случае детерминированных значений $x_i$ определяются предсказанные значения зависимой переменной $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, полная TSS, объясненная ESS и остаточная RSS суммы квадратов. Для них верно равенство

$$\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS}\,.$$

Коэффициент $R^2$ определяется равенством

$$R^2 = \frac{\mathrm{ESS}}{\mathrm{RSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} = \widehat{\mathrm{corr}}^2(x, y).$$

**Теорема.** *При выполнении условий 1), 2) и 3) для OLS-оценок параметров регрессии и для коэффициента детерминации верны все статистические свойства модели регрессии с детерминированными значениями регрессора $x_i$.*

*Замечание.* Следует отметить, что статистические свойства для модели со стохастическими регрессорами следует понимать в смысле условных распределений: $t_1|x_1 \ldots, x_n \sim t_{n-2}$ и т.д.

## 1.10. Задачи

1. Рассчитанный по выборочным данным коэффициент корреляции оказался равным $-1$. Это означает что:

   (a) между изучаемыми переменными есть слабая отрицательная линейная связь;

   (b) между изучаемыми переменными есть связь, но она не является линейной;

   (c) между изучаемыми переменными есть функциональная линейная отрицательная связь;

   (d) между изучаемыми переменными отсутствует связь;

   (e) полученное число никак не интерпретируется, допущена ошибка в вычислениях;

   (f) между переменными есть функциональная положительная связь;

   (g) между изучаемыми переменными достаточно сильная прямая линейная корреляционная зависимость;

   (h) ни один из предложенных ответов не является правильным;

2. Рассматривается КЛМПР $y_i = \theta_0 + \theta_1 x_i + \varepsilon_i$, где $i = 1, \ldots, n$, $E(\varepsilon_i) = 0$, $cov(\varepsilon_i, \varepsilon_j) = \sigma^2 \delta_{ij}$, $i, j = 1, \ldots, n$. Даны девять пар наблюдений

| $x$ | $-1$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|---|
| $y$ | 1 | $-2$ | $-1$ | $-4$ | $-4$ | $-6$ | $-5$ | $-8$ | $-7$ |

   Для модели КЛМПР найти

а) МНК-оценки параметров $\hat{\theta}_0$ и $\hat{\theta}_1$, использую матричную запись;
б) оценку дисперсии остатков;
в) оценку дисперсий и ковариации параметров;
г) наилучший точечный прогноз для значения $\hat{y}$ в точке $x = 8$;
д) дисперсию этого прогноза;
е) интервальный прогноз для значения $\hat{y}$ в точке $x = 8$;
ё) $TSS$ и $RSS$;
ж) величину $R^2$;
з) проверить гипотезу $\theta_0 = 0$ против гипотезы $\theta_0 \neq 0$;
и) построить доверительный интервал для параметра $\theta_1$.
Доверительную вероятность положить равной 0.95.

3. Наблюдения 16 пар $(x,y)$ дали следующие результаты:

$$\Sigma y^2 = 526,\ \Sigma x^2 = 657,\ \Sigma xy = 492,\ \Sigma y = 64,\ \Sigma x = 96.$$

Найдите коэффициент корреляции и оцените регрессию $y_t = \theta_0 + \theta_1 x_t + \varepsilon_t$.

4. Пусть задано $n$ наблюдений $\{x_i,y_i\}$. Для линейной функции $y_i = \theta x_i$ требуется:

1)Вывести нормальную систему линейных уравнений для нахождения параметров (оптимальной) прямой.

2)Вывести формулу для оценки $\hat{\theta}$.

3)Если по той же выборке найти оптимальную прямую со свободным членом $y_i = \theta_0 + \theta_1 x_i$, будет ли коэффициент наклона $\hat{\theta}_1$ совпадать с $\hat{\theta}$? А знаки коэффициентов наклона?

5. Получить систему нормальных уравнений для функций

$$y_i = \theta_0 + \theta_1 x + \theta_2 z,$$
$$y_i = \theta_0 + \theta_1 x + \theta_2 x^2.$$

6. Пусть $\hat{\theta}$ есть МНК-оценка коэффициента наклона линейной функции $y$ на $x$ без константы, а $\hat{\gamma}$ - МНК-оценка коэффициент наклона в линейной функции $x$ на $y$ без константы. Верно ли для этих оценок равенство

$$\hat{\theta}\hat{\gamma} = 1?$$

7. Пусть $\hat{\theta}$ есть МНК-оценка коэффициента наклона линейной функции $y$ на $x$ с константой, а $\hat{\gamma}$ - МНК-оценка коэффициент наклона в линейной функции $x$ на $y$ с константой. Верно ли для этих оценок равенство

$$\hat{\theta}\hat{\gamma} = 1?$$

8. Пусть $\hat{\theta}$ есть МНК-оценка коэффициента наклона линейной функции $y$ на $x$ с константой, а $\hat{\gamma}$ - МНК-оценка коэффициент наклона в линейной функции $x$ на $y$ с константой. Покажите, что

$$\hat{\theta}\hat{\gamma} = 1 \Longleftrightarrow \widehat{corr}(x, y) = \pm 1.$$

9. Для КЛМПР $y_i = \theta_0 + \theta_1 x_i + \varepsilon_i$ рассмотрим следующую оценку коэффициента $\theta_1$:

$$\tilde{\theta}_1 = \frac{1}{n} \sum_{t=1}^{n} \frac{y_i - \bar{y}}{x_i - \bar{x}}.$$

   а) Является ли оценка $\tilde{\theta}_1$ несмещенной?

   б) Является ли она линейной?

   в) Наилучшей (с наименьшей дисперсией)?

10. Рассмотрим линейную модель регрессии без константы $y_i = \theta x_i + \varepsilon_i$. Будут ли эти оценки коэффициента наклона $\theta$

$$\tilde{\theta} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{x_i} \quad \text{и} \quad \breve{\theta} = \frac{\sum\limits_{i=1}^{n} y_i}{\sum\limits_{i=1}^{n} x_i}$$

   а) несмещенными?

   б) линейными по $y$?

   в) наилучшими (с минимальной дисперсией)?

   г) При каких условиях они являются наилучшими среди несмещенных линейных оценок параметра $\theta$?

11. В условиях Гаусса-Маркова докажите, что

   а) $D(\varepsilon_i) = E(\varepsilon_i^2)$;

   б) $D(y_i) = D(\varepsilon_i)$ и для $i \neq j$ $cov(y_i, y_j) = 0$;

   в) матрица ковариаций имеет вид $\sigma^2 I$.

12. Доказать, что

$$\widehat{\theta}_1 = \widehat{corr}(x, y) \frac{\hat{\sigma}_y}{\hat{\sigma}_x}.$$

13. Существуют ли оценки, которые имеют меньшую дисперсию, чем МНК-оценки? А несмещенные оценки?

14. Докажите, что для линейной регрессии с константой

   а) $TSS = ESS + RSS$;

   б) $R^2 = (\widehat{corr}(x, y))^2$;

15. Рассмотрим парную модель линейной регрессии без константы $y_i = \theta x_i + \varepsilon_i$.

    а)Верно ли, что $TSS = ESS + RSS$?

    б) Покажите, что для этой модели выполнено равенство $\sum y_i^2 = \sum \hat{y}_i{}^2 + \sum e_i^2$, где $\hat{y}_i = \hat{\theta} x_i$ - предсказанные значения и $e_i = y_i - \hat{y}_i$ - остатки.

    в) Как в этом случае определить коэффициент детерминации $R^2$?

16. Проинтерпретируйте коэффициенты регрессии в линейной, полулогарифмической и лог-линейной моделях. Рассмотрите примеры

$$\widehat{Wage} = -0,62 + 0,45Educ,$$

$$\ln\widehat{Wage} = 0,584 + 0,083Educ,$$

$$\ln\widehat{Salary} = 5,267 + 0,136\ln(Sales).$$

17. Докажите, что в лог-линейной модели коэффициент $\theta_1$ есть эластичность $E_x$ $y$ по $x$.

18. Для изучения влияния образования на величину почасовой оплаты труда на основе опроса 40 человек была оценена модель регрессии (в скобках указаны стандартные ошибки коэффициентов)

$$\ln(\widehat{Wage}) = \underset{(0.02)}{2.2} + \underset{(0.04)}{0.1}\, Edu,$$

    где $Edu$ - уровень образования (в годах), $Wage$ - уровень почасовой оплаты труда.

    - Дайте интерпретацию коэффициентов модели.
    - Значимо ли уровень образования влияет на почасовую оплату труда при уровне значимости 10%? 5%? 1%?
    - Какой ожидаемый уровень почасовой оплаты труда человека с одиннадцатилетним образованием? С девятилетним образование?

19. Для изучения функции спроса на некоторый товар была оценена регрессионная модель зависимости спроса от цены (в \$)

$$\ln(\widehat{Sales}) = \underset{(0.07)}{0.91} - \underset{(0.2)}{1.21}\ln(Price) \qquad n = 25$$

    (в скобках указаны стандартные ошибки коэффициентов).

    - Дайте интерпретацию коэффициентов. Чему равна эластичность спроса по цене?

- Какой ожидаемый объем продаж при цене \$2? \$1.5?
- Постройте доверительный интервал для эластичности спроса по цене с доверительной вероятностью 90%, 95%, 98%, 99%.
- Значимо ли эластичность отличается от -1 при уровне значимости 10%? 5%? 2%? 1%?

20. (Задача 2.12 из [8])
Рассмотрим модель регрессии на константу

$$Y_t = \alpha + \varepsilon_t, \qquad t = 1, ..., n.$$

а) Найдите оценки метода наименьших квадратов для $\alpha$ и $\sigma^2$.

б) Найдите дисперсию оценки $\hat{\alpha}$.

в) Покажите, что статистика $\dfrac{\hat{\alpha} - \alpha}{s_{\hat{\alpha}}}$ имеет распределение $t(n-1)$.

г) Чему равен коэффициент детерминации $R^2$?

# Глава 2

# Многофакторная регрессия

## 2.1. Многофакторная регрессия

Естественным обобщением модели парной регрессии является модель множественной регрессии, когда рассматривается влияние нескольких факторов на зависимую переменную $y$.

Будем рассматривать следующую вероятностную модель множественной регрессии

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, \ldots, n \tag{2.1}$$

где $y$ – *зависимая переменная*, $x_1, \ldots, x_k$ – *регрессоры*, *влияющие* или *объясняющие переменные*, $\varepsilon$ – *ошибки* модели регрессии, $\beta_0, \ldots, \beta_k$ – параметры или коэффициенты в модели регрессии, $i$ – номер наблюдения. Через $m$ будем обозначать число коэффициентов регрессионной модели ($m = k+1$). Как и в случае парной регрессии сначала рассмотрим случай неслучайных (детерминированных) регрессоров, $y$ и $\varepsilon$ являются случайными величинами.

Запишем уравнение регрессии (2.1) в матричном виде. Для этого введем следующие обозначения

$$\mathbf{x}_i = \begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{ik} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}.$$

Тогда уравнение (2.1) можно записать в виде

$$y_i = \beta' \mathbf{x}_i + \varepsilon_i = \mathbf{x}_i' \beta + \varepsilon_i, \quad i = 1, \ldots, n,$$

где штрих означает операцию транспонирования матриц.

Обозначим

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Из определения видно, что $\mathbf{X}$ – матрица размера $n \times m$, первая строка матрицы $\mathbf{X}$ равна $\mathbf{x}_1'$, вторая строка равна $\mathbf{x}_2'$ и т.д. Тогда уравнения (2.1) могут быть записаны в виде одного матричного уравнения

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon.$$

## 2.2. Метод наименьших квадратов

Найдем оценку наименьших квадратов для вектора параметров $\beta$ при известных значениях зависимой переменной $y$ и регрессоров $x_1, \ldots, x_k$. Запишем систему нормальных уравнений в матричном виде. Для этого сумму квадратов отклонений перепишем в виде

$$S(\beta_0, \ldots, \beta_k) = \sum_i (y_i - \mathbf{x}_i'\beta)^2 = (\mathbf{y} - \mathbf{X}\beta)' \cdot (\mathbf{y} - \mathbf{X}\beta).$$

Тогда оценки наименьших квадратов коэффициентов регрессии находятся из условия

$$S(\beta_0, \ldots, \beta_k) \longrightarrow \min.$$

В матричном виде необходимые условия экстремума имеют вид

$$\operatorname{grad} S = \left(\frac{\partial S}{\partial \beta_0}, \frac{\partial S}{\partial \beta_1}, \ldots, \frac{\partial S}{\partial \beta_k}\right) = 2(\mathbf{X}'\mathbf{X}\beta - \mathbf{X}'\mathbf{y})' = 0$$

и система нормальных уравнений записывается в матричном виде как

$$(\mathbf{X}'\mathbf{X})\beta = \mathbf{X}'\mathbf{y}.$$

При $\det(\mathbf{X}'\mathbf{X}) \neq 0$ эта система имеет единственное решение и OLS-оценка параметров линейной модели регрессии равна

$$\widehat{\beta}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y}).$$

Можно показать, что симметричная матрица $(\mathbf{X}'\mathbf{X}) \geqslant 0$. Так как гессиан функции $S(\cdot)$ равен $\nabla^2 S = 2(\mathbf{X}'\mathbf{X}) \geqslant 0$, то $S(\cdot)$ – выпуклая функция. Следовательно, решение системы нормальных уравнений будет глобальным минимумом функции $S(\cdot)$.

*Пример.* Найти МНК-оценки параметров модели в матричном виде: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$,

где

$$Y = \begin{pmatrix} 2 \\ -1 \\ 1 \\ 3 \\ -2 \\ 0 \end{pmatrix} \quad X = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 2 \\ 1 & 2 & -1 \\ 1 & 0 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

*Решение.* Используем матричный вид формул для МНК-оценок:

$$\widehat{\beta}_{\text{мнк}} = (X^T X)^{-1}(X^T Y).$$

Подставим исходные значения:

$$\widehat{\beta}_{\text{мнк}} = \left( \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 2 \\ 1 & 2 & -1 \\ 1 & 0 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}^T \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 2 \\ 1 & 2 & -1 \\ 1 & 0 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 2 \\ 1 & 2 & -1 \\ 1 & 0 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}^T \begin{pmatrix} 2 \\ -1 \\ 1 \\ 3 \\ -2 \\ 0 \end{pmatrix}$$

$$\widehat{\beta}_{\text{мнк}} = \begin{pmatrix} 6 & 5 & 7 \\ 5 & 11 & 1 \\ 7 & 1 & 15 \end{pmatrix}^{-1} \begin{pmatrix} 3 \\ 1 \\ 3 \end{pmatrix}$$

$$\widehat{\beta}_{\text{мнк}} \approx \begin{pmatrix} 1.17 & -0.49 & -0.51 \\ -0.49 & 0.29 & 0.21 \\ -0.51 & 0.21 & 0.29 \end{pmatrix} \begin{pmatrix} 3 \\ 1 \\ 3 \end{pmatrix} = \begin{pmatrix} 1.48 \\ -0.54 \\ -0.45 \end{pmatrix}$$

## 2.3. Основные предположения. Теорема Гаусса – Маркова

Рассмотрим теперь вероятностные свойства многофакторной линейной модели регрессии. Будем предполагать выполнение следующих условий на ошибки регрессии:

1. $\mathsf{E}\varepsilon_i = 0$ для всех $i = 1, \ldots, n$.
2. $\mathrm{Var}(\varepsilon_i) = \sigma^2$ не зависит от $i$.
3. $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$ при $i \neq j$ (некоррелируемость ошибок для разных наблюдений).
4. $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, $i = 1, \ldots, n$ (нормальная распределенность ошибок регрессии).

Условия 1. – 3. на ошибки регрессии могут быть записаны в матричном виде

$$\mathsf{E}\varepsilon = 0, \quad \mathrm{Var}(\varepsilon) = \sigma^2 I,$$

где $I$ – единичная $m \times m$ матрица, а $\mathrm{Var}(\varepsilon)$ – $(n \times n)$ матрица ковариации случайного вектора $\varepsilon$.

В матричной записи условие 4. означает, что $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$, т.е. случайный вектор ошибок регрессии имеет совместное многомерное нормальное распределение с нулевым средним и ковариационной матрицей $\sigma^2 I$.

Из условия 1. получаем (для простоты опустим индекс $i$)

$$\mathsf{E}y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k = \mathbf{x}'\beta.$$

При изменении значения фактора $x_j$ на величину $\Delta_j$ среднее значение зависимой переменной изменяется на величину $\Delta\mathsf{E}y = \beta_j\Delta_j$. Коэффициент $\beta_j$ в линейной модели, таким образом, можно трактовать как «средний эффект» от увеличения на единицу значения регрессора $x_j$, т.е. как маржинальное или предельное значение (в усредненном смысле). Из условий 2. и 3. на ошибки регрессии следует, что

$$\mathrm{Var}(y_i) = \sigma^2, \quad \mathrm{cov}(y_i, y_j) = 0\,(i \neq j)$$

или в матричной записи $\mathrm{Var}(\mathbf{y}) = \sigma^2 I$.

Покажем, что OLS-оценки параметров $\widehat{\beta}_{OLS}$ являются «наилучшими» среди несмещенных линейных оценок, а именно несмещенными линейными оценками с наименьшей дисперсией или BLUE оценками (Best Linear Unbiased Estimator).

**Теорема** (Гаусс – Марков)**.** *Пусть выполнены условия 1. – 3. на ошибки линейной регрессии*

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon.$$

*Тогда OLS-оценки являются BLUE оценками.*

*Доказательство.* **1.** Докажем несмещенность. Имеем, используя детерминированность регрессоров,

$$\mathsf{E}\left(\widehat{\beta}_{OLS}\right) = \mathsf{E}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\right) = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathsf{E}\mathbf{y}) = \\ (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathsf{E}(\mathbf{X}\beta + \varepsilon) = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\beta = \beta.$$

**2.** Найдем дисперсию OLS-оценки. Из свойств дисперсий получаем выражение для $m \times m$ матрицы ковариации вектора $OLS$-оценок

$$\begin{aligned}\mathrm{Var}\left(\widehat{\beta}_{OLS}\right) &= \mathrm{Var}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\right) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathrm{Var}(\mathbf{y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2 I)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.\end{aligned}$$

Пусть $\widetilde{\beta}$ – другая несмещенная, линейная по $y_i$ оценка параметра $\beta$. Представим ее в виде

$$\widetilde{\beta} = \widehat{\beta}_{OLS} + Q\mathbf{y},$$

где $Q$ – матрица размера $m \times n$. Далее,

$$\mathsf{E}\widetilde{\beta} = \mathsf{E}\widehat{\beta}_{OLS} + Q\mathsf{E}\mathbf{y} = \beta + Q\mathbf{X}\beta.$$

Так как $\widetilde{\beta}$ – несмещенная оценка ($\mathsf{E}\widetilde{\beta} = \beta$), то $Q\mathbf{X} = 0$. Используя это наблюдение, можно показать [7], что

$$\mathrm{Var}\left(\widetilde{\beta}\right) = \mathrm{Var}\left(\widehat{\beta}_{OLS}\right) + \sigma^2 QQ'.$$

Следовательно, OLS-оценка имеет наименьшую дисперсию в классе всех линейных несмещённых оценок. □

*Замечание.* Из доказательства видно, что для несмещенности OLS-оценок достаточно выполнения условия $\mathsf{E}\varepsilon_i = 0$ на ошибки регрессии.

Если кроме условий 1. – 3. выполнено условие 4. нормальной распределенности ошибок регрессии, то случайные величины $y_i$ также имеют нормальное распределение

$$y_i \sim \mathcal{N}\left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}, \sigma^2\right)$$

или в матричной записи $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 I)$.

Следовательно, вектор OLS-оценок параметров линейной регрессии $\widehat{\beta}_{OLS}$ имеет (многомерное) нормальное распределение

$$\widehat{\beta}_{OLS} \sim \mathcal{N}\left(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right)$$

с ковариационной матрицей

$$\mathrm{Var}\left(\widehat{\beta}_{OLS}\right) = \left(\mathrm{cov}(\hat{\beta}_i, \hat{\beta}_j)\right) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

Обозначим через $b_j$ $(j = 0, \ldots, k)$ диагональные элементы матрицы $(\mathbf{X}'\mathbf{X})^{-1}$:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} b_0 & & & \star \\ & b_1 & & \\ & & \ddots & \\ \star & & & b_k \end{pmatrix}.$$

Тогда OLS-оценка $\hat{\beta}_j$ коэффициента регрессии $\beta_j$ имеет нормальное распределение

$$\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2 b_j).$$

Укажем теперь несмещенную оценку дисперсии ошибок регрессии $\sigma^2$. Обозначим через

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik} = \mathbf{x}_i' \cdot \widehat{\beta}_{OLS}$$

*предсказанные* или *подогнанные* (fitted) значения фактора $y$. В матричной записи

$$\widehat{\mathbf{y}} = \mathbf{X} \cdot \widehat{\beta}_{OLS}.$$

**Определение.** *Остатки* (residual) в линейной модели регрессии (2.1) определяются как $e_i = y_i - \hat{y}_i$.

Как и в случае парной регрессии, обозначим через

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

*остаточную сумму квадратов* регрессии. Можно показать, что

$$\mathsf{E}(\mathrm{RSS}) = (n - m)\sigma^2.$$

Следовательно, статистика

$$s^2 = \frac{\text{RSS}}{n-m} = \frac{1}{n-m}\sum_{i=1}^{n} e_i^2$$

является несмещенной оценкой дисперсии ошибок в модели линейной регрессии. Величина $\text{SER} = s = \sqrt{s^2}$ называется *стандартной ошибкой регрессии* (Standard Error of Regression).

## 2.4. Статистические свойства OLS-оценок. Доверительные интервалы и проверка гипотез

**1.** Выше мы показали, что OLS-оценки $\hat{\beta}_j$ коэффициентов линейной модели регрессии имеют нормальное распределение с дисперсией $\sigma^2 b_j$, где $\sigma^2$ – дисперсия ошибок регрессии. Матрица ковариации вектора OLS-оценок коэффициентов модели регрессии равна

$$\text{Var}\left(\widehat{\beta}_{OLS}\right) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

Однако в большинстве прикладных задач значение дисперсии ошибок неизвестно, и вместо нее используют несмещенную оценку $s^2$. Выборочная оценка дисперсии OLS-оценок, таким образом, равна

$$\widehat{\text{Var}}\left(\hat{\beta}_j\right) = s^2 b_j \quad (j = 0, \ldots, k),$$

а *выборочное стандартное отклонение* или *стандартная ошибка* коэффициента регрессии $\widehat{\beta}_j$ определяется как

$$s_j = \sqrt{\widehat{\text{Var}}\left(\hat{\beta}_j\right)} = \sqrt{s^2 b_j} = s\sqrt{b_j}.$$

Обычно при записи оцененной модели регрессии выборочные стандартные ошибки коэффициентов записываются в круглых скобках под коэффициентами. Выборочная матрица ковариации вектора оценок определяется как

$$\widehat{\text{Var}}\left(\widehat{\beta}_{OLS}\right) = s^2(\mathbf{X}'\mathbf{X})^{-1}$$

и ее диагональные элементы равны $s_j^2$ $(j = 0, 1, \ldots, k)$.

Основной результат о статистических свойствах оценок коэффициентов регрессии дается следующей теоремой.

**Теорема.** *Пусть выполнены условия 1. – 4. на ошибки регрессии. Тогда*

*a) случайные величины $s^2$ и $\widehat{\beta}_{OLS}$ независимы,*

*b) статистика $(n-m)s^2/\sigma^2$ имеет распределение $\chi^2_{n-m}$,*

*c) статистика* $t_j = (\widehat{\beta}_j - \beta_j)/s_j$ *имеет распределение* $t_{n-m}$.

Теорема доказывается так же, как и в случае парной регрессии.

**2.** Приведем формулы доверительных интервалов для коэффициентов регрессии с доверительной вероятностью $\gamma$. При выполнении условий 1. – 4. на ошибки регрессии статистика $t_j$ имеет распределение Стьюдента $t_{n-m}$. Пусть $t_{\text{кр}} = t(\alpha; n-m)$ – **двустороннее** критическое значение распределения $t_{n-m}$ с уровнем значимости $\alpha = 1 - \gamma$. Тогда

$$\mathsf{P}(|t_j| > t_{\text{кр}}) = \alpha \Longrightarrow \mathsf{P}(|t_j| < t_{\text{кр}}) = \gamma$$

Так как (аналогично случаю парной регрессии)

$$|t_j| < t_{\text{кр}} \Longleftrightarrow \widehat{\beta}_j - t_{\text{кр}} \cdot s_j < \beta_j < \widehat{\beta}_j + t_{\text{кр}} \cdot s_j,$$

то получаем выражение для доверительного интервала для коэффициента регрессии $\beta_j$ с доверительной вероятностью $\gamma$:

$$\mathsf{P}\left(\widehat{\beta}_j - t_{\text{кр}} \cdot s_j < \beta_j < \widehat{\beta}_j + t_{\text{кр}} \cdot s_j\right) = \gamma.$$

**3.** Приведем статистический критерий для проверки гипотезы

$$H_0 \ : \ \beta_j = \theta_0$$

($\theta_0$ – заданное число) против альтернативы

$$H_0 \ : \ \beta_j \neq \theta_0.$$

При справедливости нулевой гипотезы статистика

$$t = \frac{\widehat{\beta}_j - \theta_0}{s_j}$$

имеет распределение Стьюдента

$$t \underset{H_0}{\sim} t_{n-m}.$$

Пусть $t_{\text{кр}} = t(\alpha, n-m)$ – **двустороннее** критическое значение распределения $t_{n-m}$ с заданным уровнем значимости $\alpha$. Тогда при справедливости $H_0$ вероятность $\mathsf{P}(|t| > t_{\text{кр}}) = \alpha$ мала. Для проверки гипотезы имеем следующий статистический критерий:

- если $|t| > t_{\text{кр}}$, то нулевая гипотеза $H_0$ отвергается в пользу альтернативы $H_1$ при заданном уровне значимости. Также говорят, что коэффициент значимо отличается от числа $\theta_0$;
- если $|t| < t_{\text{кр}}$, то данные согласуются с нулевой гипотезой (не противоречат ей) при заданном уровне значимости. Также говорят, что коэффициент незначимо отличается от числа $\theta_0$.

*Замечание.* Несложно проверить, что $|t| < t_{\text{кр}}$ тогда и только тогда, когда число $\theta_0$ принадлежит доверительному интервалу для коэффициента $\beta_j$ с доверительной вероятностью $1-\alpha$. Таким образом, мы получаем альтернативный способ проверки нулевой гипотезы:

*гипотеза $H_0$ отвергается при заданном уровне значимости $\Longleftrightarrow$ значение $\theta_0$ не принадлежит доверительному интервалу, построенному для доверительной вероятности $1-\alpha$.*

Этот критерий бывает полезен в прикладных задачах, т.к. многие эконометрические пакеты вычисляют доверительные интервалы автоматически.

В случае *проверки значимости* коэффициента регрессии, т.е. при проверке нулевой гипотезы

$$H_0 \,:\, \beta_j = 0$$

(фактор $x_j$ не влияет на зависимую переменную) при двусторонней альтернативе $t$-статистика вычисляется как

$$t_j = \frac{\widehat{\beta}_j}{s_j}$$

и именно это значение выводится в эконометрических программах. Коэффициент $\beta_j$ **значим** (нулевая гипотеза отвергается) при $|t_j| > t_{\text{кр}}$.

**4.** Приведем статистический критерий для тестирования гипотезы

$$H_0 \,:\, \beta_j = \theta_0$$

($\theta_0$ – заданное значение) против **односторонней** альтернативы

$$H_1 \,:\, \beta_j > \theta_0.$$

Пусть $t'_{\text{кр}} = t(\alpha; n-m)$ – **одностороннее**[1] критическое значение распределения $t_{n-m}$ при заданном уровне значимости $\alpha$. Если $H_0$ верна, то статистика

$$t = \frac{\widehat{\beta}_j - \theta_0}{s_j} \underset{H_0}{\sim} t_{n-m}$$

и вероятность события $\mathsf{P}(t > t'_{\text{кр}}) = \alpha$ мала. Для проверки нулевой гипотезы против односторонней альтернативы получаем следующий статистический критерий:

- если $t > t'_{\text{кр}}$, то гипотеза $H_0$ отвергается в пользу альтернативы $H_1$ при заданном уровне значимости;
- если $t < t'_{\text{кр}}$, то данные согласуются с нулевой гипотезой при заданном уровне значимости.

Аналогично проверяется гипотеза $H_0$ против односторонней альтернативы $H_1 : \beta_1 < \theta_0$.

[1] Одностороннее критическое значение находится из условия $\mathsf{P}(t_{n-m} > t'_{\text{кр}}) = \alpha$.

*Пример.* Была оценена модель регрессии зависимости зарплаты CEO salary (тыс) от объема продаж фирмы sales (млн), доходности на собственный капитал roe, доходности акций ros и фиктивных переменных indus и finance (принадлежит ли фирма индустриальному или финансовому сектору). Зависимая переменная: $\ln(salary)$; объем выборки 209.

| $Y$ | $const$ | $ln(sales)$ | $roe$ | $ros$ | $indus$ | $finance$ |
|---|---|---|---|---|---|---|
| $coeff$ | 4.23 | 0.28 | 0.018 | 0.00033 | 0.011 | 0.1903 |
| $Std.Err$ | 0.3178 | 0.035 | 0.0041 | 0.00055 | 0.079 | 0.088 |

$$R^2 = 0,3006, \quad R^2_{adj} = 0,2834$$

Проверьте значимость коэффициентов.

*Решение.* Вычисляем для каждого коэффициента $t$-статистику $t = \frac{\widehat{\beta}_i}{s(\widehat{\beta}_i)}$.

| $Y$ | $const$ | $ln(sales)$ | $roe$ | $ros$ | $indus$ | $finance$ |
|---|---|---|---|---|---|---|
| $coeff$ | 4.23 | 0.28 | 0.018 | 0.00033 | 0.011 | 0.1903 |
| $Std.Err$ | 0.3178 | 0.035 | 0.0041 | 0.00055 | 0.079 | 0.088 |
| $t-stat$ | 13.64 | 8 | 4.39 | 0.6 | 0.14 | 2.16 |

Критическое значение находим в таблице $t_{\text{кр}}(0.05; 203) = 1.96$. Те показатели, для которых $t_{\text{расчетное}} > t_{\text{кр}}$, являются значимыми.

## 2.5. Коэффициент $R^2$. Проверка сложных гипотез о коэффициентах регрессии

**1.** Как и в случае парной регрессии введем обозначения

- $\text{TSS} = \sum (y_i - \bar{y})^2$ – общая вариация зависимой переменной $y$ (Total Sum of Squares, общая сумма квадратов);
- $\text{ESS} = \sum (\hat{y}_i - \bar{y})^2$ – вариация зависимой переменной, объясненная регрессией (Explained Sum of Squares, объясненная сумма квадратов);
- $\text{RSS} = \sum (y_i - \hat{y}_i)^2 = \sum e_i^2$ – остаточная часть вариации зависимой переменной (Residual Sum of Squares, остаточная сумма квадратов).

Для модели регрессии с включенной константой $\beta_0$ верно равенство

$$\text{TSS} = \text{ESS} + \text{RSS}\,.$$

**Определение.** Коэффициент $R^2$ определяется как доля объясненной регрессией суммы квадратов зависимой переменной в общей сумме квадратов:

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}.$$

Из определения непосредственно следует, что всегда $0 \leqslant R^2 \leqslant 1$. Кроме того, для крайних значения коэффициента детерминации имеем

- $R^2 = 0 \Leftrightarrow \mathrm{ESS} = 0 \Leftrightarrow \hat{y}_i = \bar{y} \Leftrightarrow \hat{\beta}_1 = \cdots = \hat{\beta}_k = 0$, т.е. значения регрессоров «не улучшают качество прогноза» фактора $y$
- $R^2 = 1 \Leftrightarrow \mathrm{RSS} = 0 \Leftrightarrow e_i = 0 \Leftrightarrow y_i = \hat{y}_i$, т.е. получаем «идеальную подгонку» линейной регрессии на выборочных данных.

Таким образом, коэффициент $R^2$ можно рассматривать как меру «качества подгонки» («goodness-of-fit») регрессионной модели: чем ближе значение $R^2$ к 1, тем «лучше качество подгонки» уравнения регрессии на выборочных данных.

При сравнении разных моделей регрессии необходимо учитывать два важных обстоятельства:

1. при добавлении в модель новых регрессоров коэффициент детерминации **не уменьшается** (а практически почти во всех прикладных задачах возрастает);
2. при преобразовании зависимой переменной коэффициент детерминации изменяется. Следовательно, сравнивать можно только модели с **одинаковыми зависимыми переменными**.

Если число регрессоров плюс константа $\beta_0$ равно объему выборки, то можно добиться, что $R^2 = 1$. Однако это не означает, что модель будет содержательной с экономической точки зрения.

Для устранения эффекта возрастания $R^2$ при добавлении в модель новых факторов можно использовать *скорректированный* (adjusted) на число регрессоров коэффициент $R^2$

$$R^2_{adj} = \bar{R}^2 = 1 - \frac{\mathrm{RSS}\,/(n-m)}{\mathrm{TSS}\,/(n-1)}.$$

Так как $\mathrm{RSS}\,/\,\mathrm{TSS} = 1 - R^2$, то

$$R^2_{adj} = 1 - \left(1 - R^2\right)\frac{n-1}{n-m}.$$

Из этого, в частности следует, что

$$R^2_{adj} \leqslant 1,$$

однако, в отличие от коэффициента $R^2$, скорректированный коэффициент $R^2_{adj}$ может принимать отрицательные значения. Кроме того, можно показать, что

$$R^2_{adj} \leqslant R^2.$$

**2.** Приведем статистический критерий для проверки сложной статистической гипотезы

$$H_0 \;:\; \beta_1 = \cdots = \beta_k = 0$$

т.е. проверки значимости одновременно **всех** коэффициентов при регрессорах (**все** включенные в модель регрессоры не оказывают влияния на зависимую переменную $y$). Также говорят о проверке значимости регрессии «в целом». Альтернативная гипотеза

$$H_1 \ : \ \beta_1^2 + \cdots + \beta_k^2 > 0$$

(не все коэффициенты равны нулю). Рассмотрим статистику

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-m}{m-1} = \frac{\text{ESS}}{\text{RSS}} \cdot \frac{n-m}{m-1}.$$

При справедливости нулевой гипотезы $F$-статистика имеет распределение Фишера

$$F \underset{H_0}{\sim} F_{m-1,n-m}.$$

Пусть $F_{\text{кр}} = F(\alpha; m-1, n-m)$ – критическое значение распределения Фишера $F_{m-1,n-m}$ с уровнем значимости $\alpha$. Имеем следующий статистический критерий проверки гипотезы:

- если $F > F_{\text{кр}}$, то нулевая гипотеза отвергается при заданном уровне значимости (произошло маловероятное с точки зрения нулевой гипотезы событие). Также говорят, что регрессия «в целом» значима;
- если $F < F_{\text{кр}}$, то данные согласуются с нулевой гипотезой при заданном уровне значимости. Также говорят, что регрессия «в целом» незначима.

Так как при больших значениях второй степени свободы распределение Фишера приближается распределением хи-квадрат

$$F_{m-1,n-m} \approx \frac{1}{m-1}\chi^2_{m-1}, \quad (n-m >> 1),$$

то при больших объемах выборки можно использовать приближенный (асимптотический) статистически критерий проверки нулевой гипотезы. Пусть $\chi^2_{\text{кр}} = \chi^2(\alpha, m-1)$ – критической значение распределения $\chi^2_{m-1}$ с уровнем значимости $\alpha$, тогда

- при $(m-1)F > \chi^2_{\text{кр}}$ гипотеза $H_0$ отвергается в пользу $H_1$,
- при $(m-1)F < \chi^2_{\text{кр}}$ данные согласуются с нулевой гипотезой.

*Пример.* Была оценена модель регрессии зависимости зарплаты CEO salary (тыс) от объема продаж фирмы sales (млн), доходности на собственный капитал roe, доходности акций ros и фиктивных переменных indus и finance (принадлежит ли фирма индустриальному или финансовому сектору). Зависимая переменная: ln(salary); объем выборки 209.

| $Y$ | $const$ | $ln(sales)$ | $roe$ | $ros$ | $indus$ | $finance$ |
|---|---|---|---|---|---|---|
| $coeff$ | 4.23 | 0.28 | 0.018 | 0.00033 | 0.011 | 0.1903 |
| $Std.Err$ | 0.3178 | 0.035 | 0.0041 | 0.00055 | 0.079 | 0.088 |

$$R^2 = 0,3006;\ R^2_{adj} = 0,2834$$

Проверьте значимость регрессии в целом на 5% уровне значимости и сформулируйте проверяемую статистическую гипотезу. Как интерпретируется проверяемая статистическая гипотеза?

*Решение.* Гипотеза о значимости регрессии в целом — это гипотеза об одновременном равенстве нулю всех коэффициентов при факторах регрессионной модели.

Гипотеза о том, что регрессия не значима, записывается в виде $H_0 : \beta_i = 0, i = \{1; 2; 3; 4; 5\}$. Альтернативную гипотезу можно записать так $H_1 : \beta_1^2 + \ldots + \beta_5^2 > 0$. Для проверки данной гипотезы используется следующая статистика:

$$F = \frac{RSS}{TSS} \cdot \frac{n-k-1}{k} \Leftrightarrow F = \frac{R^2}{1-R^2} \cdot \frac{n-k-1}{k}.$$

Так как по таблице Фишера $F_{\text{кр}}(0.05; 5; 203) = 2.232$, а $F_{\text{расчетное}} = \frac{0.3006}{1-0.3006} \cdot \frac{203}{5} \approx 17.45 > F_{\text{кр}}$, то гипотеза о незначимости регрессии отвергается.

**3.** Приведем теперь статистический критерий для проверки гипотезы о равенстве нулю нескольких коэффициентов в модели регрессии, т.е. проверим гипотезу о том, что несколько факторов **совместно** не влияют на зависимую переменную $y$. Пусть $q$ – число коэффициентов, значимость которых проверяется. Для определенности рассмотрим случай последних $q$ коэффициентов в модели регрессии:

$$H_0\ :\ \beta_{k-q+1} = \cdots = \beta_k = 0$$

при альтернативе

$$H_1\ :\ \beta^2_{k-q+1} + \cdots + \beta^2_k > 0$$

(не все коэффициенты равны нулю).

Обозначим $R^2_{ur}$ – коэффициент $R^2$ и $\text{RSS}_{ur}$ – остаточная сумма квадратов в модели регрессии без ограничений ($ur$=unrestricted или «длинная регрессия»)

$$\widehat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

Пусть $R^2_r$ – коэффициент $R^2$ и $\text{RSS}_r$ – остаточная сумма квадратов в модели регрессии с ограничениями ($r$ = restricted или «короткая регрессия»), налагаемыми проверяемой нулевой гипотезой, а именно в модели регрессии без учета последних $q$ факторов (факторов $x_{k-q+1}, \ldots, x_k$):

$$\widehat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_{k-q} x_{k-q}.$$

Так как регрессия без ограничений («длинная регрессия») отличается от регрессии с ограничениями («короткой регрессии») включением дополнительных регрессоров, то имеют место неравенства

$$R^2_{ur} \geqslant R^2_r, \quad \text{RSS}_{ur} \leqslant \text{RSS}_r\,.$$

Рассмотрим статистику ($n$ – объем выборки, $m$ – число коэффициентов в регрессии **без ограничений**)

$$F = \frac{R_{ur}^2 - R_r^2}{1 - R_{ur}^2} \cdot \frac{n-m}{q} = \frac{\mathrm{RSS}_r - \mathrm{RSS}_{ur}}{\mathrm{RSS}_{ur}} \cdot \frac{n-m}{q}.$$

При справедливости нулевой гипотезы $F$-статистика имеет распределение Фишера

$$F \underset{H_0}{\sim} F_{q,n-m}.$$

Пусть $F_{\text{кр}} = F(\alpha; q, n-m)$ – критическое значение распределения Фишера $F_{q,n-m}$ при уровне значимости $\alpha$. Имеем следующий статистический критерий проверки гипотезы:

- если $F > F_{\text{кр}}$, то гипотеза $H_0$ отвергается при заданном уровне значимости;
- если $F < F_{\text{кр}}$, то данные согласуются с нулевой гипотезой при заданном уровне значимости.

При больших объемах выборки можно использовать приближенный (асимптотический) статистически критерий проверки нулевой гипотезы. Пусть $\chi^2_{\text{кр}} = \chi^2(\alpha, q)$ – критической значение распределения $\chi^2_q$ с уровнем значимости $\alpha$. Тогда

- при $qF > \chi^2_{\text{кр}}$ гипотеза $H_0$ отвергается в пользу $H_1$,
- при $qF < \chi^2_{\text{кр}}$ данные согласуются с нулевой гипотезой.

**4.** Рассмотрим теперь общий случай проверки гипотезы о линейных соотношениях на коэффициенты регрессии. Пусть $q$ – число линейных ограничений на коэффициенты в линейной регрессии ($q \leqslant m$). Эти ограничения можно записать в матричном виде $R\beta = \mathbf{r}$, где $R$ – некоторая матрица размера $q \times m$, а $\mathbf{r}$ – вектор-столбец $q \times 1$. Приведем статистический критерия для проверки гипотезы

$$H_0 \ : \ R\beta = \mathbf{r}$$

против альтернативы

$$H_1 \ : \ R\beta \neq \mathbf{r}.$$

Обозначим через $R_{ur}^2$ коэффициент $R^2$ и через $\mathrm{RSS}_{ur}$ остаточную сумму квадратов в модели регрессии без ограничений ($ur$ = unrestricted) со всем рассматриваемыми факторами.

Обозначим через $R_r^2$ коэффициент $R^2$ и через $\mathrm{RSS}_r$ – остаточную сумму квадратов в модели регрессии с ограничениями ($r$ = restricted), налагаемыми проверяемой нулевой гипотезой.

Легко видеть, что имеют место неравенства

$$R_{ur}^2 \geqslant R_r^2, \quad \mathrm{RSS}_{ur} \leqslant \mathrm{RSS}_r\,.$$

Таким образом, для вычисления $F$-статистики для проверки сложной нулевой гипотезы необходимо оценить две модели регрессии: с ограничениями, налагаемыми нулевой гипотезой, и без ограничений (со всеми рассматриваемыми регрессорами). Эта $F$-статистика может быть вычислена альтернативным способом **только** по регрессии без ограничений по формуле

$$F = \frac{1}{q}\left(R\widehat{\beta}_{OLS} - \mathbf{r}\right)' \left[R \cdot \widehat{\mathrm{Var}}\left(\widehat{\beta}_{OLS}\right) \cdot R'\right]^{-1} \left(R\widehat{\beta}_{OLS} - \mathbf{r}\right) = \frac{1}{qs^2}\left(R\widehat{\beta}_{OLS} - \mathbf{r}\right)' \left[R \cdot (\mathbf{X}'\mathbf{X})^{-1} \cdot R'\right]^{-1} \left(R\widehat{\beta}_{OLS} - \mathbf{r}\right).$$

*Пример.* Была оценена регрессионная модель зависимости издержек COST энергетических компаний от объемов производства $KWH$ (100 кВт. в ч), географического положения South (бинарная переменная) и от цен на рабочую силу $PL$, топливо $PF$ и капитал $PK$. Зависимая переменная: $\ln(COST)$; объем выборки 23.

| $Y$ | $const$ | $KWH$ | $\ln(PL)$ | $\ln(PF)$ | $\ln(PK)$ | $South$ |
|---|---|---|---|---|---|---|
| $coeff$ | 2.23 | 0.002 | 0.21 | 0.31 | 0.34 | $-0.03$ |
| $Std.Err$ | 1.45 | 0.0008 | 0.07 | 0.13 | 0.07 | 0.0014 |

$$ESS = 250.4, \quad RSS = 230.6$$

Предполагается, что ошибка модели регрессии удовлетворяет условиям Гаусса-Маркова. Уровень значимости 0.05.

а) Напишите спецификацию регрессии с ограничениями для проверки статистической гипотезы $H_0 : \theta_2 = \theta_3 = \theta_4$. Дайте интерпретацию нулевой гипотезе.

б) Была оценена регрессия с ограничениями и ее остаточная сумма квадратов равна 258.2. Тестируйте гипотезу из предыдущего пункта.

*Решение.* а) Гипотеза: $H_0 : \beta_2 = \beta_3 = \beta_4$ означает, что эластичности по каждому фактору равны.

Регрессия с ограничениями имеет вид

$$\ln Y = \beta_0 + \beta_1 KWH + \beta_2(\ln(PL) + \ln(PF) + \ln(PK)) + \beta_5 South + \varepsilon$$

б) Для проверки гипотезы используется статистика: $F = \frac{(R^2_{ur} - R^2_r)q}{(1-R^2_{ur})(n-m)}$, где $q$ – количество ограничений, $m$ – количество параметров, $n$ – количество наблюдений.

$$R^2_{ur} = \frac{ESS_{ur}}{TSS} = \frac{250.4}{250.4 + 230.6} \approx 0.52$$

$$R^2_r = \frac{ESS_r}{TSS} = \frac{TSS - RSS_r}{TSS} = \frac{481 - 258.2}{481} = \frac{222.8}{481} \approx 0.46.$$

Теперь находим значение $F$-статистики $F = \frac{(0.52-0.46)2}{(1-0.52)(23-6)} \approx 0.015$.

По таблице Фишера $F_{\text{кр}}(0.05; 2; 17) \approx 3.592$. Ясно, что гипотеза не отвергается.

## 2.6. Прогнозирование в линейной модели регрессии

Рассмотрим теперь задачу прогнозирования для линейной модели регрессии. Как и в случае парной регрессии, будем различать точечный и интервальный прогноз.

Предположим, что помимо выборочных данных $(\mathbf{x}_i, y_i)_{i=1}^n$ задано также еще одно значение объясняющих переменных $\mathbf{x}_{n+1}$ и известно, что соответствующее значение зависимой переменной удовлетворяет той же линейной модели регрессии (2.1)

$$y_{n+1} = \beta_0 + \beta_1 x_{(n+1)1} + \cdots + \beta_k x_{(n+1)k} + \varepsilon_{n+1} = \mathbf{x}'_{n+1}\beta + \varepsilon_{n+1}$$

и ошибка регрессии удовлетворяет условия 1. – 4. Задача состоит в оценке величины $y_{n+1}$ через $(\mathbf{x}_i, y_i)_{i=1}^n$ и $\mathbf{x}_{n+1}$.

**1.** Как и в случае парной регрессии, рассмотрим сначала простой случай, когда значения параметров регрессии $\beta_0, \ldots, \beta_k$ и $\sigma^2$ известны точно. Тогда естественно в качестве прогноза $\hat{y}_{n+1}$ величины $y_{n+1}$ взять ее математическое ожидание

$$\hat{y}_{n+1} = \mathsf{E}(y_{n+1}) = \beta_0 + \beta_1 x_{(n+1)1} + \cdots + \beta_k x_{(n+1)k} = \mathbf{x}'_{n+1}\beta,$$

среднеквадратическая ошибка этого прогноза равна

$$\mathsf{E}\left(y_{n+1} - \hat{y}_{n+1}\right)^2 = \mathsf{E}\varepsilon_{n+1}^2 = \sigma^2.$$

Так как по предположению

$$y_{n+1} - \hat{y}_{n+1} = \varepsilon_{n+1} \sim \mathcal{N}(0, \sigma^2),$$

то для доверительного интервала с заданной доверительной вероятность $\gamma$ будем иметь следующее выражение

$$\mathsf{P}\left(\hat{y}_{n+1} - \sigma \cdot z_\gamma < y_{n+1} < \hat{y}_{n+1} + \sigma \cdot z_\gamma\right) = \gamma,$$

где $z_\gamma$ есть **двустороннее** критическое значение стандартного нормального распределения с уровнем значимости $(1-\gamma)$ и определяется из уравнения

$$\Phi(z_\gamma) = \frac{1+\gamma}{2}$$

**2.** Однако в прикладных задачах точные значения параметров регрессии как правило не известны и оцениваются по выборочным данным. Тогда естественно в формуле для прогноза заменить неизвестные значения коэффициентов их OLS-оценками:

$$\hat{y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{(n+1)1} + \cdots + \hat{\beta}_k x_{(n+1)k} = \mathbf{x}'_{n+1}\widehat{\beta}_{OLS}.$$

Так как OLS-оценки коэффициентов регрессии являются несмещенными, то

$$\mathsf{E}\hat{y}_{n+1} = \beta_0 + \beta_1 x_{(n+1)1} + \cdots + \beta_k x_{(n+1)k} = \mathsf{E}y_{n+1}.$$

Кроме того, так как OLS-оценки коэффициентов регрессии линейны по $y$, то и прогноз $\hat{y}_{n+1}$ также линеен по $y$. Так же, как в модели парной регрессии, определенное таким образом прогнозное значения является «наилучшим» в смысле среднеквадратичного отклонения.

**Теорема.** *Пусть $\widetilde{y}$ – линейный (по $y_1, \ldots, y_n$) прогноз с условием $\mathsf{E}\widetilde{y}_{n+1} = \mathsf{E}y_{n+1} = \mathbf{x}'_{n+1}\beta$. Тогда*

$$\mathsf{E}\left(\widetilde{y}_{n+1} - y_{n+1}\right)^2 \geqslant \mathsf{E}\left(\hat{y}_{n+1} - y_{n+1}\right)^2.$$

Приведем теперь формулу для доверительного интервала в случае неизвестных параметров модели регрессии. Среднеквадратическая ошибка прогноза равна.

$$\mathsf{E}\left(\hat{y}_{n+1} - y_{n+1}\right)^2 = \sigma^2\left(1 + \mathbf{x}'_{n+1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{n+1}\right)$$

Заменив $\sigma^2$ на ее оценку $s^2$ обозначим

$$\delta = \sqrt{s^2\left(1 + \mathbf{x}'_{n+1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{n+1}\right)}$$

Можно показать, что случайная величина $(\hat{y}_{n+1} - y_{n+1})/\delta$ имеет распределение Стьюдента $t_{n-m}$. Следовательно, доверительный интервал с заданной доверительной вероятностью $\gamma$ задается как

$$\mathsf{P}\left(\hat{y}_{n+1} - \delta \cdot t_\gamma < y_{n+1} < \hat{y}_{n+1} + \delta \cdot t_\gamma\right) = \gamma,$$

где $t_\gamma$ есть **двустороннее** критическое значение распределения Стьюдента $t_{n-m}$ с уровнем значимости $(1-\gamma)$.

Важно отметить, что значения регрессоров $\mathbf{x}_{n+1}$ входят как в выражение для точечного прогноза $\hat{y}_{n+1}$, так и в выражение для $\delta$, характеризующего длину доверительного интервала.

## 2.7. Нелинейные модели

Выше мы предполагали, что зависимость фактора $y$ от регрессоров описывается линейным регрессионным уравнение (2.1). Из условия 1. на ошибки регрессии следует, что

$$\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$$

т.е. среднее значение $y$ линейно зависит от объясняющих переменных. Согласно этой модели регрессии при изменении значения фактора $x_j$ на $\triangle_j$ среднее значение фактора $y$ изменяется на $\beta_j\triangle_j$ и это изменение не зависит от первоначального значения $x_j$. В частности, средний эффект (отклик) от увеличения значения фактора $x_j$ на единицу постоянен и равен $\beta_j$. Другими словами, маржинальные (предельные) величины в линейной модели регрессии постоянны и равны $\beta_j$.

Однако во многих экономических ситуациях имеет место убывание маржинальных (предельных) величин. Для описания таких ситуаций наиболее часто используются два типа регрессионных моделей.

**Полулогафмическая модель регрессии** имеет вид

$$\ln y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i,$$

Относительно ошибок регрессии предполагается, что они удовлетворяют условиям теоремы Гаусса – Маркова. Тогда наилучшими линейными оценками (BLUE-оценками) параметров регрессии являются OLS-оценки и к ним применимы все выводы линейной модели регрессии (статистические свойства коэффициентов и проч.). Далее, так как $\mathsf{E}\varepsilon_i = 0$, то

$$\mathsf{E}(\ln y_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}.$$

Пусть значение фактора $x_j$ изменяется на $\triangle_j$, Тогда ожидаемое изменение величины $\ln y$ равно

$$\triangle \mathsf{E}(\ln y) = \beta_j \triangle_j$$

Это означает (см. случай полулогарифмической парной модели регрессии), что значение фактора $y$ изменится в среднем в $\exp(\beta_j \triangle_j)$ раз.

Если значение $(\beta_j \triangle_j)$ достаточно мало, то воспользовавшись приближением $\exp(\beta_j \triangle_j) \approx 1 + \beta_j \triangle_j$, получаем, что значение фактора $y$ изменяется (в первом приближении) на $(\beta_j \triangle_j) \cdot 100\%$ процентов.

*Пример.* Рассматривается зарплатная модель регрессии

$$\widehat{\ln(wage)} = \beta_0 + \beta_1 ed + \beta_2 age + \beta_3 age^2 + \beta_4 imm,$$

где $ed$ - уровень образования, $age$ - возраст, $imm$ - фиктивная переменная, равная 1 для иммигрантов.

а) Пусть $\widehat{\beta}_4 = -0.11$. Дайте интерпретацию полученной оценки.

б) Пусть $s_4 = 0.09$. Значима ли разница в оплате труда между иммигрантами и местными рабочими? Рассмотрите уровень значимости 1%, 5%, 10%. $n = 2003$.

Другой исследователь оценил следующую модель

$$\widehat{\ln(wage)} = \underset{(0.149)}{1.154} + \underset{(0.003)}{0.0493} ed + \underset{(0.007)}{0.069} age - \underset{(0.00008)}{0.0007} age^2 + \underset{(0.0803)}{0.0963} imm - \underset{(0.0073)}{0.0314} ed \cdot imm.$$

в) Какой ожидаемый эффект от дополнительного года образования для иммигранта?

г) Какой ожидаемый уровень оплаты двадцатилетнего местного рабочего с уровнем образования 10 лет?

д) Как проинтерпретировать коэффициент при $Age^2$?

*Решение.* а) У иммигрантов заработная плата изменяется на $e^{-0.11}$. Можно сказать, что зарплата иммигрантов меньше приблизительно на 11%.

б) $t_{\text{расчетное}} = \frac{|-0.11|}{0.09} = 1.22$, а $t_{\text{кр}}(0.01; 1998) = 2.57$, $t_{\text{кр}}(0.05; 1998) = 1.96$, $t_{\text{кр}}(0.1; 1998) = 1.64$. На всех уровнях гипотезу о незначимости следует принять.

в) Ожидаемый эффект от дополнительного года образования для иммигранта составляет $e^{0.0314}$, то есть вырастает приблизительно на 3.14%.

г) Ожидаемый уровень оплаты двадцатилетнего местного рабочего с уровнем образования 10 лет составляет $ln(wage) = 1.154 + 0.0493 \cdot 10 + 0.069 \cdot 20 - 0.0007 \cdot (20^2) = 2.747$ Откуда $wage = e^{2.747} \approx 15.6$.

д) Так как коэффициент при $Age^2$ отрицателен, то можно говорить об эффекте старения, то есть об уменьшении роста зарплаты с возрастом.

**Лог-линейная модель регрессии** имеет вид

$$\ln y_i = \beta_0 + \beta_1 \ln x_{i1} + \cdots + \beta_k \ln x_{ik} + \varepsilon_i$$

Относительно ошибок регрессии предполагается, что они удовлетворяют условиям теоремы Гаусса – Маркова. Тогда наилучшими линейными оценками (BLUE-оценками) параметров регрессии являются OLS-оценки и к ним применимы все выводы линейной модели регрессии (статистические свойства коэффициентов и проч.). Далее, так как $\mathsf{E}\varepsilon_i = 0$, то

$$\mathsf{E}(\ln y_i) = \beta_0 + \beta_1 \ln x_{i1} + \cdots + \beta_k \ln x_{ik}.$$

Пусть значение фактора $x_j$ изменяется в $p_j$ раз ($p_j > 0$), тогда ожидаемое изменение величины $\ln y$ равно

$$\triangle \mathsf{E}(\ln y) = \beta_j \ln p_j$$

Это означает (см. случай лог-линейной парной модели регрессии), что значение фактора $y$ в среднем изменится в $p_j^{\beta_j}$ раз.

Если $p_j = 1 + r_j$ и $r_j$ достаточно мало, то $p_j^{\beta_j} \approx 1 + r_j\beta_j$, т.е. при изменении регрессора $x_j$ на $r_j \cdot 100\%$ процентов, значение $x$ изменяется (в первом приближении) на $(\beta_j \cdot r_j) \cdot 100\%$ процентов.

В лог-линейной модели регрессии коэффициент $\beta_j$ есть не что иное как коэффициент **эластичности** $y$ по переменной $x_j$. В самом деле,

$$E_{x_j} = \frac{\partial y}{\partial x_j} \cdot \frac{x_j}{y} = \frac{\partial(\ln y)}{\partial(\ln x_j)} = \beta_j$$

*Пример.* В Eviews была оценена производственная функция Кобба-Дугласа:

$$Dependent\ variable : \ln(Q);\ Sample : 1142$$

| $Y$ | $const$ | $\ln(K)$ | $\ln(L)$ | $\ln(H)$ |
|---|---|---|---|---|
| $coeff$ | 3.45 | 0.47 | 0.12 | 0.34 |
| $Std.Err$ | 1.45 | 0.23 | 0.56 | 0.13 |
| $t - Stat$ | | | | |
| $Prob.$ | 0.021 | 0.0461 | 0.8312 | 0.0116 |

а) Дайте интерпретацию коэффициентов модели.

б) Заполните пропуски.

в) Проверьте значимость коэффициентов при уровне значимость 0.05.

г) Запишите оцененное уравнение регрессии.

д) Дайте экономическую интерпретацию гипотезе $H_0 : \theta_1 + \theta_2 + \theta_3 = 1$.

е) Пусть F-статистика для гипотезы из предыдущего пункта равна 1.76. Тестируйте гипотезу. Вычислите p-значение.

*Решение.* а) Все коэффициенты модели - это эластичность. Свободный член модели сообщает о выпуске, при нулевом использовании факторов производства (только совсем не понятно, как этот выпуск получается).

б)

| | | | | |
|---|---|---|---|---|
| $t-Stat$ | 2.37 | 2.04 | 0.21 | 2.6 |

в) Для проверки значимости сравним p-значение с 0.005:

| | | | | |
|---|---|---|---|---|
| $Prob.$ | 0.021 | 0.0461 | 0.8312 | 0.0116 |

Как видно из таблицы все коэффициенты значимы за исключением $\beta_2$.

г) Оцененное уравнение регрессии имеет вид

$$\ln(y) = 3.45 + 0.47\ln(K) + 0.12\ln(L) + 0.34\ln(H).$$

д) Если $\theta_1 + \theta_2 + \theta_3 = 1$, то сумма эластичностей равна 1, что означает постоянную отдачу от масштаба.

е) По таблице Фишера находим p-значение $p-value = 1-F^{-1}(1.76) = 0.06$. Так как $0.05 < p - value < 0.1$, то гипотезу не отвергаем на 5% уровне, но отвергаем на 10% уровне значимости.

**Другие примеры нелинейных моделей** В предыдущих примерах рассматривались модели, в которых $y$ или $\ln y$ линейно зависит от регрессоров $x$ или их логарифмов $\ln x$. Однако в некоторых ситуациях линейной зависимости недостаточно и необходимо рассматривать нелинейную зависимость от объясняющих переменных, но линейную относительно параметров. К таким моделям применимы все выводы множественной модели регрессии, при этом каждое слагаемое должно рассматриваться как отдельный фактор.

На необходимость включения нелинейных членов может указывать анализ графиков зависимости $y$ от регрессоров.

Рассмотрим пример

*Пример* (Wage-equation). Рассмотрим зависимость уровня почасовой оплаты труда $wage$ от возраста $age$. В качестве зависимой переменной естественно рассматривать фактор $\ln(wage)$. Согласно полулогарифмической модели

$$\ln(wage) = \beta_0 + \beta_1 age + \varepsilon$$

с увеличением возраста уровень почасовой оплаты будет расти (если $\hat{\beta}_1 > 0$), что не соответствует реальности: до определенного возраста зарплата будет расти, а потом снижаться. Другими словами, эта полулогарифмическая модель не учитывает старение человека. Чтобы учесть старение индивидуума, введем в модель регрессии фактора $age^2$:

$$\ln(wage) = \beta_0 + \beta_1 age + \beta_2 age^2 + \varepsilon.$$

Естественно ожидать, что после оценки параметров модели коэффициент $\widehat{\beta}_2$ будет отрицательный и значим. В этом случае легко найти возраст, при котором будет (в среднем!) максимальный уровень почасовой оплаты:

$$age_{\max} = -\frac{\widehat{\beta}_1}{2\widehat{\beta}_2}.$$

## 2.8. Стохастические регрессоры

Мы использовали вероятностную модель множественной регрессии, в которой значения объясняющих факторов считались неслучайными (детерминированными). Однако в некоторых приложениях значения регрессоров необходимо считать случайными. Например, в ситуации когда их значения не могут быть получены точно и измерены с некоторыми случайными ошибками. В этом случае статистические выводы должны быть несколько скорректированы.

Рассмотрим следующую вероятностную модель

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u \tag{2.2}$$

где $y, x_1, \ldots, x_k, u$ – случайные величины, причем $y$ и $x_j$ «наблюдаемы», а $u$ «ненаблюдаемо». Случайная величина (ошибка) $u$, как и раньше, описывает влияние факторов, не включенных в модель. Для удобства записи обозначим

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}.$$

Тогда модель регрессии (2.2) можно записать в матричном виде

$$y = \beta_0 + \mathbf{x}'\beta + u.$$

Относительно ошибки $u$ будем предполагать выполнения следующих условий:

1) $\mathsf{E}(u|x_1, \ldots, x_k) = 0$ (в матричном виде $\mathsf{E}(u|\mathbf{x}) = 0$),

2) $\mathsf{E}(u^2|x_1, \ldots, x_k) = \sigma^2$ (в матричном виде $\mathsf{E}(u^2|\mathbf{x}) = \sigma^2$),

3) $u|x_1, \ldots, x_k \sim \mathcal{N}(0, \sigma^2)$ (в матричном виде $u|\mathbf{x} \sim \mathcal{N}(0, \sigma^2)$).

При выполнении условия 1) очевидно

$$\mathsf{E}(y|x_1, \ldots, x_k) = \mathsf{E}(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k = \beta_0 + \mathbf{x}'\beta.$$

**Предложение.** *Если выполнены условия 1) и 2), то*

*a)* $\mathrm{Var}(u|x_1, \ldots, x_k) = \sigma^2$,

*b)* $\mathsf{E}u = 0$ *и* $\mathrm{Var}(u) = \sigma^2$,

*c)* $\mathrm{cov}(x_j, u) = 0$, $j = 1, \ldots, k$.

Рассмотрим задачу оценивания параметров $\beta_0, \ldots, \beta_k$ и $\sigma^2$ на основе выборочных данных. Пусть $\{y_i, x_{i1}, \ldots, x_{ik}\}_{i=1}^n$ – **случайная выборка** факторов. Обозначим

$$\mathbf{x}_i' = \begin{pmatrix} x_{i1} & x_{i2} & \cdots & x_{ik} \end{pmatrix}, \quad i = 1, \ldots, n.$$

Для OLS-оценок коэффициентов модели регрессии верны равенства:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \dots - \hat{\beta}_k \bar{x}_k$$

$$\widehat{\beta}_{OLS} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} = \widehat{\Sigma}_x^{-1} \cdot \begin{pmatrix} \widehat{\mathrm{cov}}(y, x_1) \\ \widehat{\mathrm{cov}}(y, x_2) \\ \vdots \\ \widehat{\mathrm{cov}}(y, x_k) \end{pmatrix},$$

где $\widehat{\Sigma}_x$ – выборочная ковариационная матрица.

Основной результат дается следующей теоремой.

**Теорема** (Гаусс – Марков)**.** *Пусть для линейной модели* (2.2) *выполнены условия 1) и 2),* $\det \Sigma_x \neq 0$ *и* $(y_i, \mathbf{x}_i)$ *– случайная выборка. Тогда OLS-оценки* $\hat{\beta}_0, \dots, \hat{\beta}_k$ *коэффициентов регрессии* $\beta_0, \dots, \beta_k$ *будут линейными несмещенными оценками с минимальной дисперсией*[2]*, т.е. BLUE оценками. Кроме того, эти оценки состоятельны, т.е.*

$$\hat{\beta}_j \xrightarrow[\mathsf{P}]{} \beta_j, \quad j = 0 \dots, k$$

*при* $n \longrightarrow +\infty$

Так же, как в случае детерминированных регрессоров, определяются предсказанные значения $\hat{y}_i$, остатки, полная TSS, объясненная ESS и остаточная RSS суммы квадратов. Для них верно равенство

$$\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS}\,.$$

Коэффициент $R^2$ определяется равенством

$$R^2 = \frac{\mathrm{ESS}}{\mathrm{RSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}.$$

**Теорема.** *При выполнении условий 1), 2) и 3) для OLS-оценок параметров регрессии и для коэффициента детерминации верны все статистические свойства модели регрессии с детерминированными регрессорами.*

*Замечание.* Следует отметить, что статистические свойства для модели со стохастическими регрессорами следует понимать в смысле условных распределений: $t_j | \mathbf{x}_1 \dots, \mathbf{x}_n \sim t_{n-m}$ и т.д.

Более подробно о регрессии со стохастическими регрессорами написано в книгах [1, 2].

## 2.9. Задачи

Рассматривается КЛММР $y_i = \theta_0 + \theta_1 x_{i1} + \ldots + \theta_k x_{ik} + \varepsilon_i$, где $i = 1, \dots, n$, $E(\varepsilon_i) = 0$, $cov(\varepsilon_i, \varepsilon_j) = \sigma^2 \delta_{ij}$, $i, j = 1, \dots, n$.

[2]Имеется в виду условная дисперсия $\mathrm{Var}(\cdot | \mathbf{x}_1, \dots, \mathbf{x}_n)$.

1. В матричном виде

   а) запишите сумму квадратов отклонений $S$;

   б) выписать необходимые условия экстремума $gradS = \vec{0}$ и систему нормальных уравнений;

   в) найти МНК-оценки параметров и доказать их несмещенность;

   г) найти дисперсию МНК-оценок.

2. Доказать, что $R^2 \geq R^2_{adj}$.

3. Почему для проверки значимости регрессии "в целом" недостаточно последовательно проверить значимость каждого коэффициента в отдельности?

4. Покажите, что в регрессии $y_i$ на прогнозные значения $\hat{y}_i$ и константу свободный член равен 0, а угловой коэффициент равен 1 и $R^2 = \widehat{corr}^2(y; \hat{y})$.

5. Рассмотрим регрессионную модель зависимости оценки за финальный экзамен exam от оценки за промежуточный срез midterm и времени, отведенного на выполнение экзаменационного задания time: $exam = \theta_1 midterm + \theta_2 time + error$:
   1)Выведете систему нормальных уравнений для нахождения оценок наименьших квадратов коэффициентов модели.
   2)Найдите формулы для оценок наименьших квадратов коэффициентов модели.

6. (задача 3.1 из [8]) Рассмотрим уравнения

$$\ln y_t = \beta_1 + \beta_2 \ln \omega_t + \beta_3 s_t + \varepsilon_t,$$

$$\ln(y_t/\omega_t) = \gamma_1 + \gamma_2 \ln \omega_t + \gamma_3 s_t + \varepsilon_t, \qquad t = 1, ..., n,$$

   где $y_t$ - годовой доход $t$-го индивидуума, $\omega_t$ - число его рабочих недель в году, $s_t$ - полное число лет, потраченных им на образование.
   а) Покажите, что для соответствующих МНК-оценок выполнены соотношения:
   $\hat{\gamma}_1 = \hat{\beta}_1$, $\hat{\gamma}_3 = \hat{\beta}_3$, $\hat{\gamma}_2 = \hat{\beta}_2 - 1$.
   б) Покажите, что остатки этих регрессий совпадают.

7. Была оценена регрессионная модель зависимости издержек (фактор $COST$) энергетических компаний от объемов производства $KWH$ (100 кВт. в ч), географического положения South (бинарная переменная) и от цен на рабочую силу $PL$, топливо $PF$ и капитал $PK$. Зависимая переменная: $\ln(COST)$; объем выборки 23, уровень значимости 0.05.

| $Y$ | $const$ | $KWH$ | $\ln(PL)$ | $\ln(PF)$ | $\ln(PK)$ | $South$ |
|---|---|---|---|---|---|---|
| $coeff$ | 2.23 | 0.002 | 0.21 | 0.31 | 0.34 | $-0.03$ |
| $Std.Err$ | 1.45 | 0.0008 | 0.07 | 0.13 | 0.07 | 0.0014 |

$ESS = 250.4$; $RSS = 230.6$

a) Дайте интерпретацию коэффициентов модели.
b) Тестируйте значимость влияния объемов производства. Какая альтернатива предпочтительней: односторонняя или двухсторонняя? Ответ поясните.
c) Вычислите p-значение для проверки значимости влияния объемов производства.
d) Значимо ли эластичность издержек по цене на топливо отличается от 0.25?
e) Вычислите коэффициент $R^2$ и дайте его интерпретацию.
f)Проверьте значимость регрессии в целом и сформулируйте проверяемую статистическую гипотезу. Как интерпретируется проверяемая статистическая гипотеза?
g) Вычислите p-значение для проверки значимости регрессии в целом.
h) Напишите спецификацию регрессии с ограничениями для проверки статистической гипотезы $H_0 : \theta_2 = \theta_3 = \theta_4$. Дайте интерпретацию нулевой гипотезе.
i) Была оценена регрессия с ограничениями и ее остаточная сумма квадратов равна 258.2. Тестируйте гипотезу из предыдущего пункта.
j) Вычислите p-значение для тестирования гипотезы из пункта h).

8. Была оценена модель регрессии зависимости стоимости дома *price* (в \$) от площади дома *area* (кв. м), числа комнат *rooms*, площади участка *land* (кв. м), числа ванных комнат *baths* и возраста дома *age*

Зависимая переменная: *price*; объем выборки 321

| | const | *area* | *rooms* | *land* | *baths* | *age* | $age^2$ |
|---|---|---|---|---|---|---|---|
| coeff | 12396.19 | 29.91 | 1793.846 | 0.041 | 6818.556 | -664.38 | 2.36 |
| S. E. | 15138.56 | 3.357 | 2462.598 | 0.043 | 3831.34 | 176.2 | 1.13 |

$R^2 = 0.5407 \quad R^2_{adj} = 0.5319$

(a) Дайте интерпретацию коэффициентов при факторах *area*, *rooms*, *land* и *baths*.

(b) Тестируйте значимость влияния числа ванных комнат. Какая альтернатива предпочтительней: односторонняя или двухсторонняя? Ответ поясните.

(c) Дайте интерпретацию $R^2$.

(d) Тестируйте значимость регрессии «в целом».

9. По четырехнедельным историческим данным была оценена регрессионная модель зависимости спроса на мороженное *consumption* (в пинтах на человека) от среднего недельного дохода домохозяйства *income* (в \$), цены за пинту *price* (в \$) и средней температуры *temp* (по Фаренгейту).

Зависимая переменная: *consumption*; объем выборки 30

| | const | *income* | *price* | *temp* |
|---|---|---|---|---|
| coeff | 0.1973149 | 0.0033078 | −1.044413 | 0.0034584 |
| Std. Err | 0.2702161 | 0.0011714 | 0.834357 | 0.0004455 |

$R^2 = 0.7890 \quad R^2_{adj} = 0.6866$

(a) Дайте интерпретацию коэффициентов при факторах *income*, *price* и *temp*.

(b) Проверьте значимость каждого из коэффициентов. Какие можно сделать выводы?

(c) Дайте интерпретацию $R^2$.

(d) Тестируйте значимость регрессии «в целом».

10. Была оценена модель регрессии зависимости продолжительности недельного сна *sleep* (в мин) от недельной занятости *totwrk* (в мин), возраста *age*, уровня образования *educ*, почасовой оплаты труда *hrwage*, числа детей *yngkid* в возраста меньше 3 лет и бинарной переменной *male* (гендерный фактор)

Зависимая переменная: *sleep*; объем выборки 532

| | *c* | *totwrk* | *male* | *age* | *educ* | *hrwage* | *yngkid* |
|---|---|---|---|---|---|---|---|
| coeff | 3593.091 | -0.157 | 42.53 | 1.61 | -10.59 | 0.13 | 67.43 |
| S. E. | 139.41 | 0.02 | 41.65 | 1.74 | 7.14 | 5.42 | 54.31 |

$R^2 = 0.1154 \quad R^2_{adj} = 0.1053$

(a) Дайте интерпретацию всех коэффициентов.

(b) Тестируйте значимость влияния числа детей. Какая альтернатива предпочтительней: односторонняя или двухсторонняя? Ответ поясните.

(c) Проверьте значимость влияния гендерного фактора. Какие можно сделать выводы?

(d) Дайте интерпретацию $R^2$.

(e) Тестируйте значимость регрессии «в целом».

11. Была оценена регрессионная модель зависимости издержек (фактор $COST$) энергетических компаний от объемов производства $KWH$ (100 кВт. в ч), географического положения $South$ (бинарная переменная) и от цен на рабочую силу $PL$, топливо $PF$ и капитал $PK$.

Зависимая переменная: $\ln(COST)$; объем выборки 23 (уровень значимости 5%).

| | const | $KWH$ | $\ln(PL)$ | $\ln(PF)$ | $\ln(PK)$ | $South$ |
|---|---|---|---|---|---|---|
| coefficient | 2.3 | 0.002 | 0.21 | 0.31 | 0.34 | −0.03 |
| Std. Error | 1.45 | 0.0008 | 0.07 | 0.13 | 0.07 | 0.0014 |

ESS = 250.4 RSS = 230.6

a) Дайте интерпретацию коэффициентов модели.

b) Тестируйте значимость влияния объемов производства. Какая альтернатива предпочтительней: односторонняя или двухсторонняя? Ответ поясните.

c) Вычислите $p$-значение для проверки значимости влияния объемов производства.

d) Значимо ли эластичность издержек по цене на топливо отличается от 0.25?

e) Вычислите коэффициент $R^2$ и дайте его интерпретацию.

f) Проверьте значимость регрессии «в целом» и сформулируйте проверяемую статистическую гипотезу. Как интерпретируется проверяемая статистическая гипотеза?

g) Вычислите $p$-значение для проверки значимости регрессии «в целом».

h) Напишите спецификацию регрессии с ограничениями для проверки статистической гипотезы $H_0 : \beta_2 = \beta_3 = \beta_4$. Дайте интерпретацию нулевой гипотезе.

i) Была оценена регрессия с ограничениями и ее остаточная сумма квадратов равна 258.2. Тестируйте гипотезу из предыдущего пункта.

j) Вычислите $p$-значение для тестирования гипотезы из пункта 11h).

12. Была оценена регрессионная модель зависимости цены дома $Price$ (в \$1000) от его площади $hsize$ (в м$^2$), площади участка вокруг дома $lsize$ (в м$^2$), числа ванных комнат $bath$, числа спален $bdr$ и бинарной переменной $poor$ (равна единице, если состояние дома оценивается как плохое).

Зависимая переменная: $Price$; объем выборки 33 (уровень значимости 5%).

| | const | $hsize$ | $lsize$ | $bath$ | $bdr$ | $poor$ | |
|---|---|---|---|---|---|---|---|
| coefficient | 121.3 | 0.13 | 0.001 | 12.1 | 0.42 | −50.7 | ESS = 350.4 |
| Std. Error | 12.2 | 0.012 | 0.00076 | 4.13 | 1.3 | 19.7 | RSS = 533.2 |

a) Дайте интерпретацию коэффициентов модели.

b) Тестируйте значимость влияния числа спален. Какая альтернатива предпочтительней: односторонняя или двухсторонняя? Ответ поясните.

c) Вычислите $p$-значение для проверки значимости влияния числа спален.

d) Значимо ли предельный эффект от изменения площади дома отличается от 0.2?

e) Вычислите коэффициент $R^2$ и дайте его интерпретацию.

f) Проверьте значимость регрессии «в целом» и сформулируйте проверяемую статистическую гипотезу. Как интерпретируется проверяемая статистическая гипотеза?

g) Вычислите $p$-значение для проверки значимости регрессии «в целом».

h) Напишите спецификацию регрессии с ограничениями для проверки статистической гипотезы $H_0 : \beta_3 = \beta_4$. Дайте интерпретацию нулевой гипотезе.

i) Была оценена регрессия с ограничениями и ее остаточная сумма квадратов равна 560.3. Тестируйте гипотезу из предыдущего пункта.

j) Вычислите $p$-значение для тестирования гипотезы из пункта 12h).

13. На основе опроса 25–35 летних была оценена регрессионная модель зависимости почасовой оплаты $Wage$ от уровня образования $educ$, возраста $age$, членства в профсоюзе $union$ (бинарная переменная), пола $female$ (бинарная переменная) и семейного статуса $married$ (бинарная переменная)

Зависимая переменная: $\ln(Wage)$; объем выборки 26

| | const | $educ$ | $age$ | $union$ | $female$ | $married$ |
|---|---|---|---|---|---|---|
| coefficient | 1.82 | 0.031 | 0.018 | 0.039 | −0.04 | −0.03 |
| Std. Error | 0.49 | 0.011 | 0.006 | 0.013 | 0.001 | 0.04 |

$R^2 = 0.58$

a) Дайте интерпретацию коэффициентов модели.

b) Тестируйте значимость влияния возраста. Какая альтернатива предпочтительней: односторонняя или двухсторонняя? Ответ поясните.

c) Вычислите $p$-значение для проверки значимости влияния возраста.

d) Значимо ли отдача от образования отличается от 0.05?

e) Дайте интерпретацию коэффициента $R^2$.

f) Проверьте значимость регрессии «в целом» и сформулируйте проверяемую статистическую гипотезу. Как интерпретируется проверяемая статистическая гипотеза?

g) Вычислите $p$-значение для проверки значимости регрессии «в целом».

h) Напишите спецификацию регрессии с ограничениями для проверки статистической гипотезы $H_0 : \beta_4 = \beta_5 = 0$. Дайте интерпретацию нулевой гипотезе.

i) Была оценена регрессия с ограничениями и для нее коэффициент $R^2$ равен 0.045. Тестируйте гипотезу из предыдущего пункта.

j) Вычислите $p$-значение для тестирования гипотезы из пункта 13h).

14. (задача 3.4 из [8]) Регрессия зависимой переменной $y$ на три независимые переменные на основе $n = 30$ наблюдений дала следующие результаты:

| $y$ | $= 25.1$ | $+ 1.2x_1$ | $+ 1.0x_2$ | $- 0.50x_3$ |
|---|---|---|---|---|
| Стандартные ошибки | (2.1) | (1.5) | (1.3) | (0.060) |
| $t$ – значения | (11.9) | ( ) | ( ) | ( ) |
| 95% – доверительные границы | $\pm(4.3)$ | ( ) | ( ) | ( ) |

а) Заполните пропуски.
б) Истинны или ложны следующие утверждения (если ложны, исправьте их):
1) Оценка коэффициента при $x_1$ есть 1.2. Другие исследователи могут собрать другие данные и построить другие оценки этого коэффициента. Распределение этих оценок сосредоточено вокруг истинного значения 1.2. Поэтому оценка называется несмещенной.
2) Если есть априорная уверенность в том, что $x_1$ не влияет на $y$, то представляется разумным отвергнуть нулевую гипотезу $\mathbf{H_0} : \beta_1 = 0$ на 5%-уровне значимости.
3) Если есть априорная уверенность в том, что $x_2$ влияет на $y$, то представляется более разумным использовать оценку 1.0, чем принимать гипотезу $\mathbf{H_0} : \beta_2 = 0$.

15. (задача 3.5 из [8]) Бюджетное обследование пяти случайно выбранных семей дало следующие результаты (в тыс. руб.):

| Семья | Накопления, $S$ | Доход, $Y$ | Имущество, $W$ |
|---|---|---|---|
| 1 | 3.0 | 40 | 60 |
| 2 | 6.0 | 55 | 36 |
| 3 | 5.0 | 45 | 36 |
| 4 | 3.5 | 30 | 15 |
| 5 | 1.5 | 30 | 90 |

а) Оцените регрессию $S$ на $Y$ и $W$.
б) Спрогнозируйте накопления семьи, имеющей доход 4- тыс. руб. и имущество стоимостью 25 тыс.руб.
в) Предположим, что доход семьи возрос на 10 тыс.руб., в то время как стоимость имущества не изменилась. Оцените, как возрастут ее накопления.
г) Оцените, как возрастут накопления семьи, если ее доход вырос на 5 т.р., а стоимость имущества увеличилась на 15 т. р.
д) Найдите сумму квадратов остатков и постройте оценку дисперсии регрессии.

16. (задача 3.6 из [8]) Рассмотрим регрессию $s_t = \beta_1 + \beta_2 y_t + \beta_3 \omega_t + \varepsilon_t$ из предыдущей задачи.
а) Постройте 95%- доверительное множество для $\beta_2$ и $\beta_3$; $\beta_2$; $\beta_3$; $\beta_1$ и $\beta_2$.
б) Проверьте с 5%-уровнем значимости следующие гипотезы:
   (a) $\beta_2 = 0$ и $\beta_3 = 0$;
   (b) $\beta_3 = 0$ (стоимость имущества несущественна);
   (c) $\beta_2 = 0$ (величина дохода несущественна);
   (d) $\beta_2 = 1$ (таким мог быть ответ вашего коллеги на вопрос о зависимости накопления от дохода);

(e) $\beta_2 = 1.57$ (такое значение коэффициента $\beta_2$ могло быть с высокой степенью надежности установлено для другой страны, и вас интересует вопрос, верно ли это для вашей страны).

(f) $\beta_2 = -5\beta_3$ (т.е. эффект дохода противоположен эффекту богатства в фиксированной пропорции).

в) Пусть некоторая семья имеет доход $Y = 30$ тыс.руб. и имущество стоимостью $W = 52.5$ тыс.руб. Чему равна прогнозная величина ее накоплений? В каком смысле эта семья может рассматриваться как средняя между семьями 4 и 5? Почему прогнозная величина ее накоплений не есть среднее между 3.5 и 1.5 тыс.руб.? Постройте 95%-доверительный интервал для прогнозной величины накоплений этой семьи.

17. *

Для предприятий отрасли имеются данные $(Q_i, L_i, K_i)$, где $Q_i$ - объем выпуска продукции, $L_i$ - трудовые ресурсы, $K_i$ - капитал предприятия, $i = 1, \ldots, n$. Воспользовавшись прологарифмированной моделью производственной функции Кобба-Дугласа $\ln Q_i = \theta_0 + \theta_1 \ln L_i + \theta_2 \ln K_i + \varepsilon_i$, $(E(\varepsilon_i) = 0;\ cov(\varepsilon_i; \varepsilon_j) = \sigma^2 \delta_{ij},\ i, j = 1, \ldots, n)$:

1) определить конкретное выражение для матрицы $R$ и вектор-столбца $r$, участвующих в общей форме $R\Theta = r$ записи гипотезы $H_0 : \theta_1 = \theta_2$ и описать процедуру статистической проверки;

2) выписать явный вид выражения оценок через исходные данные, в случае справедливости гипотезы $H_0$;

3) выписать явный вид оценки эластичности выпуска по труду.

18. *

С целью исследования зависимости заработной платы от ряда факторов проводилось обследование работающих, в ходе которого регистрировались значения: $y_i$ (тыс. руб.) – месячный заработок $i$-го респондента, $x_i^{(1)}$ (годы) – продолжительность его обучения, $x_i^{(2)}$ (баллы) – результат его профессионального тестирования (в 100-балльной шкале), $x_i^{(3)}$ (равен 1, если респондент мужского пола, и нулю – в противном случае), $x_i^{(4)}$ (равен 1, если респондент женат/замужем, и нулю – в противном случае).

По результатам $n = 125$ обследованных респондентов с помощью МНК были оценены два варианта регрессионных зависимостей

(1) $y_i = \theta_0(1) + \theta_1(1)x_i^{(1)} + \theta_2(1)x_i^{(2)} + \varepsilon_i$

(2) $y_i = \theta_0(2) + \theta_1(2)x_i^{(1)} + \theta_2(2)x_i^{(2)} + \theta_3(2)x_i^{(3)} + \theta_4(2)x_i^{(4)} + \varepsilon_i$

(в обоих вариантах остатки $\varepsilon_i$ полагались нормальными и удовлетворяющими требованиям классической линейной модели). Результаты оценивания

приведены в таблице ($j$ - номер варианта модели):

| $j$ | $\hat{\theta}_0(j)$ | $\hat{\theta}_1(j)$ | $\hat{\theta}_2(j)$ | $\hat{\theta}_3(j)$ | $\hat{\theta}_4(j)$ | $\hat{R}^2(j)$ |
|---|---|---|---|---|---|---|
| 1 | $6,20\ (0,92)$ | $2,25\ (0,15)$ | $0,36\ (0,07)$ | – | – | $0,21$ |
| 2 | $4,63\ (0,58)$ | $1,85\ (0,14)$ | $0,29\ (0,06)$ | $6,22\ (0,7)$ | $2,12\ (0,85)$ | $0,36$ |

(в скобках справа от значений $\hat{\theta}_l(j)$ приведены значения стандартных среднеквадратических ошибок их оценивания). Требуется:

1) сделать выводы о статистической значимости (незначимости) включения в модель каждой из анализируемых переменных;

2) используя оценки по моделям (1) и (2), проверить гипотезу о совместной статистической незначимости влияния на заработную плату пола респондента и его семейного статуса (при уровне значимости $\alpha = 0,05$);

3) используя оценки по модели (1), проверить гипотезу о совместной статистической незначимости влияния на заработную плату респондента длительности обучения и результатов профессионального теста (при уровне значимости $\alpha = 0,05$).

4) Проинтерпретировать значение оценки $\hat{\theta}_3(2)$.

5) Выписать уравнение зависимости месячного заработка для незамужних женщин.

6) Построить интервальную оценку для коэффициента $\theta_2(1)$ в модели (1) при доверительной вероятности $P_0 = 0,95$.

19. *

По результатам $(x_i^{(1)}, x_i^{(2)}, y_i)$, – $i = 1, 2, \ldots, 58$, – обследования 58-ми однотипных предприятий проведен статистический анализ производственной функции Кобба-Дугласа

$$y_i = \theta_0 (x_i^{(1)})^{\theta_1} (x_i^{(2)})^{\theta_2} e^{\varepsilon_i},$$

описывающей зависимость объема выпуска продукции ($y_i$) от основных факторов производства: капитала ($x_i^{(1)}$) и труда ($x_i^{(2)}$), в предположении постоянства отдачи от масштаба производства, т.е. при гипотетичном условии

$$\theta_1 + \theta_2 = 1 \qquad (*)$$

(случайные остатки $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{58}$ предполагаются статистически независимыми и одинаково $(0, \sigma^2)$-нормально распределенными).

Применение МНК и прологарифмированной модели

$$\ln y_i = \theta_0' + \theta_1 \ln x_i^{(1)} + \theta_2 \ln x_i^{(2)} + \varepsilon_i, \quad i = 1, 2, \ldots, 58 \qquad (1)$$

($\theta_0' = \ln \theta_0$) с учетом ограничения $(*)$ дало следующие результаты

$$\hat{\theta}_0' = 3,21;\ \hat{\theta}_1 = 0,63;\ \hat{\theta}_2 = 0,37;\ \hat{\sigma}^2 = 1,82;\ \hat{R}^2 = 0,71.$$

Требуется:

1) Доказать, что при соблюдении условия $(*)$ увеличение в $k$ раз основных факторов производства приводит к увеличению среднего выпуска в то же самое количество раз (при вычислении из (1) $E(y|x^{(1)}, x^{(2)})$ воспользоваться приблизительным равенством $e^{\varepsilon} = 1 + \varepsilon$).

2) Определить конкретные выражения для матрицы $R$ и вектора $r$, участвующих в общей форме $R\Theta = r$ записи ограничения $(*)$, и описать процедуру проверки гипотезы $H_0 : \theta_1 + \theta_2 = 1$.

3) Выписать формулы для вычисления оценок коэффициентов $\theta_0'$, $\theta_1$, $\theta_2$ в уравнении (1) с учетом ограничения $(*)$; требуемые формулы получить с помощью замены в (1) коэффициента $\theta_2$ его выражением через $\theta_1$ (с учетом ограничения $(*)$).

4) Проанализировать эластичности объема выпуска по труду $Э_{y/x^{(1)}}$ и по капиталу $Э_{y/x^{(2)}}$. С целью получения выводов о приоритетных направлениях инвестирования в производство.

5) Проинтерпретировать значение коэффициента детерминации $\widehat{R}^2$.

# Глава 3

# Разные аспекты линейной регрессии

## 3.1. Фиктивные переменные

При рассмотрении линейных регрессионных моделей не накладывалось никаких ограничений на возможные значения, принимаемые регрессорами. Так, включенные в модель регрессоры могут принимать как непрерывные, так дискретные значения. Но стоит отметить, что при нормальной распределенности ошибок регрессии зависимая переменная должна принимать непрерывные значения.

Часто возникает необходимость учитывать в регрессионной модели факторы, носящие не количественный, а качественный характер. Например, учесть различие в уровне оплаты труда между мужчинами и женщинами. Для учета таких факторов вводятся *фиктивные* (бинарные) *переменные* (dummy variable), принимающие два значения:

- 1, если качественный признак присутствует;
- 0, если качественный признак отсутствует.

*Пример* (Wage–equation). Рассмотрим модель регрессии, описывающую зависимость почасовой оплаты труда $wage$ от стажа $age$:

$$\widehat{\ln(wage)} = \beta_0 + \beta_1 age + \beta_2 age^2.$$

Для исследования влияния пола человека на уровень оплаты введем фиктивную переменную

$$male = \begin{cases} 1, & \text{для мужчин} \\ 0, & \text{для женщин} \end{cases}$$

и изменим спецификацию модели регрессии

$$\widehat{\ln(wage)} = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 male.$$

Тогда средний уровень оплаты труда для женщины равен

$$\widehat{\ln(wage)} = \beta_0 + \beta_1 age + \beta_2 age^2,$$

а для мужчины равен

$$\widehat{\ln(wage)} = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3.$$

Таким образом, коэффициент $\beta_3$ представляет собой (среднюю) разницу в оплате труда между мужчинами и женщинами при прочих равных условиях. Проверка значимости коэффициента $\beta_3$ означает проверку гипотезы $H_0 : \beta_3 = 0$ об отсутствии разницы в оплате труде между мужчинами и женщинами.

*Замечание.* Можно было бы включить в модель бинарную переменную $male$, принимающую любые два значения (не обязательно 0 и 1). Однако в этом случае коэффициент $\beta_3$ не имел бы простой интерпретации.

Фиктивные переменные могут быть включены в модель регрессии разными способами, что позволяет исследовать разные тонкие эффекты влияния качественных факторов на зависимую переменную.

*Пример* (Wage–equation). Рассмотрим модель регрессии, описывающую влияние высшего образование (бинарная переменная $hedu$, равная 1 при наличии высшего образования) на почасовую оплату труда (переменная $wage$):

$$\widehat{\ln wage} = \beta_0 + \beta_1 hedu.$$

Тогда коэффициент $\beta_0$ есть (средний) уровень оплаты человека без высшего образования, а коэффициент $\beta_1$ показывает (среднее) увеличение оплаты труда при наличии высшего образования (естественно ожидать, что $\hat{\beta}_1 > 0$). Эта модель не учитывает разницу в уровне оплаты труде между мужчинами и женщинами.

Как в предыдущем примере введем фиктивную переменную $male$, равную единице для мужчин, и рассмотрим регрессионную модель

$$\widehat{\ln wage} = \beta_0 + \beta_1 hedu + \beta_2 male.$$

Согласно этой модели (средняя) разница в уровне оплаты труда между мужчинами и женщинами с одинаковым уровнем образования постоянна и равна $\beta_2$. Однако наличие высшего образование дает **одинаковое** (среднее) увеличение уровня почасовой оплаты как для мужчин, так и для женщин на величину $\beta_1$, т.е. эффект от наличия высшего образования постоянен и не зависит от пола.

Рассмотрим теперь следующую модель

$$\widehat{\ln wage} = \beta_0 + \beta_1 hedu + \beta_2 male + \beta_3 (hedu \cdot male)$$

Согласно этой модели (средняя) разница в уровне оплаты труда между мужчинами и женщинами без высшего образования равна $\beta_2$. При наличии высшего образования уровень оплаты женщины увеличивается (в среднем) на величину $\beta_1$, а для мужчины – на величину $\beta_1 + \beta_3$. Следовательно, в этой модели и базовый уровень оплаты, и эффект от наличия высшего образования зависят от пола. В рамках этой модели проверка значимости коэффициента $\beta_3$ означает, что проверятся гипотеза об отсутствии разницы в увеличении оплаты труда между мужчинами и женщинами при наличии высшего образования.

В некоторых задачах качественный признак может иметь несколько «значений» или «градаций». В этом случае удобно ввести несколько фиктивных переменных: если качественный фактор имеет $l$ градаций, то в модель включаются $l-1$ фиктивная переменная, отвечающая каким-либо $l-1$ уровням фактора. Такая ситуация возникает, например, если учитывать в модели регрессии качественный фактор сезонности. Удобно рассмотреть это на примере.

*Пример.* Рассмотрим регрессионную модель зависимости объема продаж зонтов от цены

$$\ln(Q) = \beta_0 + \beta_1 \ln(P) + \text{error}\,.$$

Естественно полагать, что объем продаж зонтов имеет выраженную сезонность, т.е. в модель необходимо включить (качественный) фактор сезонности. Так как сезонов четыре, то включим в модель три фиктивные переменные, например: $Spr$ (весна), $Summ$ (лето) и $Fall$ (осень). Рассмотрим уточненную модель

$$\ln(Q) = \beta_0 + \beta_1 \ln(P) + \beta_2 Spr + \beta_3 Summ + \beta_4 Fall + \text{error}$$

Тогда зимой зависимость между объемом продаж и ценой описывается уравнением $\widehat{\ln(Q)} = \beta_0 + \beta_1 \ln P$, а весной – уравнением $\widehat{\ln(Q)} = \beta_0 + \beta_1 \ln P + \beta_2$. Следовательно, весной объем продаж изменяется (по отношению к зиме) на величину коэффициента $\beta_2$. Аналогичный смысл имеют коэффициенты $\beta_3$ и $\beta_4$.

*Замечание.* В предыдущем примере статистическая гипотеза

$$H_0 : \beta_2 = \beta_4$$

означает, что уровень продаж весной и осенью, по сравнению с зимним периодом, изменяется на одну и ту же величину.

*Замечание.* В предыдущем примере выбор зимы как «базового» сезона, изменения по отношению к которому показывали коэффициенты при фиктивных переменных, был полностью произволен. Можно было бы выбрать **любые** три сезон, ввести соответствующие фиктивные переменные и включить их в модель регрессии. Выбор «базового» сезона может быть продиктован какими-то экономическими соображениями или удобством. С точки зрения статистических свойств все эти модели одинаковы.

*Замечание.* Если качественный фактор имеет $l$ «значений» (градаций) и ввести в модель $l$ фиктивных переменных, то очевидно, что сумма этих переменных будет равна 1. Если в модель включена константа, то определитель матрицы системы нормальных уравнений будет равен нулю и система нормальных уравнений будет иметь **бесконечное число решений**, т.е. OLS-оценки коэффициентов регрессии определены неоднозначно.

С помощью фиктивных переменных можно моделировать кусочно - линейные регрессионные модели, применяемые, в частности, для описания структурных изменений.

*Пример.* Пусть на основе временных рядов оценивается зависимость между двумя факторами

$$y_t = \beta_0 + \beta_1 x_t + \text{error}, \quad t = 1, \ldots, n$$

Если в момент времени $t_0$ $(1 < t_0 < n)$ произошло событие, влияющее на экономическую конъюнктуру и, возможно, на структурные изменения в экономике, то как отразить его влияние в модели регрессии? Определим фиктивную переменную

$$d_t = \begin{cases} 0, & t = 1, \ldots, t_0 - 1 \\ 1, & t = t_0, \ldots, n \end{cases}$$

и введем ее в модель регрессии:

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 d_t + \beta_3 (x_t \cdot d_t) + \text{error}\,.$$

Тогда до наступления события зависимость между факторами описывается соотношением

$$\hat{y}_t = \beta_0 + \beta_1 x_t \quad t = 1, \ldots, t_0 - 1,$$

а после наступления события зависимость будет иметь вид

$$\hat{y}_t = \beta_0 + \beta_2 + (\beta_1 + \beta_3) x_t \quad t = t_0, \ldots, n.$$

Коэффициенты $\beta_2$ и $\beta_3$ показывают изменения, соответственно, константы и коэффициента наклона прямой после наступления события, влияющего на структурные изменения количественной зависимости между факторами.

*Пример.* Была оценена регрессионная модель зависимости издержек COST энергетических компаний от объемов производства $KWH$ (100 кВт. в ч), географического положения South (бинарная переменная) и от цен на рабочую силу $PL$, топливо $PF$ и капитал $PK$. Зависимая переменная: $\ln(COST)$; объем выборки 23.

| $Y$ | $const$ | $KWH$ | $\ln(PL)$ | $\ln(PF)$ | $\ln(PK)$ | $South$ |
|---|---|---|---|---|---|---|
| $coeff$ | 2.23 | 0.002 | 0.21 | 0.31 | 0.34 | $-0.03$ |
| $Std.Err$ | 1.45 | 0.0008 | 0.07 | 0.13 | 0.07 | 0.0014 |

$$ESS = 250.4, \quad RSS = 230.6$$

Предполагается, что ошибка модели регрессии удовлетворяет условиям Гаусса-Маркова. Уровень значимости 0.05.

a) Дайте интерпретацию коэффициента при бинарной переменной $South$ и протестируйте её значимость.

b) Значимо ли эластичность издержек по цене на топливо отличается от 0.25?

c) Проверьте значимость регрессии в целом и сформулируйте проверяемую статистическую гипотезу. Вычислите p-значение.

*Решение.* a) Оценка коэффициента при переменной $South$ равна $-0.03$. Так как модель для этого фактора является полулогарфмической, то если $South = 1$ (предприятие находится на юге), издержки падают на 3% по сравнению со случаем $South = 0$.

Этот коэффициент значим при любом разумном уровне значимости, потому что t-статистика равна $t = \frac{-0.03}{0.0014} \approx 21.43$.

b) Гипотеза о том, что эластичность издержек по цене на топливо не отличается от 0.25 выглядит так $H_0 : \beta_3 = 0.25$ . Альтернативная гипотеза $H_1 : \beta_3 \neq 0.25$.

Для проверки гипотезы требуется вычислить значение $t$-статистика:

$$t = \frac{|\widehat{\beta}_3 - \beta_3|}{s(\widehat{\beta}_3)} = \frac{|0.31 - 0.25|}{0.13} \approx 0.46.$$

Теперь сравним данное значение с критическим значением: $t(\alpha, n-k-1) = t(0.05, 23-5-1) = t(0.05, 17) = 2.11$. Так как $t_{\text{кр}} > t_{\text{расчетное}}$, поэтому основная гипотеза не отвергается.

c) Вычислим коэффициент

$$R^2 = \frac{ESS}{TSS} = \frac{ESS}{ESS + RSS} = \frac{250.4}{250.4 + 230.6} = 0.52$$

Основная гипотеза о том, что регрессия в целом не значима имеет вид $H_0 : \beta_1 = \ldots = \beta_5 = 0$. Альтернативная гипотеза - гипотеза значима, $H_1 : \beta_1^2 + \ldots \beta_5^2 \neq 0$.

Для проверки данной гипотезы используется следующая статистика:

$$F = \frac{RSS}{TSS} \cdot \frac{n-k-1}{k} \Leftrightarrow F = \frac{R^2}{1-R^2} \cdot \frac{n-k-1}{k} = \frac{0.52}{1-0.52} \cdot \frac{23-6-1}{6} = 2.8$$

Сравним с критическим значением $F(m-1; n-m) = F(5; 17) = 2.21$ $F_{\text{кр}} < F_{\text{расчетное}}$, поэтому гипотеза о незначимости регрессии отвергается. $p - value = 1 - F^{-1}(2.21) = 1 - 0.97 = 0.03$.

*Пример.* На основе квартальных данных с 2001 по 2008 год было получено следующее уравнение регрессии, описывающее зависимость цены $y_t$ на товар от нескольких факторов:

$$\widehat{y}_t = 9.5 + \underset{(0.01)}{0.2}\, x_t + \underset{(0.002)}{0.6}\, w_t + \underset{(0.02)}{0.7}\, z_t; \quad \text{ESS} = 62.5 \quad \text{RSS} = 44.7.$$

Когда в уравнение были добавлены фиктивные переменные, соответствующие первым трем кварталам года, величина ESS выросла до 76.4. Напишите спецификацию уравнения регрессии с учетом сезонности. Вычислите скорректированный коэффициент детерминации для обеих моделей регрессии. Улучшило ли включение сезонности качество модели?

*Решение.* Новая модель имеет вид $\hat{y}_t = \beta_0 + \beta_1 x_t + \beta_2 w_t + \beta_3 z_t + \beta_4 \cdot d_1 + \beta_5 \cdot d_2 + \beta_6 \cdot d_3$, где $d_i$ - это бинарная переменная, принимающая значения 1, если наблюдение соответствует сезону $i$, и 0 - в противном случае. Обратите внимание, что принадлежность к четвёртому сезону определяется равенством $d_1 = d_2 = d_3 = 0$.

Для тестирования улучшения необходимо проверить гипотезу: $H_0 = \beta_4 = \beta_5 = \beta_6 = 0$, вычислим статистику: $F = \frac{(ESS_R - ESS_{UR})/q}{ESS_{UR}/(n-m)} \sim F(\alpha; q; n-m)$, $q$-количество фиктивных переменных, $m$-количество параметров, $n$-количество наблюдений.

$$F = \frac{(ESS_R - ESS_{UR})/q}{ESS_{UR}/(n-m)} = \frac{(76.4 - 62.5)/3}{76.4/(32-7)} \approx 1.516.$$

Так как критическое значение равно $F_{\text{кр}}(\alpha; q; n-m) = F_{\text{кр}}(0.05; 3; 25) \approx 2.991$, то гипотеза об отсутствии сезонности не отвергается.

## 3.2. Тест Чоу

Тест Чоу (Chow's test) используется для проверки **однородности** двух выборок, а именно проверяется нулевая гипотеза, что две выборки описываются одним и тем же уравнением регрессии.

Пусть имеется две выборки факторов $y, x_1, \ldots, x_k$ объемом $n_1$ и $n_2$ и в каждой зависимость $y$ от регрессоров описывается уравнением регрессии

$$\begin{aligned} y_i &= \mathbf{x}_i'\beta + \varepsilon_i, & \mathrm{Var}(\varepsilon_i) &= \sigma_1^2 & i &= 1 \ldots n_1, \\ y_i &= \mathbf{x}_i'\gamma + \nu_i, & \mathrm{Var}(\nu_i) &= \sigma_2^2 & i &= n_1 + 1 \ldots n_1 + n_2. \end{aligned}$$

Пусть $\mathrm{RSS}_1$ и $\mathrm{RSS}_2$ – остаточные суммы квадратов в модели регрессии, оцененной по первой и второй выборкам соответственно. Обозначим через RSS остаточную сумму квадратов в модели регрессии, оцененной по **объединенной** выборке объема $n_1 + n_2$.

Статистический критерий проверки (сложной) нулевой гипотезы об однородности выборок

$$H_0 \ : \ \beta = \gamma, \ \sigma_1^2 = \sigma_2^2$$

основан на $F$-статистике

$$F = \frac{\mathrm{RSS} - (\mathrm{RSS}_1 + \mathrm{RSS}_2)}{\mathrm{RSS}_1 + \mathrm{RSS}_2} \cdot \frac{n_1 + n_2 - 2m}{m}$$

При справедливости нулевой гипотезы

$$F \underset{H_0}{\sim} F_{m, n_1+n_2-2m}$$

Следовательно, при заданном уровне значимости $\alpha$ нулевая гипотеза отвергается при $F > F_{\text{кр}}$, где $F_{\text{кр}} = F(\alpha; m, n_1 + n_2 - 2m)$

При больших объемах выборки можно использовать приближенный (асимптотический) статистически критерий проверки нулевой гипотезы. Пусть $\chi^2_{\text{кр}} = \chi^2(\alpha, q)$ – критической значение распределения $\chi^2_q$ с уровнем значимости $\alpha$, тогда

- при $qF > \chi^2_{\text{кр}}$ гипотеза $H_0$ отвергается в пользу $H_1$,
- при $qF < \chi^2_{\text{кр}}$ данные согласуются с нулевой гипотезой.

*Пример.* Рассмотрим модель зависимости логарифма зарплаты $\ln(Wage)$ от уровня образования $edu$, возраста $age$, $age^2$, места жительства $city$ (бинарная переменная, 1 если вырос в городе) и пола $male$

$$\widehat{\ln(wage)} = \beta_0 + \beta_1 edu + \beta_2 age + \beta_3 age^2 + \beta_4 male + \beta_5 city + \beta_6 (city \cdot male),$$

Модель регрессии была отдельно оценена по выборкам из 35 иммигрантов и из 23 граждан страны и были получены остаточные суммы квадратов $RSS_{imm} = 40.2$ и $RSS_{native} = 60.2$. Остаточная сумма квадратов в регрессии, оцененной по объединенной выборке, равна 140.6. Тестируйте гипотезу об отсутствии дискриминации в оплате труда между иммигрантами и гражданами страны. Уровень значимости 5

*Решение.* Так как $RSS = 140.6$, $RSS_1 = 40.2$, $RSS_2 = 60.2$, $n_1 = 35$, $n_2 = 23$, $m =$, то можем применить тест Чоу для проверки гипотезы

$$H_0 \ : \ \beta = \gamma, \ \sigma_1^2 = \sigma_2^2$$

Вычисли значение $F$-статистике

$$F = \frac{RSS - (RSS_1 + RSS_2)}{RSS_1 + RSS_2} \cdot \frac{n_1 + n_2 - 2m}{m} = \\ = \frac{140.6 - (40.2 + 60.2)}{40.2 + 60.2} \cdot \frac{35 + 23 - 14}{7} \approx 887 \quad (3.1)$$

Находим критическое значение

$$F_{\text{кр}} = F_{\alpha;m;n_1+n_2-2m} = F_{0.05;7;44} \approx 2.249$$

Следовательно, при уровне значимости 5% гипотеза отвергается, то есть подтверждается дискриминация.

## 3.3. Мультиколлинеарность

Как уже отмечалось, система нормальных уравнений для модели регрессии имеет единственное решение при $\det(\mathbf{X}'\mathbf{X}) \neq 0$ или $\det \widehat{\Sigma}_x \neq 0$.

Если $\det(\mathbf{X}'\mathbf{X}) = 0$ (или $\det \widehat{\Sigma}_x = 0$), то говорят, что между регрессорами модели есть *чистая мультиколлинеарность*. Это означает, что в выборке один из регрессоров **линейно выражается** через остальные и в этом случае OLS-оценки коэффициентов определены неоднозначно.

Однако в прикладных задачах явление чистой мультиколлинеарности может встречаться крайне редко. Чаще возникает ситуация, когда один из регрессоров «хорошо линейно приближается» остальными регрессорами. Такое явление называется *мультиколлинеарностью*. Рассмотрим, как это **качественно** влияет на оценки коэффициентов регрессии.

Можно доказать следующее

**Предложение.** *Для дисперсии OLS-оценки коэффициентов регрессии верно равенство*

$$\operatorname{Var}\left(\widehat{\beta}_j\right) = \frac{\sigma^2}{(1 - R_j^2)\,\mathrm{TSS}_j} \quad (j = 1, \ldots, k), \quad (3.2)$$

*где*

$$\mathrm{TSS}_j = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$

*есть общая вариация фактора $x_j$, а коэффициент $R_j^2$ вычислен в линейной регрессии фактора $x_j$ на константу и остальные регрессоры модели.*

*Замечание.* Другими словами, коэффициент $R_j^2$ показывает, «насколько хорошо» фактор $x_j$ линейно приближается остальными регрессорами.

**Следствие.** *Для выборочной дисперсии оценок коэффициентов регрессии верно равенство*

$$s_j^2 = \widehat{\mathrm{Var}}\left(\widehat{\beta}_j\right) = \frac{s^2}{(1 - R_j^2)\,\mathrm{TSS}_j} \quad (j = 1, \ldots, k)$$

Таким образом, «большие» значения $R_j^2$ (т.е. «близкие» к единице) приводят (при прочих равных) к «большой» выборочной стандартной ошибке коэффициента $\beta_j$ и, следовательно, к «большому» доверительному интервалу. Таким образом, «точность» оценки коэффициента снижается, и мы можем сделать вывод о незначимости коэффициента, хотя из экономических соображений фактор должен влиять на зависимую переменную. Более того, может даже случиться, что **все** коэффициенты в модели незначимы, но регрессия «в целом» значима. Кроме того, из-за «низкой» точности мы можем получить «неправильный» знак OLS-оценки коэффициента регрессии.

*Замечание.* Из формулы (3.2) видно, что к «большим» стандартным ошибкам коэффициента также приводят и «малые» значения $\mathrm{TSS}_j$.

*Пример.* Рассмотрим двухфакторную модель

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u.$$

Тогда коэффициент $R_1^2$ вычисляется в регрессии

$$x_1 = \gamma_0 + \gamma_1 x_2 + \mathrm{error},$$

а коэффициент $R_2^2$ – в регрессии

$$x_2 = \delta_0 + \delta_1 x_1 + \mathrm{error}$$

Из соотношений для парной модели регрессии получаем

$$R_1^2 = R_2^2 = \widehat{\mathrm{corr}}^2(x_1, x_2).$$

Следовательно, чем «сильнее» коррелируют регрессоры (т.е. чем «ближе» по абсолютной величине к единице коэффициент $\widehat{\mathrm{corr}}(x_1, x_2)$), тем больше выборочные стандартные ошибки коэффициентов (при прочих равных условиях).

Важно понимать, что явление мультиколлинеарности носит **не количественный, а качественный характер**. В самом деле, в каком смысле следует понимать «коэффициент $R_j^2$ близок к единице» или «большая стандартная ошибка коэффициента»? Как следствие, когда следует считать, что в модели регрессии мы имеем «проблему мультиколлинеарности»? Среди эконометристов нет единого мнения на этот счет. Некоторые эконометрические пакеты для каждого фактора

вычисляют показатель $VIF_j = 1/(1 - R_j^2)$ (Variance Inflation Factor), и некоторые эконометристы предлагают считать, что в модели регрессии «присутствует проблема мультиколлинеарности для фактора $x_j$», если показатель $VIF_j > 20$ (см. [23], Гл. 4.9.1 и ссылки). В книге [31] Гл. 3.4 указывается на пороговое значение 10 для показателя $VIF_j$. Эти пороговые значения для $VIF_j$ предлагаются, исходя из эмпирического анализа выборочных данных.

Частым признаком мультиколлиенарности является незначимость всех коэффициентов по отдельности, при этом вся модель "в целом" значима.

Другой подход к мультиколлинеарности основан на состоятельности оценок коэффициентов регрессии: так как выборочные дисперсии коэффициентов (а следовательно, и стандартные ошибки) по вероятности стремятся к нулю, то с ростом объема выборки мы получаем «более точные» оценки коэффициентов. В связи с этим некоторые эконометристы предлагают рассматривать мультиколлинеарность как «проблему малых выборок» (micronumerosity), т.е. причиной «большой» стандартной ошибки коэффициента следует считать не «большое» значение $R_j^2$ (или $VIF_j$), а малый объем выборки (см. [31])

Таким образом, явление и проблема мультиколлинеарности не имеет строго общепризнанного определения и трактовки. Как следствие, существует различные подходы к устранению мультиколлинеарности. Приведем некоторые их них:

1. По возможности включать в модель «слабокоррелированные» между собой влияющие переменные (но в первую очередь при отборе влияющих переменных руководствоваться экономической целесообразностью и экономической теорией!).

2. Если есть возможность, то увеличить объем выборки (что, конечно же, не всегда представляется возможным).

3. Попытаться исключить из модели фактор, который «подозревается» в причине мультиколлинеарности. Однако к исключению факторов надо подходить крайне осторожно, т.к. исключение существенного фактора приводит к смещению оценок коэффициентов (omitted variable problem, см. следующую главу).

4. Изменить спецификацию модели.

5. Оставить модель регрессии «как есть» (например, если экономическая теория говорит в пользу именно такой модели регрессии). Ведь МНК-оценки будут иметь минимальную дисперсию среди линейных несмещённых оценок и интервальные прогнозы от мультиколлинеарности никак не страдают.

6. Использовать метод главных компонент (см. [1]).

7. Можно пожертвовать несмещённостью, чтобы уменьшить дисперсию коэффициентов. Например, использовать ридж-регрессию или гребневую регрессию, в которой при поиске оценок минимизируется такая величина $\sum(y_i - \hat{y}_i)^2 + \lambda \sum \hat{\beta}_j^2$. В матричном виде оценка имеет вид $\widehat{\beta} = (\mathbf{X}'\mathbf{X} + \lambda D)^{-1}(\mathbf{X}'\mathbf{y})$

для вектора коэффициентов регрессионного уравнения $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, где $D$ – диагональная матрица, состоящая из диагональных элементов матрицы $\mathbf{X}'\mathbf{X}$.

*Замечание.* При оценке регрессий с совершенной мультиколлинеарностью в компьютерных программах, будет выдано сообщение об ошибке или автоматически удалена одна из переменных.

## 3.4. Спецификация модели регрессии

При изучении вероятностных и статистических свойств OLS-оценок коэффициентов регрессии мы предполагали, что нам **точно** известна вероятностная модель, описывающая зависимость между факторами (известна **точная** спецификация модели).

Однако в прикладных задачах точная спецификация модели регрессии, как правило, неизвестна и выборочные данные мы пытаемся «подогнать» под подходящую регрессионную модель.

Рассмотрим, как влияют ошибки спецификации модели регрессии на вероятностные и статистические свойства OLS-оценок коэффициентов модели. Для определенности будем рассматривать модель регрессии со стохастическими регрессорами.

### 3.4.1. Невключение в модель значимого фактора

Для простоты изложения рассмотрим проблему невключения значимого фактора на следующем примере: пусть «истинная модель» (true model) описывается уравнение

$$y = \beta_0 + \beta_1 x + \beta_2 z + u$$

и ошибка регрессии $u$ удовлетворяет стандартным условиям регрессионной модели.

Однако исследователю неизвестна «истинная модель», и он оценивает «укороченную» модель регрессии без учета фактора $z$:

$$y = \beta_0 + \beta_1 x + w,$$

ошибка которой равна $w = \beta_2 z + u$. Рассмотрим, как это повлияет на вероятностные и статистические свойства оценок коэффициентов регрессии. OLS-оценка коэффициента $\beta_1$ в «укороченной регрессии» равна

$$\hat{\beta}_1 = \frac{\widehat{\mathrm{cov}}(x, y)}{\widehat{\mathrm{Var}}(x)}.$$

По теореме Слуцкого при $n \to +\infty$

$$\hat{\beta}_1 \xrightarrow[\mathsf{P}]{} \frac{\mathrm{cov}(x, y)}{\mathrm{Var}(x)}.$$

Согласно «истинной» модели регрессии и свойствам ошибок регрессии ($\operatorname{cov}(x, u) = 0$)

$$\operatorname{cov}(y, x) = \operatorname{cov}(\beta_0 + \beta_1 x + \beta_2 z + u, x) = \\ \beta_1 \operatorname{cov}(x, x) + \beta_2 \operatorname{cov}(z, x) + \operatorname{cov}(u, x) = \beta_1 \operatorname{Var}(x) + \beta_2 \operatorname{cov}(z, x).$$

Следовательно,

$$\hat{\beta}_1 \xrightarrow[\mathsf{P}]{} \beta_1 + \beta_2 \frac{\operatorname{cov}(z, x)}{\operatorname{Var}(x)}.$$

Таким образом, если $\beta_2 \operatorname{cov}(z, x) \neq 0$, то OLS-оценка коэффициента $\beta_1$ в «недооцененной» модели **несостоятельная** и асимптотически **смещена**. Можно также показать, что выборочная стандартная ошибка коэффициента $\beta_1$, вычисленная на основе «недооцененной» модели регрессии, меньше выборочной стандартной ошибки, вычисленной на основе «истинной» модели регрессии.

Аналогичный результат верен и для OLS-оценки коэффициента $\beta_0$ в «укороченной» регрессии.

**Теорема** (Omitted variable problem)**.** *Пусть модель зависимости $y$ от факторов $x$ и $z$ описывается уравнением*

$$y = \beta_0 + \beta_1 x + \beta_2 z + u, \tag{3.3}$$

*где ошибка $u$ удовлетворяет всем условиям теоремы Гаусса-Маркова. Пусть $\hat{\beta}_0$ и $\hat{\beta}_1$ – OLS-оценки коэффициентов $\beta_0$ и $\beta_1$, полученные при рассмотрении «укороченой» модели*

$$y = \beta_0 + \beta_1 x + w. \tag{3.4}$$

*Тогда*

1. *если $\beta_2 \operatorname{cov}(z, x) \neq 0$, то оценки $\hat{\beta}_0$ и $\hat{\beta}_1$* ***несостоятельны*** *и* ***асимптотически смещены****.*

2. *Выборочные стандартные ошибки коэффициентов $\beta_0$ и $\beta_1$, вычисленные в рамках модели* (3.4)*,* ***меньше*** *выборочных стандартных ошибок тех же коэффициентов, но вычисленных в рамках «истинной» модели* (3.3)*.*

3. *$\mathsf{E}s^2 > \sigma^2$, т.е. $s^2$ – смещенная оценка дисперсии ошибок регрессии* (3.3)*.*

**Следствие.** *При невключении в модель регрессии влияющей переменной, т.е. при оценивании «укороченной» модели* (3.4)*, $t$- и $F$- статистики для проверки статистических гипотез уже не имеют распределений Стьюдента и Фишера. Следовательно, в это случае статистические выводы, основанные на их применении, неверны.*

*Замечание.* Теорема и ее следствие верны для произвольной модели регрессии, в том числе и с неслучайными факторами, и произвольного числа невключенных факторов (подробнее см. [7, 23]).

### 3.4.2. Включение в модель незначимого фактора

Рассмотрим теперь обратную ситуацию. Пусть «истинная» модель зависимости (true model) описывается уравнением

$$y = \beta_0 + \beta_1 x + u,$$

и ошибка $u$ удовлетворяет стандартным условиям регрессионной модели. Тогда $\beta_1 = \operatorname{cov}(y, x)/\operatorname{Var}(x)$.

Однако исследователю неизвестна «истинная» модель, и он оценивает «переопределенную» («расширенную») модель

$$y = \gamma_0 + \gamma_1 x + \gamma_2 z + w, \tag{3.5}$$

включая в нее «лишний» (irrelevant) фактор $z$, **некоррелирующий** с $x$ и $y$ (т.е. $\operatorname{cov}(x, z) = \operatorname{cov}(y, z) = 0$). Тогда

$$\gamma_1 = \frac{\operatorname{cov}(y, x)\operatorname{Var}(z) - \operatorname{cov}(x, z)\operatorname{cov}(y, z)}{\operatorname{Var}(x)\operatorname{Var}(z) - \operatorname{cov}^2(x, z)} = \frac{\operatorname{cov}(y, x)}{\operatorname{Var}(x)} = \beta_1$$
$$\gamma_2 = \frac{\operatorname{cov}(y, z)\operatorname{Var}(x) - \operatorname{cov}(x, z)\operatorname{cov}(y, x)}{\operatorname{Var}(x)\operatorname{Var}(z) - \operatorname{cov}^2(x, z)} = 0.$$

и $w = u$, т.е. ошибки в модели (3.5) также удовлетворяет стандартным условиям регрессионной модели.

По теореме Гаусса-Маркова OLS-оценки $\hat{\gamma}_1$ и $\hat{\gamma}_2$ будут линейными, несмещенными и состоятельными[1] оценками коэффициентов $\gamma_1 = \beta_1$ и $\gamma_2 = 0$ в модели (3.5). В случае нормальной распределенности ошибки регрессии к оценкам $\hat{\gamma}_1$ и $\hat{\gamma}_2$ применимы все статистические выводы стандартной модели регрессии.

Но можно показать, что стандартная ошибка OLS-оценки $\hat{\beta}_1$ (в «истинной» модели) **меньше** стандартной ошибки OLS-оценки $\hat{\gamma}_1$ в «расширенной» модели (3.5), т.е. оценка $\hat{\gamma}_1$ не будет наилучшей. Как следствие, доверительный интервал для коэффициента $\beta_1$ в «истинной» модели будет у́же, чем доверительный интервал для коэффициента $\gamma_1 = \beta_1$ в модели (3.5) с включенной невлияющей («лишней») переменной $z$. Другими словами, оценка $\hat{\gamma}_1$ «менее точная» (имеет бо́льшую дисперсию), чем оценка $\hat{\beta}_1$.

В общем случаем верна следующая теорема

**Теорема** (Overspecification problem)**.** *При включении в модель факторов, не коррелирующих с объясняющими и с зависимой переменными, OLS-оценки коэффициентов остаются линейными, несмещенными и состоятельными, но не наилучшими (не с минимальной дисперсией). В случаем нормальной распределенности ошибок к OLS-оценкам в «расширенной» модели применимы все статистические выводы стандартной модели регрессии.*

*Замечание.* При включении в модель незначимых объясняющих переменных OLS-оценки становятся «менее точными», т.к. возрастает стандартная ошибка коэффициентов. Это соответствует общему принципу математической статистики: чем

[1] Так как рассматриваем модель со стохастическими регрессорами.

больше коэффициентов в вероятностной модели нужно оценить, тем «хуже» (при одинаковом объеме выборки) статистические свойства выборочных коэффициентов.

### 3.4.3. Сравнение вложенных моделей

Рассмотрим вопрос о сравнении двух вложенных моделей регрессии (nested models)

$$(A) : y = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + u$$
$$(B) : y = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \sum_{j=1}^{l} \theta_j z_j + w.$$

Вложенность означает, то модель $(B)$ отличается от модели $(A)$ включением дополнительных влияющих переменных $z_1, \ldots, z_l$.

Так как всегда $R_A^2 \leqslant R_B^2$, то критерий сравнения моделей на основе коэффициента $R^2$ не является содержательным.

Для выбора между этими моделями приведем три критерия:

a) Критерий скорректированного коэффициента $R^2$: выбирается та модель, для которой показатель $\bar{R}^2$ больше.

b) Тестировать значимость совместного влияния на зависимую переменную факторов $z_1, \ldots, z_l$, т.е. в модели $(B)$ тестировать статистическую гипотезу

$$H_0 \ : \ \theta_1 = \cdots = \theta_l = 0.$$

Если гипотеза отвергается, то выбор в пользу модели $(B)$. Если данные согласуются с нулевой гипотезой, то выбираем модель $(A)$.

c) Сравнение моделей на основе информационных критериев Akaike или Shwarz.

### 3.4.4. Информационные критерии

Как уже отмечалось, при добавлении в модель регрессии новых объясняющих переменных коэффициент $R^2$, показывающий «качество подгонки» модели, не убывает (а практически всегда возрастает). Поэтому сравнивать разные модели регрессии по критерию $R^2$ некорректно. Для сравнения моделей регрессии вводился скорректированный (на число коэффициентов) коэффициент детерминации $\bar{R}^2$: он в некотором смысле вводил «штрафы» за включение в модель дополнительных влияющих переменных. Однако некоторые эконометристы полагают этот штраф недостаточно «большим» ([23], Гл. 3.5.1).

Наряду с показателем $\bar{R}^2$ для сравнения регрессионных моделей с **одинаковыми зависимыми переменными** (и оцененных по одним и тем же выборочным

данным!) применяются т.н. *информационные критерии*, связанные с методом максимального правдоподобия оценки параметров модели регрессии. Наиболее часто используются информационный критерий Акаике (предложен H. Akaike) и байесовский критерий (предложен Шварцом, G. Shwarz, и иногда называется критерий Шварца).

Важно понимать, что сравнение регрессионных моделей как по скорректированному коэффициенту $\bar{R}^2$, так и по информационным критериям не есть тестирование статистических гипотез. В них не участвует ни уровень значимости, ни критические значения подходящих распределений. Процедура сравнения и выбора модели проста: разные модели регрессии сравниваются по той или иной формальной числовой характеристике.

**Информационный критерий Акаике** Для линейной модели регрессии вычисляется показатель $AIC$ (Akaike Information Criteria)

$$AIC = \ln\left(\frac{\mathrm{RSS}}{n}\right) + \frac{2m}{n}.$$

Выбор делается в пользу модель с **минимальным** показателем AIC.

Для небольших выборок иногда предлагают использовать скорректированный показатель

$$AIC_c = AIC + \frac{2m(m+1)}{n-m-1}.$$

*Замечание.* В общем случае для произвольной вероятностной модели показатель $AIC$ вычисляется как

$$AIC = 2m - \ln(L),$$

где $L$ – максимальное значение функции правдоподобия для оцененной модели, а $m$ – число параметров модели.

**Байесовский критерий (критерий Шварца)** Для линейной модели регрессии вычисляется показатель $BIC$ (Bayesian Information Criteria)

$$BIC = \ln\left(\frac{\mathrm{RSS}}{n}\right) + \frac{m\ln n}{n}.$$

Выбор делается в пользу модель с **минимальным** показателем $BIC$.

Из определения видно, что байесовский критерий дает больший «штраф» за включение в модель дополнительных факторов, чем показатель $AIC$.

*Замечание.* Иногда для байесовского информационного критерия используют обозначение $SIC$ (Shwarz Information Criteria).

### 3.4.5. Сравнение невложенных моделей

Рассмотрим вопрос о сравнении двух невложенных моделей регрессии (non-nested models) с одинаковыми зависимыми переменными. Приведем процедуру

для тестирования модели

$$(A):\ y = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + u$$

против альтернативной модели

$$(B):\ y = \gamma_0 + \sum_{j=1}^{l} \gamma_j z_j + w,$$

причем в моделях могут быть совпадающие регрессоры.

**Критерий $\bar{R}^2$** Выбирается та модель, для которой скорректированный коэффициент $R^2$ больше.

**Тест Davidson – MacKinnon [11, 12, 22]** Пусть $\widehat{y}_i^{(B)}$ – предсказанные значения в модели $(B)$. В рамках вспомогательной регрессии

$$y = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \theta_B \cdot \widehat{y}_i^{(B)} + \text{error}$$

проверяется значимость коэффициента $\theta_B$. Если коэффициент незначим, то данные согласованы с моделью $(A)$.

Наоборот, пусть $\widehat{y}_i^{(A)}$ – предсказанные значения в модели $(A)$. Если во вспомогательной регрессии

$$y = \gamma_0 + \sum_{j=1}^{l} \gamma_j z_j + \theta_A \cdot \widehat{y}_i^{(A)} + \text{error}$$

коэффициент $\theta_A$ незначим, то данные согласованы с моделью $(B)$.

*Замечание.* Тест Davidson – MacKinnon имеет следующие недостатки:

1. И в первом, и во втором случаях данные могут быть согласованы с моделями $(A)$ и $(B)$ соответственно.

2. И в первом и во втором случаях тест отвергает модели $(A)$ и $(B)$ соответственно.

В первом случае модели можно сравнить по $\bar{R}^2$ или на основе информационных критериев (Akaike или Shwarz).

## 3.5. Выбор функциональной формы зависимости

Рассмотрим проблему выбора функционального вида модели регрессии при заданном наборе объясняющих переменных. При первичном анализе данных часто полезен визуальный анализ графиков рассеяния $y$ vs объясняющие переменные.

### 3.5.1. Тест на функциональную форму

Для тестирования гипотезы о **линейной спецификации** модели регрессии (на необходимость включить в модель нелинейные слагаемые)

$$H_0 \ : \ y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

часто используется RESET-тест (Regression Equation Specification Error Test [26])

Пусть $\hat{y}_i$ – подогнанные значения в линейной модели регрессии, задаваемой нулевой гипотезой. Во вспомогательной модели регрессии (число $M \geqslant 2$ – параметр теста)

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \theta_2 \widehat{y}^2 + \cdots + \theta_q \widehat{y}^M + \text{error}$$

тестируется гипотеза

$$H_0' \ : \ \theta_2 = \cdots = \theta_M = 0.$$

Исходная нулевая гипотеза отвергается если отвергается гипотеза $H_0'$.

*Замечание.* Обычно RESET-тест применяется при небольших значениях $M = 2, 3, 4$ (зависит от объемы выборки). Предпочтительней использовать значения $M = 3, 4$ [31].

*Замечание.* RESET-тест может отвергать нулевую гипотезу о линейной спецификации в случае невключения в линейную модель значимого регрессора.

*Пример.* Рассмотрим регрессионную модель зависимости зарплаты Wage от уровня школьного образования Educ (в годах), возраста Age, наличия высшего образования Heduc (фиктивная переменная), школьного образования родителей Feduc, Meduc и пола Male:

$$\widehat{\ln(wage)} = \beta_0 + \beta_1 Educ + \beta_2 Age + \beta_3 Heduc + \beta_4 Feduc + \beta_5 Meduc + \beta_6 Male,$$

$$n = 500, \quad R^2 = 0.302$$

a) Пусть $\hat{\beta}_3 = 0.34$. Дайте интерпретацию полученной оценке.

b) Пусть $\hat{\beta}_6 = 0.02$ и $s_6 = 0.002$. Дайте интерпретацию полученной оценке. Значима ли дискриминация в оплате труде между мужчинами и женщинами на 1-процентном уровне значимости?

c) Для тестирования справедливости линейной спецификации обычно используют RESET-тест. Напишите спецификацию вспомогательной регрессии для применения RESET-теста (с включением вспомогательных переменных до 4-й степени). Сформулируйте проверяемую статистическую гипотезу.

d) Пусть во вспомогательной регрессии $R^2 = 0.361$. Тестируйте нулевую гипотезу из пункта c) при уровне значимости 0.01. Какой можно сделать вывод?

e) Можно ли доверять выводам, сделанным в пп. a) и b)?

*Решение.* a) При наличии высшего образования заработная плата увеличится на $e^{0.34}$.

b) Оценка $\hat{\beta}_6 = 0.02$ говорит о том, что мужчины в среднем получают приблизительно на 2% больше, чем женщины. Чтобы проверить значима ли дискриминация

в оплате, вычислим значение статистики $t = \frac{\hat{\beta}_6}{s_6} = \frac{0.02}{0.002} = 10$ и сравним с критическим значением $t_{\text{кр}} = 2.58$. Так как $t_{\text{кр}} < t_{\text{расчетное}}$, то коэффициент является значимым.

c) Вспомогательная регрессия для применения RESET-теста (с включением вспомогательных переменных до 4-й степени) имеет вид

$\widehat{\ln(wage)} = \beta_0 + \beta_1 Educ + \beta_2 Age + \beta_3 Heduc + \beta_4 Feduc + \beta_5 Meduc + \beta_6 Male + \theta_2 \widehat{y}^2 + \theta_3 \widehat{y}^3 + \theta_4 \widehat{y}^4$.

И тестируется гипотеза

$$H_0' \ : \ \theta_2 = \theta_3 = \theta_4 = 0.$$

d) Протестируем нулевую гипотезу с помощью статистики

$$F = \frac{R_{ur}^2 - R_r^2}{1 - R_{ur}^2} \cdot \frac{n-m}{q}.$$

Так как во вспомогательной регрессии $R_{ur}^2 = 0.361$, а $R_r^2 = 0.302$, $q = 3$, $n = 500$, $m = 10$, то

$$F = \frac{0.361 - 0.302}{1 - 0.361} \cdot \frac{490}{3} \approx 15.$$

Так как $F_{\text{кр}}(0.01; 3; 490) \approx 3.821$, то гипотеза о линейной функциональной форме отвергается.

e) Выводы для пунктов a) и b) не надёжны.

### $y$ **или** $\ln y$ **?**

Основная модели регрессии – линейная – характеризуется линейной зависимостью $\mathsf{E}y_i$ от значений объясняющих переменных и постоянством маржинальных значений. Коэффициент $\beta_j$ есть средний эффект от увеличения на единицу значения регрессора $x_j$. Выбор в пользу этой модели можно сделать, если из экономических соображения можно ожидать постоянство маржинальных значений и диаграммы рассеяния $y$ vs $x_j$ показывают «в среднем» линейную зависимость $y$ от объясняющих переменных. Кроме того модель должна давать содержательный экономический прогноз. Например, во многих экономических задачах значения объясняющей переменной должны быть положительными.

*Пример.* Рассмотрим модели зависимость цены подержанного автомобиля $Price$ от его возраста $Age$. Так как (в среднем) цена автомобиля с возрастом уменьшается, то при оценке парной линейной модели регрессии

$$Price = \beta_0 + \beta_i Age + \varepsilon$$

мы получим $\hat{\beta}_1 < 0$ (обратная линейная зависимость). Но тогда для достаточно старых автомобилей мы можем получить предсказанные значения $\widehat{Price}_i < 0$ (при $Age_i >> 1$). Кроме того, естественно ожидать, что эффект от увеличения возраста машины на один год будет непостоянен: $+1$ год для однолетней машины и для 10-ти летней будет давать (в среднем) разное уменьшение цены. Таким образом, предпочтительней использовать другую функциональную форму для описания зависимости цена автомобиля от его возраста.

В случае когда линейная модель из каких-либо соображения не подходит для описания зависимости (отрицательные предсказанные значения, непостоянство маржинальных значений или другое), то исправить ситуацию может помочь переход к логарифму зависимой переменной и рассмотрению полулогарифмических и лог-линейных моделей. В этих моделях предсказанные значения зависимой переменной будут положительными (т.к. по модели получаем значения $\widehat{\ln y_i}$ и $\widehat{y_i} = \exp\left(\widehat{\ln y_i}\right) > 0$).

Лог-линейная модель подходит для описания зависимости с постоянной эластичностью по объясняющим переменным. Напомним: это означает, что эффект от увеличения значения объясняющей переменной $x_j$ на 1% постоянен и приводит к изменению (в среднем) значения зависимой переменной на $\beta_j \cdot 100\%$.

Полулогарифмическая модель подходит когда можно ожидать, что при увеличении значения влияющей переменной $x_j$ на единицу значение зависимой переменной увеличится в $\exp(\beta_j)$ раз.

Существуют формальные тесты для сравнения линейной, полулогарифмической и лог-линейной моделей.

**1.** Так как полулогарифмическая

$$\ln y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

и лог-линейная модели

$$\ln y = \beta_0 + \beta_1 \ln x_1 + \cdots + \beta_k \ln x_k + w$$

являются **невложенными моделями** с одинаковой зависимой переменной, то для их сравнения можно применять критерий $\bar{R}^2$, тест Davidson – MacKinnon, информационные критерии.

**2.** Для сравнения линейной

$$(Linear) \ : \ y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

и лог-линейной

$$(Loglin) \ : \ \ln y = \beta_0 + \beta_1 \ln x_1 + \cdots + \beta_k \ln x_k + w$$

моделей тест Davidson – MacKinnon, $\bar{R}^2$ и информационные критерии **неприменимы**, так как в этих моделях разные зависимые переменные. Приведем описание формального теста для сравнения этих моделей.

**PE-тест** (MacKinnon, White, Davidson) Пусть $\hat{y}_i$ и $\widehat{\ln y_i}$ – предсказанные значения в моделях $(Linear)$ и $(Loglin)$ соответственно. Во вспомогательной регрессии

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \theta_{Lin} \cdot \left(\ln(\hat{y}) - \widehat{\ln y}\right) + \text{error}$$

проверяется значимость коэффициента $\theta_{Lin}$. Если коэффициент незначим, то данные согласованы с моделью $(Linear)$.

Наоборот, во вспомогательной регрессии

$$\ln y = \beta_0 + \beta_1 \ln x_1 + \cdots + \beta_k \ln x_k + \theta_{Log} \cdot \left(\hat{y} - \exp\left(\widehat{\ln y}\right)\right) + \text{error}$$

проверяется значимость коэффициента $\theta_{Log}$. Если коэффициент незначим, то данные согласованы с моделью $(LogLin)$.

## 3.6. Задачи

1. На основе квартальных данных с 2003 по 2008 год было получено следующее уравнение регрессии, описывающее зависимость цены на товар $y_t$ от нескольких факторов:

$$\hat{y}_t = 3.5 + \underset{(0.001)}{0.4}\, x_t + \underset{(0.01)}{1.1}\, w_t; \quad ESS = 70.4; \quad RSS = 40.5$$

   Когда в уравнение были добавлены фиктивные переменные, соответствующие первым трем кварталам года, величина $ESS$ выросла до 86.4. Напишите спецификацию уравнения регрессии с учетом сезонности. Сформулируйте и проверьте гипотезу о наличии сезонности (уровень значимости 5%).

2. Рассмотрим модель регрессии зависимость логарифма зарплаты $\ln(Wage)$ от уровня образования $edu$, возраста $age$, $age^2$, места жительства $city$ (бинарная переменная, 1 если вырос в городе) и пола $male$

$$\widehat{\ln(wage)} = \theta_0 + \theta_1 edu + \theta_2 age + \theta_3 age^2 + \theta_4 male + \theta_5 city + \theta_6 (city \cdot male),$$

   Модель регрессии была отдельно оценена по выборкам из 35 иммигрантов и из 23 граждан страны и были получены остаточные суммы квадратов $RSS_{imm} = 40.2$ и $RSS_{native} = 60.2$. Остаточная сумма квадратов в регрессии, оцененной по объединенной выборке, равна 140.6. Тестируйте гипотезу об отсутствии дискриминации в оплате труда между иммигрантами и гражданами страны. Уровень значимости 5%.

3. Рассмотрим модель зависимости почасовой оплаты труда Wage от уровня школьного образования $educ$, пола $male$ (бинарная переменная), места жительства $smsa$ (бинарная переменная, равная 1 для мегаполиса) и бинарной переменной $imm$ (1, если иммигрант)

$$Wage = \underset{(1.2)}{11.4} + \underset{(0.0006)}{0.002}\, educ + \underset{(0.17)}{0.4}\, male + \underset{(0.78)}{0.6}\, (male \cdot smsa) - \underset{(1.04)}{2.3}\, imm + \underset{(1.34)}{0.078} smsa$$

   Какая средняя разность в уровне почасовой оплаты между мужчинами и женщинами в мегаполисе? В малом городе? Какая средняя разность в уровне оплаты между мужчинами, живущих в мегаполисе и в малом городе? Для женщин? Значима ли разность в уровне оплаты между женщинами, живущими в мегаполисе и в малом городе? Уровень значимости 5%, а объём выборки $n = 33$.

4. Рассмотрим функцию спроса с сезонными переменными $Spr$ (весна), $Summ$ (лето) и $Fall$ (осень):

$$\widehat{\ln(Q)} = \beta_0 + \beta_1 \ln(P) + \beta_2 Spr + \beta_3 Summ + \beta_4 Fall \quad R^2 = 0.367 \quad n = 20$$

Напишите спецификацию регрессии с ограничениями для проверки статистической гипотезы $H_0 : \beta_2 = \beta_4$. Дайте интерпретацию проверяемой гипотезе. Пусть для регрессии с ограничениями был вычислен $R_r^2 = 0.238$. Тестируйте нулевую гипотезу при уровне значимости 5%.

5. Рассмотрим функцию спроса с сезонными переменными $Spr$ (весна), $Summ$ (лето) и $Fall$ (осень):

$$\widehat{\ln(Q)} = \beta_0 + \beta_1 \ln(P) + \beta_2 Spr + \beta_3 Summ + \beta_4 Fall \quad R^2 = 0.247 \quad n = 24$$

Напишите спецификацию регрессии с ограничениями для проверки статистической гипотезы $H_0 : \beta_2 = 0, \beta_3 = \beta_4$. Дайте интерпретацию проверяемой гипотезе. Пусть для регрессии с ограничениями был вычислен $R_r^2 = 0.126$. Тестируйте нулевую гипотезу при уровне значимости 5%.

6. Рассматривается зарплатная модель регрессии

$$\widehat{\ln(wage)} = \theta_0 + \theta_1 ed + \theta_2 age + \theta_3 age^2 + \theta_4 imm,$$

где $ed$ - уровень образования, $age$ - возраст, $imm$ - фиктивная переменная, равная 1 для иммигрантов.

а) Пусть $\widehat{\theta_4} = -0.11$. Дайте интерпретацию полученной оценки.

б) Пусть $s_4 = 0.09$. Значима ли разница в оплате труда между иммигрантами и местными рабочими? Рассмотрите уровень значимости 1%, 5%, 10%. $n = 2003$.

Другой исследователь оценил следующую модель

$$\widehat{\ln(wage)} = \underset{(0.149)}{1.15} + \underset{(0.003)}{0.0493} ed + \underset{(0.007)}{0.069} age - \underset{(0.00008)}{0.0007} age^2 + \underset{(0.0803)}{0.0963} imm - \underset{(0.0073)}{0.031}\, ed \cdot imm,$$

$$R^2 = 0.209$$

в) Какой ожидаемый эффект от дополнительного года образования для иммигранта?

г) Какой ожидаемый уровень оплаты двадцатилетнего местного рабочего с уровнем образования 10 лет?

7. Исследуется влияние образования $Educ$, пола $Male$ и возраста $Age$ на почасовую оплату труда $Wage$ для возрастной группы 25-75 лет. Рассматривается модель регрессии

$$ln(Wage) = \theta_0 + \theta_1 Educ + \theta_1 Educ + \theta_2 Male + \theta_3 Age + \theta_4 Age^2 + error$$

С какой целью в модель регрессии был включен квадрат возраста? Какой ожидаемый знак у оценки $\theta_4$? Как можно объяснить включение именно квадрата возраста (а не, например, третьей степени)? Как можно объяснить использования в качестве зависимой переменной логарифма зарплаты?

8. Предложите и экономически обоснуйте модель регрессии для оценки влияния на почасовую оплату $wage$ возраста $age$, гендерного признака $female$ и недельной занятости $hours$

   1. Для возрастной группы 30-37 лет;

   2. Для возрастной группы 30-80 лет Дайте интерпретацию параметров модели.

9. Рассмотрим модель регрессии зависимости цены автомобиля $Price$ от его возраста $Age$. Какая модель предпочтительней:линейная или полулогарифмическая? Какие преимущества у полулогарифмической модели? Эконометрист полагает, что скорость падения цены автомобиля с возрастом уменьшается. Как учесть это предположение в модели регрессии?

10. Предложите и экономически обоснуйте модель регрессии для оценки влияния на зарплату CEO (фактор $salary$) объема продаж фирмы $sales$, доходности на собственный капитал $roe$ и доходности акций $ros$ если

    1. Модель регрессии рассматривается только для фирм с объемом продаж 200 млн. долл.- 250 млн. долл. в год.

    2. Модель регрессии рассматривается для фирм с объемами продаж 10 млн. долл. - 1 млрд. долл. в год.

    Дайте интерпретацию параметров модели.

11. Рассматриваются следующие регрессионные модели зависимости зарплаты CEO ($S$) от годового уровня продаж фирмы ($Sales$), дохода на собственный капитал $roe$ и доходности акций $ros$ (n=7340):

$$\widehat{\ln(s)} = \theta_0 + \theta_1 \ln(Sales) + \theta_2 roe + \theta_3 ros$$
$$\hat{S} = \theta_0 + \theta_1 Sales + \theta_2 Sales^2 + \theta_3 roe + \theta_4 roe^2 + \theta_5 ros$$
$$\text{Для первой модели } R^2 = 0.2129$$
$$\text{Для второй модели } R^2 = 0.2139$$

    Можно ли сравнить эти модели по критерию $R^2$? Ответ обоснуйте. Если можно, то в пользу какой модели можно сделать выбор?

12. В декабре 2007 были оценены (на одних и тех же данных) две модели регрессии зависимости зарплаты CEO $S$ от годового уровня продаж фирмы $sales$, дохода на собственный капитал $roe$ и доходности акций $ros$:

Модель 1: зависимая переменная $\ln(S)$, объем выборки 27

| $Y$ | $cosnt$ | $\ln(sales)$ | $roe$ | $ros$ |
|---|---|---|---|---|
| $coeff$ | 2.4 | 0.14 | 0.008 | 0.0013 |
| $Std.Error$ | 1.02 | 0.013 | 0.0023 | 0.0003 |

$$R^2 = 0.101$$

Модель 2: зависимая переменная $\ln(S)$, объем выборки 27

| $Y$ | $cosnt$ | $\ln(sales)$ | $roe$ | $ros$ | $roe^2$ | $ros^2$ |
|---|---|---|---|---|---|---|
| $coeff$ | 2.31 | 0.18 | 0.006 | 0.0009 | $-0.00003$ | $-0.0000078$ |
| $Std.Error$ | 1.02 | 0.026 | 0.0024 | 0.00034 | 0.00004 | 0.0000063 |

$$R^2 = 0.127$$

a) По каким критериям можно сравнить эти модели? Укажите не менее двух и дайте их описание.
b) Какая модель предпочтительней? Ответ обоснуйте.

13. *

Рассматривается задача построения линейной регрессионной модели, описывающей зависимость месячного объема потребляемой в регионе электроэнергии ($y$) от объема выработанной в данном регионе за тот же месяц промышленной продукции ($x$)

$$y_i = \theta_0 + \theta_1 x_i + \varepsilon_i, \tag{1}$$

по месячным исходным статистическим данным (за год): $(x_i, y_i)$, $i = 1, \ldots, 12$. При этом первое наблюдение ($i = 1$) соответствовало декабрю месяцу, а последнее – ноябрю следующего года.

Как видоизменить модель (1) и подходящим образом ввести фиктивные переменные отдельно в каждом из следующих сценариев:

1) Начиная с сентября месяца, в регионе введены новые промышленные мощности, энергоемкость которых существенно меньше по сравнению со средней энергоемкостью ранее существовавших мощностей.

2) Начиная с июля месяца, по решению администрации региона часть вырабатывающих продукцию предприятий переводится в другой район.

3) Установлено, что имеют место **сезонные (трехмесячные)** эффекты в динамике поведения $y_i$. Как это учесть в модели (1) с помощью введения фиктивных переменных?

4) Можно ли решить каждую из задач 1), 2) и 3) без использования фиктивных переменных, и если "да", то в чем эти решения будут уступать подходу с фиктивными переменными?

# Глава 4

# Гетероскедастичность и автокоррелированность ошибок

## 4.1. Гетероскедастичность ошибок регрессии

**Определение.** *Гетероскедастичностью*[1] (heteroskedasticity) называется нарушение условия постоянства дисперсии ошибок в линейной модели регрессии

$$\mathrm{Var}(\varepsilon_i) \neq \mathrm{const} \quad \text{или} \quad \mathrm{Var}(u|x_1, \ldots, x_k) \neq \mathrm{const},$$

но при этом считается, что остальные условия на ошибки регрессии (нулевое среднее и некоррелируемость в случае детерминированных регрессоров) выполнены.

*Замечание.* В случае постоянства дисперсий ошибок говорят о *гомоскедастичности* (homoskedasticity) или *однородности* ошибок модели регрессии.

Непостоянство ошибок регрессии означает, что для одних значений объясняющих переменных «разброс» значений зависимой переменной будет больше, а для других значений – меньше. Часто гетероскедастичность ошибок регрессии возникает при построении регрессионных моделей для неоднородных данных.

*Пример.* Рассмотрим регрессионную модель зависимости уровня оплаты труда от числа сотрудников в фирме. Тогда естественно ожидать, что чем больше число сотрудников, т.е. чем больше фирма, тем больше разброс зарплаты сотрудников. Таким образом, с возрастанием значения объясняющей переменной можно ожидать возрастание дисперсий ошибок регрессии.

Гетероскедастичность следующим образом влияет на статистические свойства OLS-оценок коэффициентов регрессии.

**Теорема.** *Если в теореме Гаусса – Маркова* $\mathrm{Var}(\varepsilon_i) \neq \mathrm{const}$ *(в модели стохастических регрессоров* $\mathrm{Var}(u|\mathbf{x}) \neq \mathrm{const}$*), то OLS-оценки коэффициентов регрессии*

[1] Иногда употребляют термин *неоднородность ошибок регрессии.*

*будут несмещенными и состоятельными (в случае стохастических регрессоров), но* ***не наилучшими*** *(не BLUE оценки).*

*Если ошибки регрессии нормально распределены, то в общем случае $t$- и $F$-статистики* ***не имеют*** *распределений Стьюдента и Фишера соответственно.*

*Замечание.* Таким образом, при гетероскедастичности ошибок регрессии OLS-оценки по-прежнему остаются несмещенными и состоятельными, но **не наилучшими**, т.е. они вполне пригодны для вычисления предсказанных значений зависимой переменной. Самое важное, что при гетероскедастичности статистические выводы, основанные на использовании $t$- и $F$-статистик, уже могут быть неверны. Например, доверительные интервалы для коэффициентов регрессии не соответствуют заявленным уровням значимости

Рассмотрим теперь две задачи: статистические тесты на обнаружение гетероскедастичности и разные способы корректировки модели регрессии на гетероскедастичность.

### 4.1.1. Тесты на гетероскедастичность

Приведем несколько наиболее употребительных тестов для выявления гетероскедастичности ошибок регрессии. Важно отметить, что большинство тестов ориентированы на выявление гетероскедастичности той или иной априорной структуры.

**Графический анализ** Достаточно эффективным методом предварительного анализа однородности ошибок модели регрессии в случае больших выборок является визуальный анализ графиков остатков. Как правило, рассматриваются следующие графики:

1. Остатки в зависимости от оцененных значений: $e$ vs $\hat{y}$.

2. Остатки в зависимости от отдельных объясняющих переменных: $e$ vs $x$.

3. Остатки в зависимости от номера наблюдений: $e_t$ vs $t$ (только в случае временных рядов).

Обычно на неоднородность ошибок указывает «неравномерность разброса» остатков на графике относительно оси абсцисс. Например, если с ростом значения объясняющей переменной $x_j$ увеличивается разброс остатков, это может указывает на то, что с увеличением значения объясняющей переменной растет и дисперсия ошибок в модели регрессии.

**Тест Goldfeld – Quandt [19]** Это тест применяется в случае, если есть априорные предположения о зависимости дисперсии ошибок от значения одной из объясняющих переменных[2] (например, фактора $x_j$) вида $\sigma_i^2 = \sigma^2 x_{ij}^2$ с неизвестным

[2]В тесте неявно предполагается, что значения этой объясняющей переменной должны быть **положительными**.

параметром $\sigma$. Такие предположения вполне обоснованы, например, в модели зависимости уровня зарплаты от числа сотрудников в фирме.

В модели с детерминированными регрессорами проверяется нулевая гипотеза

$$H_0 \,:\, \sigma_1^2 = \cdots = \sigma_n^2$$

против альтернативы

$$H_1 \,:\, \sigma_i^2 = \sigma^2 x_{ij}^2 \; (i = 1, \ldots, n),$$

где $\mathrm{Var}(\varepsilon_i) = \sigma_i^2$.

Для применения теста вначале необходимо упорядочить выборочные данные по возрастанию фактора $x_j$. После упорядочивания отбрасывается $r$ центральных наблюдений и получаем разделение выборки на две группы, как мы предполагаем, с «малыми» и «большими» дисперсиями. Отбрасывание центральных значений проводится для «более надежного» разделения на две группы. Пусть $n_1$ – число наблюдений в первой группе (с «малыми» дисперсиями), а $n_2$ – число наблюдений во второй группе (с «большими» дисперсиями). Очевидно, $n_1 = n_2 = (n - r)/2$

Далее, модель регрессии отдельно оценивается по первой и по второй группе наблюдений. Обозначим через $\mathrm{RSS}_1$ и $\mathrm{RSS}_2$ соответствующие остаточные суммы квадратов, а через $\mathrm{SER}_1$ и $\mathrm{SER}_2$ соответствующие стандартные ошибки регрессии. В качестве статистики для проверки нулевой гипотезы берется дисперсионное $F$-отношение

$$F = \frac{\mathrm{SER}_2^2}{\mathrm{SER}_1^2} = \frac{\mathrm{RSS}_2 \,/(n_2 - m)}{\mathrm{RSS}_1 \,/(n_1 - m)} = \frac{\mathrm{RSS}_2}{\mathrm{RSS}_1}.$$

При справедливости нулевой гипотезы $F$-статистика имеет распределение Фишера

$$F \underset{H_0}{\sim} F(n_1 - m, n_2 - m).$$

Следовательно, при заданном уровне значимости $\alpha$ гипотезу $H_0$ следует отвергается при $F > F_{\text{кр}}$ (т.е. при больших значениях статистики), где критическое значение $F_{\text{кр}} = F(\alpha; n_1 - m, n_2 - m)$.

*Замечание.* Число $r$ исключаемых наблюдений не должно быть «очень большим» и «очень маленьким» и нет формальных критериев выбора числа отбрасываемых наблюдений. Формально тест работает и в случае $r = 0$, но в этом случае его мощность ниже [19].

*Замечание.* Важно отметить, что это **точный** тест, т.е. применим при любом объеме выборки.

*Замечание.* Тест Goldfeld – Quandt применим также в ситуации, когда дисперсия ошибки принимает только два возможных значения. В этом случае очевидным образом происходит разделение **всех** выборочных данных на две группы[3] (вообще говоря, разного) объема $n_1$ и $n_2$ соответственно ($n_1 + n_2 = n$). Пусть $\mathrm{RSS}_1 < \mathrm{RSS}_2$. Тогда $F$-статистика теста равна

$$F = \frac{\mathrm{SER}_2^2}{\mathrm{SER}_1^2} = \frac{\mathrm{RSS}_2 \,/(n_2 - m)}{\mathrm{RSS}_1 \,/(n_1 - m)}.$$

[3]Легко понять, что в этой ситуации нет необходимости в отбрасывании значений для «более надежного» разделения на группы.

*Пример.* Исследуется зависимость логарифма цены коттеджа ($\ln Price$) от его площади ($Square$), удаленности от МКАД ($Dist$), площади приусадебного участка ($LSize$) и количества этажей ($Floor$). Было высказано предположение, что с увеличением площади коттеджа дисперсия ошибок регрессии возрастает. Для проверки этого пред- положения отдельно оценили регрессию по 42 коттеджам небольшой площади и по 42 коттеджам большой площади (всего в выборке 100 коттеджей) и получили остаточные суммы квадратов $RSS_1 = 11.2$ и $RSS_2 = 27.2$ соответственно.

а) Напишите спецификацию уравнения регрессии.

б) Можно ли сделать вывод о возрастании дисперсий ошибок регрессий при увеличении площади коттеджа? Уровень значимости 10%.

*Решение.* Воспользуемся тестом Голдфелда-Квандта для проверки гипотезы о гомоскедастичности, то есть проверяем гипотезу

$$H_0 : \ \sigma_1^2 = \ldots = \sigma_n^2$$

Используем $F$-статистику $F = \frac{RSS_2/(n_2-m)}{RSS_1/(n_1-m)}$,где $RSS_1 = 11.2$, $RSS_2 = 27.2$, $n_1 = n_2 = 42$, $m = 5$.

$$F = \frac{27.2/(42-5)}{11.2/(42-5)} = \frac{27.2}{11.2} \approx 2.43$$

А критическое значение $F_{\text{кр}}(0.05; 37; 37) \approx 1.884$, поэтому гипотеза о гомоскедастичности отвергается.

**Тест White [30]** Этот тест применяется в случае, когда есть априорные предположения, что гетероскедастичность обусловлена зависимостью дисперсии ошибки от объясняющих переменных. В случае стохастических регрессоров проверяется гипотеза

$$H_0 \ : \ \mathrm{Var}(u|x_1, \ldots, x_k) \equiv \sigma^2$$

против альтернативы

$$H_1 \ : \ \mathrm{Var}(u|x_1, \ldots, x_k) = \sigma^2(x_1, \ldots, x_k).$$

В модели с детерминированными регрессорами проверяется нулевая гипотеза

$$H_0 \ : \ \sigma_1^2 = \cdots = \sigma_n^2$$

Для проверки нулевой гипотезы White (1980) предложил следующую двухшаговую процедуру:

1. находим OLS-остатки $\{e_i\}_{i=1}^n$ в исходной модели регрессии;
2. Вычисляем коэффициент $R_0^2$ во вспомогательной регрессии $e_i^2$ на константу, регрессоры, их квадраты и попарные произведения:

$$e^2 = \delta_0 + \sum_{i=1}^{k} \delta_i x_i + \sum_{i \leqslant j}^{k} \delta_{ij} x_i x_j + \text{error} =$$
$$\delta_0 + \delta_1 x_1 + \cdots + \delta_{11} x_1^2 + \cdots + \delta_{12} x_1 x_2 + \cdots + \text{error} \quad (4.1)$$

При справедливости нулевой гипотезы о гомоскедастичности ошибок статистика $nR_0^2$ для проверки значимости вспомогательной регрессии (4.1) «в целом» асимптотически (т.е. при больших объемах выборки) имеет распределение хи-квадрат

$$nR_0^2 \underset{H_0}{\approx} \chi_K^2,$$

где число степеней свободы $K$ равно числу регрессоров во **вспомогательной** регрессии (4.1). Таким образом, при заданном уровне значимости $\alpha$ (асимптотически) нулевая гипотеза о гомоскедастичности отвергается при $nR_0^2 > \chi_{\text{кр}}^2$, где критическое значение $\chi_{\text{кр}}^2 = \chi^2(\alpha; K)$.

Несложно показать, что $K = (k^2 + 3k)/2$, где $k$, как обычно, число регрессоров в **исходной** модели регрессии.

*Замечание.* Важно отметить, что вспомогательная регрессия (4.1) не имеет никакой экономической интерпретации, а нужна **только для вычисления** коэффициента $R_0^2$.

Тест White является наиболее общим тестом для проверки гетероскедастичности ошибки регрессии, однако он имеет следующие недостатки:

- даже при небольшом количестве регрессоров в исходной модели во вспомогательной модели, оцениваемой на втором шаге, может быть «много» (относительно объема выборки) оцениваемых коэффициентов, что уменьшает мощность теста;

- если нулевая гипотеза отвергается, то не дается указаний на функциональную форму гетероскедастичности (на вид функции $\sigma^2(\mathbf{x})$).

- тест не учитывает зависимость дисперсии ошибок от невключенных в модель регрессоров.

*Замечание.* Для проверки значимости вспомогательной регрессии (4.1) «в целом» также можно использовать $F$-статистику, асимптотически имеющую распределение Фишера.

Если во вспомогательной модели (4.1) «много» (относительно объема выборки) коэффициентов, то можно воспользоваться следующей модификацией теста White [31]:

1. Находим остатки $\{e_i\}_{i=1}^n$ и предсказанные значения $\hat{y}_i$ в исходной модели регрессии.

2. Оцениваем вспомогательную регрессию $e_i^2$ на $\hat{y}_i$ и $\hat{y}_i^2$:

$$e_i^2 = \delta_0 + \delta_1 \hat{y}_i + \delta_2 \hat{y}_i^2 + \text{error} \tag{4.2}$$

   и находим коэффициента $R_0^2$.

При справедливости нулевой гипотезы о гомоскедастичности ошибок статистика $nR_0^2$ для проверки значимости вспомогательной регрессии (4.2) «в целом» асимптотически (т.е. при больших объемах выборки) имеет распределение хи-квадрат

$$nR_0^2 \underset{H_0}{\approx} \chi_2^2$$

Таким образом, при заданном уровне значимости $\alpha$ (асимптотически) нулевая гипотеза о гомоскедастичности отвергается при $nR_0^2 > \chi^2_{\text{кр}}$, где критическое значение $\chi^2_{\text{кр}} = \chi^2(\alpha; 2)$.

*Пример.* Исследуется зависимость логарифма цены коттеджа ($\ln Price$) от его площади ($Square$) и удаленности от МКАД ($Dist$) на основе данных по 800 коттеджам. Для проверки гипотезы о постоянстве дисперсий ошибок регрессии был применен тест Уайта.

а) Напишите спецификацию вспомогательной модели регрессии.

б) Пусть во вспомогательной модели регрессии $R_0^2 = 0.075$. Можно ли сделать вывод о гетероскедастичности в ошибках регрессии? Уровень значимости 1%.

*Решение.* а) Вспомогательная модель регрессии имеет вид

$$e^2 = \delta_0 + \delta_1 Square + \delta_2 Dist + \delta_{11}^2 Square^2 + \delta_{12} Square \cdot Dist + \delta_{22}^2 Dist^2 + error$$

б) Так как во вспомогательной модели регрессии $R_0^2 = 0.075$ и объем выборки $n = 800$ достаточно большой, то можно воспользоваться статистикой

$$nR_0^2 \underset{H_0}{\approx} \chi_K^2,$$

где число степеней свободы $K = 5$ равно числу регрессоров во вспомогательной регрессии. Так как

$$nR_0^2 = 800 \cdot 0.075 = 60, \text{ а } \chi_5^2(0.01) \approx 15.086,$$

то на уровне значимости 1% делаем вывод о наличии гетероскедастичности в ошибках.

**Тест Breusch – Pagan [10]** Это тест применяется, если есть априорные предположения, что дисперсия ошибок регрессии зависит от факторов $\mathbf{z}' = (z_1, \ldots, z_p)$

$$\sigma_i^2 = f(\gamma_0 + \mathbf{z}_i'\gamma) = f(\gamma_0 + z_{i1}\gamma_1 + \cdots + z_{ip}\gamma_p) \tag{4.3}$$

с некоторой **неизвестной** гладкой функцией $f(\cdot) > 0$. Проверяется нулевая гипотеза

$$H_0 \ : \ \sigma_i^2 \equiv \sigma^2 = f(\gamma_0)$$

против альтернативы

$$H_1 \ : \ \sigma_i^2 = f(\gamma_0 + \mathbf{z}_i'\gamma).$$

Другими словами, в модели гетероскедастичности (4.3) проверяется гипотеза

$$H_0' \ : \ \gamma_1 = \cdots = \gamma_p = 0.$$

Тест Breusch–Pagan основан на следующей двухшаговой процедуре:

1. В исходной модели регрессии находим OLS-остатки $e_i$ и обозначим

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2.$$

2. Вычисляем объясненную сумму квадратов ESS во вспомогательной регрессии

$$\left(\frac{e_i^2}{\widehat{\sigma}^2}\right) = \gamma_0 + \mathbf{z}_i'\gamma + \text{error} = \gamma_0 + z_{i1}\gamma_1 + \cdots + z_{ip}\gamma_p + \text{error}\,.$$

При справедливости нулевой гипотезы (т.е. при гомоскедастичности ошибок регрессии в модели (4.3)) $LM$-стастистика[4]

$$LM = \text{ESS}\,/2$$

асимптотически имеет распределение хи-квадрат

$$LM \underset{H_0}{\approx} \chi_p^2.$$

Следовательно, для больших выборок при заданном уровне значимости $\alpha$ нулевая гипотеза о гомоскедастичности ошибок регрессии отвергается при $LM > \chi^2_{\text{кр}}$, где критическое значение $\chi^2_{\text{кр}} = \chi^2(\alpha; p)$.

*Замечание.* Существует другая форма теста Breusch–Pagan, эквивалентная изложенной. Она основана на статистике $nR_0^2$, где коэффициент $R_0^2$ вычисляется по вспомогательной регрессии

$$e_i^2 = \gamma_0 + z_{i1}\gamma_1 + \cdots + z_{ip}\gamma_p + \text{error}\,.$$

При справедливости нулевой гипотезы статистика $nR_0^2$ асимптотически имеет распределение хи-квадрат

$$nR_0^2 \underset{H_0}{\approx} \chi_p^2.$$

*Пример.* По методу наименьших квадратов была оценена регрессионная модель зависимости недельной занятости hours от почасовой оплаты труда *wage* (в \$100) и уровня образования *educ*. Зависимая переменная: *hours*; объем выборки 5321.

| $Y$ | $const$ | $wage$ | $wage^2$ | $educ$ |
|---|---|---|---|---|
| $coeff$ | 23.4 | 0.84 | $-0.008$ | $-0.04$ |
| $Std.Error$ | 1.02 | 0.013 | 0.0027 | 0.023 |

$ESS = 380.12$; $RSS = 130.23$.

Статистика теста Breusch-Pagan'а на зависимость дисперсии ошибки модели регрессии от почасовой оплаты и пола человека равна 8.75.

a) Можно ли для прогнозирования занятости использовать полученные оценки коэффициентов? Ответ обоснуйте.

b) Можно ли проверять значимость влияния уровня образования на занятость с помощью стандартной t-статистики? Ответ поясните.

[4] $LM$ = Lagrange Multiplier, статистика множителей Лагранжа.

*Решение.* a) Статистика теста Breusch-Pagan'a асимптотически имеет распределение $\chi^2_2$, а критическое значение равно $\chi^2_{\text{кр}}(0.05; 2) \approx 5.991$. Поэтому гипотеза о гомоскедастичности ошибок отвергается. Поэтому использовать МНК-оценки можно,но они уже не являются наилучшими.

b) При гетероскедастичности остатков статистические выводы, основанные на использовании $t$-статистики, могут быть неверны.

**Ошибка спецификации** Важно отметить, что иногда положительные выводы тестов о наличии гетероскедастичности ошибок регрессии связаны с неправильной спецификацией модели. Избавиться от неоднородности ошибок в этом случае можно с помощью изменения спецификации модели. Например, осуществить переход к логарифмам объясняющих переменных или зависимой переменных, включить в модель нелинейные члены (например, квадраты объясняющих переменных). Как правило, такие ситуации хорошо различимы при графическом анализе данных.

### 4.1.2. Корректировка на гетероскедастичность

При гетероскедастичности ошибок регрессии МНК-оценки уже не будут наилучшими, и, что наиболее важно, статистические выводы, основанные на применении $t$- и $F$-статистик, могут привести к неправильным выводам. Поэтому для получения «лучших» оценок и исследования статистических свойств модели регрессии необходимо произвести корректировку модели на гетероскедастичность.

#### Взвешенный метод наименьших квадратов

Рассмотрим модель с детерминированными регрессорами

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

и предположим, что ошибки удовлетворяют условиям

- $\mathsf{E}\varepsilon_i = 0$;
- $\operatorname{cov}(\varepsilon_i, \varepsilon_j) = 0$ при $i \neq j$;
- $\operatorname{Var}(\varepsilon_i) = \sigma_i^2$ и дисперсии $\sigma_i^2$ **известны**.

Таким образом, нам известна **полная информация** о структуре неоднородности (гетероскедастичности) ошибок регрессии.

Разделим уравнение регрессии на $\sigma_i$:

$$\frac{y_i}{\sigma_i} = \beta_0 \frac{1}{\sigma_i} + \beta_1 \frac{x_{i1}}{\sigma_i} + \cdots + \beta_k \frac{x_{ik}}{\sigma_i} + \varepsilon_i' \quad \left(\varepsilon_i' = \frac{\varepsilon_i}{\sigma_i}\right).$$

Тогда ошибки $\varepsilon_i'$ в преобразованном уравнении регрессии уже **будут гомоскедастичны**, так как

$$\operatorname{Var}(\varepsilon_i') = \operatorname{Var}\left(\frac{\varepsilon_i}{\sigma_i}\right) = \frac{1}{\sigma_i^2} \operatorname{Var}(\varepsilon_i) = 1.$$

*Замечание.* Важно отметить, что так как дисперсии ошибок в общем случае непостоянны, то преобразованная модель регрессии в общем случае может быть моделью **без константы**.

Таким образом, ошибки в преобразованном уравнении регрессии удовлетворяют всем условиям теоремы Гаусса – Маркова и, следовательно, наилучшими линейными оценками коэффициентов (BLUE – оценками) будут OLS-оценки, получающиеся минимизацией суммы квадратов отклонений в преобразованном уравнении:

$$S = \sum_{i=1}^{n} \left( \frac{y_i}{\sigma_i} - \beta_0 \frac{1}{\sigma_i} - \beta_1 \frac{x_{i1}}{\sigma_i} - \dots - \beta_k \frac{x_{ik}}{\sigma_i} \right)^2 =$$

$$\sum_{i=1}^{n} \frac{1}{\sigma_i^2} \left( y_i - \beta_0 - \beta_1 x_{i1} - \dots - \beta_k x_{ik} \right)^2 \longrightarrow \min$$

**Определение.** Пусть заданы положительные числа $\{w_i\}_{i=1}^n$ (веса). Оценки коэффициентов модели регрессии, полученные минимизацией *взвешенной суммы квадратов отклонений*

$$\sum_{i=1}^{n} w_i \left( y_i - \beta_0 - \beta_1 x_{i1} - \dots - \beta_k x_{ik} \right)^2 \longrightarrow \min,$$

называются *WLS-оценками* (WLS = Weighted Least Squares) или *оценками взвешенного метода наименьших квадратов.*

Итак, мы доказали:

**Предложение.** *В случае гетероскедастичности ошибок в линейной модели регрессии наилучшими несмещёнными линейными оценками коэффициентов будут WLS-оценки с весами* $w_i = 1/\sigma_i^2$.

*Замечание.* Так как ошибки в преобразованном уравнении регрессии гомоскедастичны, то в случае нормальной распределенности ошибок к ней применимы все статистические выводы многофакторной модели регрессии, возможно **без константы**. В последнем случае коэффициент $R^2$ **не имеет смысла**.

Таким образом, в случае известных дисперсий ошибок для исследования статистических свойств модели регрессии необходимо рассматривать WLS-оценки коэффициентов регрессии. Однако во многих прикладных задачах дисперсии ошибок $\sigma_i^2$ **неизвестны**, и их нужно включать в число оцениваемых параметров. Но в этом случае число неизвестных параметров модели будет превышать объем выборки $n$, и представляется невозможным получить удовлетворительные оценки параметров в общем случае. Поэтому для получения «хороших» оценок коэффициентов необходимо накладывать дополнительные ограничения на структуру гетероскедастичности ошибок регрессии.

### Известна функциональная форма гетероскедастичности с точностью до множителя

Предположим, что нам известна функциональная форма зависимости дисперсии ошибок от объясняющих переменных **с точностью до неизвестного множителя** $\sigma^2$:

$$\mathrm{Var}(\varepsilon_i) = \sigma^2 h(x_1, \ldots, x_k), \tag{4.4}$$

где $h(\mathbf{x}) = h(x_1, \ldots, x_k) > 0$ – **известная функция**. По аналогии с предыдущим разделом преобразуем уравнение регрессии разделив его на $\sqrt{h(\mathbf{x}_i)}$:

$$\frac{y_i}{\sqrt{h(\mathbf{x}_i)}} = \beta_0 \frac{1}{\sqrt{h(\mathbf{x}_i)}} + \beta_1 \frac{x_{i1}}{\sqrt{h(\mathbf{x}_i)}} + \cdots + \beta_k \frac{x_{ik}}{\sqrt{h(\mathbf{x}_i)}} + \varepsilon_i' \left( \varepsilon_i' = \frac{\varepsilon_i}{\sqrt{h(\mathbf{x}_i)}} \right).$$

Тогда ошибки $\varepsilon_i'$ в преобразованном уравнении регрессии уже будут **гомоскедастичны**, так как

$$\mathrm{Var}(\varepsilon_i') = \mathrm{Var}\left( \frac{\varepsilon_i}{\sqrt{h(\mathbf{x}_i)}} \right) = \frac{1}{h(\mathbf{x}_i)} \mathrm{Var}(\varepsilon_i) = \sigma^2.$$

Следовательно, верно следующее

**Предложение.** *В случае гетероскедастичности ошибок вида* (4.4) *в линейной модели регрессии наилучшими линейными оценками коэффициетов будут WLS-оценки с весами* $w_i = 1/h(\mathbf{x}_i)$.

*Замечание.* Если ошибки нормально распределены, то к модели регрессии применимы все статистические выводы многофакторной модели регрессии, возможно **без константы**. В последнем случае коэффициент $R^2$ в преобразованной модели регрессии уже **не имеет смысла**.

*Замечание.* Часто вид функциональной зависимости $h(\cdot)$ дисперсии ошибки от объясняющих переменных можно попытаться «угадать» исходя их визуального анализа графиков остатков $e$ vs $x_j$. В некоторых случаях эта зависимость может быть описана степенной функцией $h(\cdot) = x_j^\gamma$ с **известным** параметром $\gamma > 0$ (дисперсия зависит только от одного из регрессоров).

### Доступный взвешенный метод наименьших квадратов

В некоторых случаях точна форма гетероскедастичности не очевидна и трудно найти точный вид функции $h(\cdot)$ из предыдущего раздела. Однако иногда удается «угадать» вид функции $h(\cdot)$ с точностью до параметров. Тогда можно оценить эти параметры функции $h(\cdot)$ по выборочным данным, вычислить $\hat{h}_i$ – оценки значений $h(\mathbf{x}_i)$ и использовать эти значения для получения WLS-оценок коэффициентов регрессии. Описанная процедура носит название *Доступного взвешенного метода наименьших квадратов* (FWLS = Feasible Weighted Least Squares). При общих предположениях в модели регрессии со стохастическими регрессорами

FWLS-оценки параметров регрессии являются состоятельными и асимптотически более эффективными, чем OLS-оценки. Однако они могут быть **смещенными**.

Есть много способов моделирования гетероскедастичности. Опишем процедуру FWLS на одном примере модели со стохастическими регрессорами, а именно предположим что гетероскедастичность имеет следующий вид

$$\mathrm{Var}(u|x_1,\ldots,x_k) = \sigma^2 h(\mathbf{x}) = \sigma^2 \exp(\delta_1 x_1 + \cdots + \delta_k x_k) \tag{4.5}$$

где параметры $\delta_1,\ldots,\delta_k$ и $\sigma^2$ **неизвестны**. Прологарифмировав равенство (4.5), получим

$$\ln\Big(\mathrm{Var}(u|x_1,\ldots,x_k)\Big) = \delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k.$$

Тогда процедура FLWS для корректировки гетероскедастичности состоит в следующем:

1. находим OLS-остатки $e_i$ в исходной модели регрессии;
2. оцениваем вспомогательную регрессию

   $$\ln\left(e^2\right) = \delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k + \mathrm{error}$$

   и получаем предсказанные значения $\hat{g}_i$.
3. оцениваем исходное уравнение регрессии по методу WLS с весами $w_i = 1/\hat{h}_i$, где $\hat{h}_i = \exp(\hat{g}_i)$.

*Замечание.* Наряду с гетероскедастичностью экспоненциального вида (4.5) можно рассматривать и другие функциональные формы зависимости, например, линейную зависимость

$$\mathrm{Var}(u|x_1,\ldots,x_k) = \delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k$$

или зависимость степенного характера

$$\mathrm{Var}(u|x_1,\ldots,x_k) = \sigma^2 x_1^{\gamma_1}\cdots x_k^{\gamma_k}.$$

Основной недостаток линейной модели состоит в том, что предсказанные веса $w_i = 1/\hat{h}_i$ могут оказаться отрицательными.

### Стандартные ошибки в форме Уайта

Как отмечалось, в случае гетероскедастичности ошибок регрессии основной недостаток OLS-оценок коэффициентов регрессии состоит в том, что статистические выводы, основанные на применении $t$- и $F$-статистик, уже неверны.

H. White [30] предложил использовать **устойчивые к гетероскедастичности** скорректированные стандартные ошибки (heteroskedasticity–robust standard errors) коэффициентов. $t$-статистики, вычисленные обычным способом по скорректированным стандартным ошибкам, имеют (асимптотически) нужное распределение Стьюдента.

**1.** Приведем формулу для стандартной ошибки в форме White для парной регрессии

$$y = \beta_0 + \beta_1 x + u.$$

Пусть $e_i$ – OLS-остатки в модели регрессии. Тогда выборочная дисперсия оценки $\hat{\beta}_1$, устойчивая к гетероскедастичности **любого вида**, вычисляется по формуле (сравните с выборочной OLS дисперсией):

$$\widehat{\mathrm{Var}}_w\left(\hat{\beta}_1\right) = \frac{\sum_i (x_i - \bar{x})^2 e_i^2}{\left(\sum_i (x_i - \bar{x})^2\right)^2}.$$

Для проверки гипотезы

$$H_0 \,:\, \beta_1 = \theta_0$$

надо использовать статистику

$$t = \frac{\hat{\beta}_1 - \theta_0}{\sqrt{\widehat{\mathrm{Var}}_w\left(\hat{\beta}_1\right)}},$$

имеющую (асимптотически), при справедливости нулевой гипотезы, распределение $t_{n-2}$. Также стандартные ошибки в форме White'а необходимо использовать для построение доверительных интервалов для коэффициентов.

**2.** В общем случае выборочные дисперсии (квадраты стандартных ошибок) оценок коэффициентов, устойчивые в гетероскедастичности любого вида, находятся как диагональные элементы выборочной матрицы ковариации вектора $\widehat{\beta}_{OLS}$ OLS-оценок коэффициентов регрессии, устойчивой к гетероскедастичности **любого вида**

$$\widehat{\mathrm{Var}}_w\left(\widehat{\beta}_{OLS}\right) = \frac{1}{n}\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2 \mathbf{x}_i \mathbf{x}_i'\right)\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1},$$

где $e_i$ – OLS-остатки в модели регрессии. Также скорректированные выборочные дисперсии коэффициентов могут быть вычислены по формуле

$$\widehat{\mathrm{Var}}_w\left(\hat{\beta}_j\right) = \frac{\sum_{i=1}^{n} e_i^2 \hat{u}_{ij}^2}{\mathrm{RSS}_j},$$

где $\hat{u}_{ij}$ – остатки в линейной модели регрессии фактора $x_j$ на **остальные регрессоры**, а $\mathrm{RSS}_j$ – остаточная сумма квадратов в этой регрессии.

*Замечание.* Многие современные эконометрические пакеты содержат опции по расчету стандартных ошибок коэффициентов в форме White.

**3.** Для проверки сложной гипотезы о линейных ограничениях на коэффициенты регрессии

$$H_0 \,:\, R\beta = \mathbf{r}$$

используется статистика

$$F = \frac{1}{q}\left(R\widehat{\beta}_{OLS} - \mathbf{r}\right)' \left[R \cdot \widehat{\mathrm{Var}}_w\left(\widehat{\beta}_{OLS}\right) \cdot R'\right]^{-1} \left(R\widehat{\beta}_{OLS} - \mathbf{r}\right),$$

устойчивая к гетероскедастичности любого вида. При справедливости нулевой гипотезы эта статистика асимптотически имеет распределение

$$F \underset{H_0}{\approx} F_{q,n-m},$$

где $q$ – число линейных ограничений на коэффициенты регрессии. При заданном уровне значимости $\alpha$ гипотеза $H_0$ асимтотически отвергается при $F > F_{\text{кр}}$, где критическое значение $F_{\text{кр}} = F(\alpha; q, n-m)$.

Статистика $\chi^2 = qF$ при справедливости нулевой гипотезы асимптотически имеет распределение $\chi^2_q$. Следовательно, нулевая гипотеза асимптотически отвергается при $\chi^2 > \chi^2_{\text{кр}}$, где критическое значение $\chi^2_{\text{кр}} = \chi^2(\alpha; q)$.

## 4.2. Корреляция во времени ошибок регрессии

Одним из условий на ошибки линейной модели регрессии была их некоррелируемость в различных наблюдениях. Неформально это означает, что «данные одного наблюдения не влияют на данные других наблюдений». При работе с пространственными выборками (cross–sectional data) это условие можно считать выполненным исходя из априорных соображений.

Однако при построении регрессионных моделей для временных рядов условие некоррелируемости ошибок может нарушаться. Это может быть связано с тем, что на зависимую переменную влияют не только значения регрессоров в соответствующий период времени, но и их значения в прошлые периоды времени (т.н. «эффект памяти»).

**Определение.** *Автокорреляцией* (serial correlation) *порядка $p$ в ошибках модели регрессии* называется зависимость ошибок вида

$$\varepsilon_t = \rho_1\varepsilon_{t-1} + \cdots + \rho_p\varepsilon_{t-p} + u_t,$$

где $u_t$ удовлетворяет условиям теоремы Гаусса – Маркова.

Можно доказать, что при автокорреляции первого порядка OLS-оценки коэффициентов регрессии буду несмещенными и состоятельными, однако оценки дисперсий коэффициентов регрессии будут смещены вниз. Как следствие, $t$- и $F$- статистики уже не имеют распределения Стьюдента и Фишера соответственно. Более того, (в случае автокорреляции первого порядка) доверительные интервалы, построенные по стандартным формулам, будут иметь доверительную вероятность меньше заявленной. Говоря нестрого, по методу OLS мы получаем «более оптимистичную» картину, чем есть на самом деле.

### 4.2.1. Автокорреляция первого порядка

Одной из наиболее часто используемых моделей автокорреляции является модель автокорреляции первого порядка

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t, \quad |\rho| < 1 \tag{4.6}$$

где $\{u_t\}$ удовлетворяют условиям Гаусса – Маркова.

В модели автокорреляции первого порядка несложно показать, что

$$\operatorname{cov}(\varepsilon_t, \varepsilon_{t-l}) = \rho^l \operatorname{Var}(u), \quad l = 0, 1, 2, \ldots$$

откуда получаем, что

$$\operatorname{corr}(\varepsilon_t, \varepsilon_{t-1}) = \frac{\operatorname{cov}(\varepsilon_t, \varepsilon_{t-1})}{\operatorname{Var}(u)} = \rho.$$

Из этого соотношения становится понятным смысл термина автокорреляция: параметр $\rho$ есть не что иное как коэффициента корреляции между соседними ошибками в модели регрессии.

В зависимости от знака параметра $\rho$ говорят о положительной или отрицательной автокорреляции в ошибках регрессии. В случае положительной автокорреляции в ряду остатков $e_1, \ldots, e_n$ наблюдается «тенденция к сохранению знака», а в случаем отрицательной автокорреляции – «тенденция к смене знака».

*Замечание.* Модель (4.6) зависимости (автокоррелированности) ошибок регрессии называется моделью авторегрессии первого порядка и обозначается AR(1). Условие $|\rho| < 1$ означает, что модель AR(1) задает стационарный временной ряд.

#### Тесты на автокорреляции

Опишем несколько наиболее часто используемых тестов на выявление автокорреляции первого порядка. Во всех тестах в рамках модели автокорреляции (4.6) проверяется гипотеза

$$H_0 \,:\, \rho = 0$$

против альтернативы

$$H_1 \,:\, \rho \neq 0.$$

**Тест Durbin – Watson [15, 16]** Тест основан на статистике

$$\mathrm{DW} = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}, \quad 0 \leqslant \mathrm{DW} \leqslant 4.$$

Можно показать, что

$$\mathrm{DW} \xrightarrow[\mathsf{P}]{} 2(1-\rho).$$

Следовательно, если автокорреляция отсутствует, то значение статистики DW должно быть «близко» к 2. Положительной автокорреляции соответствуют «малые» значения статистики, а отрицательной – «большие» значения статистики

DW. Более точно, из специальной таблицы теста при заданных уровне значимости, $n$ и $k$ берутся два критических значения $d_u > d_l$ (верхнее и нижнее, u=upper, l=low). Имеем следующий статистический критерий:

- при $d_u < \mathrm{DW} < 4 - d_u$ данные согласуются с нулевой гипотезой об отсутствии автокорреляции первого порядка;
- при $\mathrm{DW} < d_l$ гипотеза $H_0$ отвергается в пользу $H_1$, есть **положительная** автокорреляция в ошибках;
- при $4 - d_l < \mathrm{DW} < 4$ гипотеза $H_0$ отвергается в пользу $H_1$, есть **отрицательная** автокорреляция в ошибках;
- при $d_l < \mathrm{DW} < d_u$ или $4 - d_u < \mathrm{DW} < 4 - d_l$ **неопределенность**, нельзя сделать выбор ни в пользу $H_0$, ни в пользу $H_1$.

Особенностью теста является наличие «зон неопределенности», когда нулевая гипотеза не принимается и не отклоняется. Это вызвано тем, что распределение статистики DW «не свободно» от выборочных данных.

Важно отметить **условия применимости** теста Durbin – Watson:

1. В модель регрессии **включена константа** $\beta_0$.
2. Регрессоры **не коррелируют** с ошибками регрессии (strictly exogenous), в частности среди объясняющих факторов не должно быть лаговых значений зависимой переменной.

*Пример.* По квартальным данным за 2002-2008 года была оценена регрессионная модель, связывающая количество вакансий ($y_t$), уровень безработицы ($u_t$) и объем ВНП ($z_t$)

$$\widehat{\ln y_t} = 3.4 + 0.3\ln(u_t) + 0.03\ln(z_t), \quad R^2 = 0.69, \quad DW = 3.22$$

Будут ли ошибки в модели регрессии автокоррелированны (первого порядка)? Если да, то автокорреляция положительная или отрицательная?

*Решение.* Объём выборки равен $n = 28$ (7 лет по 4 квартала), а число регрессоров равно $k = 2$. Находим границы критической области по таблице Дарбина-Уотсона $d_l \approx 1.255$ и $d_u \approx 1.56$. В нашем случае $DW = 3.22$, поэтому $DW \in (4 - d_l; 4) = (2.745; 4)$ и гипотеза $H_0$ отвергается в пользу $H_1$. Делаем вывод о наличии отрицательной автокорреляции в ошибках.

**Асимптотические тесты** лишены недостатков теста Durbin – Watson, но применимы только при больших объемах выборки. Тесты основаны на следующей двухшаговой процедуре:

1. Вычисляем OLS-остатки $e_t$ в исходной модели регрессии.

2. Оцениваем вспомогательную регрессию $e_t$ на константу, лаговое значение остатка $e_{t-1}$ и **все** регрессоры исходной модели

$$e_t = \gamma_0 + \gamma_1 e_{t-1} + \sum_{j=1}^{n} \delta_j x_j + \text{error}\,. \tag{4.7}$$

**Тест множителей Лагранжа** ($LM$-test) основан на статистике

$$LM = (n-1)R_0^2,$$

где коэффициент $R_0^2$ вычисляется во вспомогательной регрессии (4.7). При справедливости нулевой гипотезы при больших объемах выборки статистика имеет распределение

$$LM \underset{H_0}{\approx} \chi_1^2.$$

Следовательно, при заданном уровне значимости нулевую гипотезу нужно отвергать при $LM > \chi^2_{\text{кр}}$, где $\chi^2_{\text{кр}} = \chi^2(\alpha; 1)$ есть критическое значение распределения $\chi_1^2$.

*Замечание.* Тест множителей Лагранжа является частным случаем асимптотического теста Breusch – Godfrey на автокорреляцию произвольного порядка.

*Замечание.* Важно отметить, что оба описанных теста требуют **постоянства дисперсий** ошибок регрессии. Существуют модификации этих тестов, устойчивые к гетероскедастичности ошибок регрессии.

### Корректировка модели регрессии на автокорреляцию

Для простоты процесс корректировки модели регрессии на автокорреляцию ошибок первого порядка (4.6) рассмотрим на примере однофакторной модели регрессии

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t.$$

**Значение $\rho$ известно.** Умножим уравнение для наблюдения $t-1$ на $\rho$, вычтем из уравнения для наблюдения $t$ и с учетом (4.6) получаем:

$$y_t - \rho y_{t-1} = \beta_0(1-\rho) + \beta_1(x_t - \rho x_{t-1}) + u_t$$

Для корректировки первого наблюдения умножим уравнение при $t = 1$ на $\sqrt{1-\rho^2}$:

$$\sqrt{1-\rho^2}y_1 = \sqrt{1-\rho^2}\beta_0 + \sqrt{1-\rho^2}\beta_1 x_1 + \sqrt{1-\rho^2}\varepsilon_1$$

Введем новые факторы

$$\tilde{y}_t = \begin{cases} y_t - \rho y_{t-1}, & t = 2, \ldots, n \\ \sqrt{1-\rho^2}y_1, & t = 1. \end{cases}$$

$$\tilde{x}_t = \begin{cases} x_t - \rho x_{t-1}, & t = 2, \ldots, n \\ \sqrt{1-\rho^2}x_1, & t = 1. \end{cases}$$

$$\tilde{z}_t = \begin{cases} 1-\rho, & t = 2, \ldots, n \\ \sqrt{1-\rho^2}, & t = 1. \end{cases}$$

Тогда преобразованное уравнение регрессии можно записать в виде

$$\tilde{y}_t = \beta_0 \tilde{z}_t + \beta_1 \tilde{x}_t + u_t, \quad t = 1, \ldots, n$$

Ошибки в преобразованном уравнении буду удовлетворять всем условиям теоремы Гаусса –Маркова. Следовательно, OLS-оценки коэффициентов $\beta_0$ и $\beta_1$ в преобразованном уравнении будут BLUE-оценками и для них верны стандартные статистические выводы модели без константы.

Описанное преобразование называется процедурой Prais-Winsten (1954), а полученные оценки называются иногда оценками Prais-Winsten.

*Замечание.* Если опустить преобразование уравнения регрессии с $t = 1$, т.е. оценивать только регрессию

$$y_t - \rho y_{t-1} = \beta_0(1-\rho) + \beta_1(x_t - \rho x_{t-1}) + u_t, \quad t = 2, \ldots, n,$$

то полученное преобразование называется преобразованием Cochrane – Orcutt (1949). При таком преобразовании объем выборки, на основе которой производится оценивание коэффициентов, уменьшается на единицу. Очевидно, для больших выборок обе оценки будут мало отличаться. Так как предпочтительней работать с моделью регрессии с включенной константой, то при больших выборках предпочтительней провести преобразование Cochrane – Orcutt.

**Значение $\rho$ неизвестно.** Однако во многих прикладных задачах значение параметра $\rho$, как правило, неизвестно, и мы можем получить только его оценку по выборочным данным. В этом случае обычно применяется итерационная двухшаговая процедура

1. Для OLS-остатков $e_t$ находим оценку параметра $\rho$ в модели $e_t = \rho e_{t-1} + \text{error}$:

$$\widehat{\rho} = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n-1} e_t^2}.$$

2. Проводим преобразование Prais – Winsten или преобразование Cochrane – Orcutt с $\rho = \widehat{\rho}$ и находим OLS-оценки коэффициентов регрессии.

3. Находим остатки в преобразованной модели регрессии и повторяем процедуру снова, начиная с п. 1.

Как правило, процесс заканчивается, когда новые значения коэффициентов «мало» отличаются от предыдущих или фиксируется количество итераций.

*Замечание.* Фактически описанная процедура представляют собой доступный метод наименьших квадратов (FWLS) в условиях автокорреляции.

*Замечание.* Многие эконометрические программные пакеты автоматически проводят корректировку модели регрессии на автокорреляцию произвольного порядка.

### 4.2.2. Автокорреляция произвольного порядка

Рассмотрим теперь модель автокорреляции ошибок регрессии произвольного порядка $p$:

$$\varepsilon_t = \rho_1 \varepsilon_{t-1} + \cdots + \rho_p \varepsilon_{t-p} + w_t, \tag{4.8}$$

и предполагается, что $u_t$ удовлетворяют условиям Гаусса – Маркова. Модель (4.8) зависимости ошибок регрессии есть модель $AR(p)$ авторегрессии порядка $p$ и для получения «хороших» статистических свойств оценок регрессии на модель (4.8) необходимо накладывать условия, гарантирующие стационарности временного ряда $\varepsilon_t$.

Рассмотрим тесты на обнаружение автокорреляции .

#### Тесты на автокорреляцию произвольного порядка

В рамках модели автокорреляции ошибок регрессии (4.8) (порядок автокорреляции $p$ **фиксирован**) проверяется нулевая гипотеза

$$H_0 \ : \ \rho_1 = \cdots = \rho_p = 0$$

против альтернативы

$$H_1 \ : \ \rho_1^2 + \cdots + \rho_p^2 > 0$$

Тесты на автокорреляцию порядка $p$ основаны на следующей двухшаговой процедуре

1. Вычисляем OLS-остатки $e_t$ в исходной модели регрессии.
2. Оцениваем вспомогательную регрессию $e_t$ на константу, лаговые значения остатков $e_{t-1}, \ldots, e_{t-p}$ и **все** регрессоры исходной модели

$$e_t = \gamma_0 + \gamma_1 e_{t-1} + \cdots + \gamma_p e_{t-p} + \sum_{j=1}^{k} \delta_j x_j + \text{error}\,. \tag{4.9}$$

**F-тест** основан на обычной проверке значимости совместного влияния лаговых переменных на $e_t$ в регрессии (4.9), т.е. на проверке гипотезы

$$H_0' \ : \ \gamma_1 = \cdots = \gamma_p = 0.$$

$H_0$ отвергается если отвергается $H_0'$.

**Тест множителей Лагранжа** или тест Breusch – Godfrey основан на статистике множителей Лагранжа

$$LM = (n-p)R_0^2,$$

где $R_0^2$ – коэффициент $R^2$ во вспомогательной регрессии (4.9). При справедливости нулевой гипотезы об отсутствии автокорреляции ошибок регрессии и при больших объемах выборки эта статистика имеет распределение хи-квадрат

$$LM \underset{H_0}{\approx} \chi_p^2.$$

Следовательно, нулевая гипотеза об отсутствии автокорреляции порядка $p$ отвергается при $LM > \chi^2_{\text{кр}}$, где критическое значение $\chi^2_{\text{кр}} = \chi^2(\alpha; p)$.

*Замечание.* Важно отметить, что оба описанных теста требуют **гомоскедастичности** ошибок регрессии. Существуют модификации этих тестов, устойчивые к гетероскедастичности ошибок регрессии.

*Пример.* Рассмотрим кривую Филлипса с запаздыванием, связывающую темпы инфляции $inf_t$ с уровнем безработицы $unem_t$, уровнем безработицы с лагом один $unem_{t-1}$ и уровнем инфляции с лагом один $inf_{t-1}$

$$\widehat{inf_t} = 1.23 + 0.12inf_{t-1} - 0.12unem_t - 0.002unem_{t-1}, \quad R^2 = 0.12 \quad n = 60.$$

a) Опишите двухшаговую процедуру тестирования ошибок регрессии на автокорреляцию третьего порядка и напишите спецификацию вспомогательной модели регрессии.

b) Пусть во вспомогательной регрессии $R_0^2 = 0.17$. Можно ли сделать вывод об автокоррелированности (третьего порядка) ошибок регрессии в кривой Филлипса? Уровень значимости 10%.

c) Можно ли в данной модели регрессии для тестирования ошибок регрессии на автокорреляцию первого порядка использовать тест Дарбина-Уотсона?

*Решение.* a) Тест на автокорреляцию порядка $p$ основаны на следующей двухшаговой процедуре:

1. вычисляем OLS-остатки $e_t$ в исходной модели регрессии;

2. оцениваем вспомогательную регрессию $e_t$ на константу, лаговые значения остатков $e_{t-1}, e_{t-2}, e_{t-3}$ и все регрессоры исходной модели

$$e_t = \gamma_0 + \gamma_1 e_{t-1} + \gamma_2 e_{t-2} + \gamma_3 e_{t-3} + \delta_1 inf_{t-1} + \delta_2 unem_t + \delta_3 unem_{t-1} + error$$

И проверяется нулевая гипотеза

$$H_0 \ : \ \gamma_1 = \gamma_2 = \gamma_3 = 0$$

b) Так как коэффициент во вспомогательной регрессии равен $R_0^2 = 0.17$, а $n = 60$ (выборка достаточно большая), то можно воспользоваться тестом множителей Лагранжа, основанном на статистике

$$LM = (n-3)R_0^2.$$

Статистика имеет асимптотическое распределение $\chi^2_3$. Значение статистики равно $LM = 57 \cdot 0.017 = 0.969$, а критическое значение $\chi^2_3(0.1) \approx 6.251$. Поэтому на уровне значимости 10% гипотеза не отвергается, и можно сделать вывод об отсутствии автокоррелированности (третьего порядка) ошибок регрессии в кривой Филлипса.

c) Тест Дарбина-Уотсона использовать нельзя, потому среди регрессоров есть лаговое значение зависимой переменной.

Аналогично случаю автокорреляции первого порядка проводится корректировка модели на автокорреляцию произвольного порядка. Многие современные эконометрические пакеты позволяют проводить корректировку модели на автокорреляцию произвольного порядка.

**Неправильная спецификация.** Как и в модели первого порядка, иногда выводы об автокорреляция ошибок регрессии могут быть связаны с неправильной спецификацией модели.

## 4.3. Корректировка модели на автокорреляцию и гетероскедастичность

**1.** Как уже отмечалось, и в случае гетероскедастичности, и в случае автокоррелированности ошибок регрессии основной недостаток OLS-оценок коэффициентов регрессии состоит в том, что к ним неприменимы статистические выводы, основанные на использовании стандартных $t$- и $F$-статистик. Однако в обоих случаях OLS-оценки остаются несмещенными и состоятельными (но не наилучшими) и вполне пригодны для вычисления предсказанных значений зависимой переменной.

В предыдущих разделах были описаны разные подходы к корректировки модели на гетероскедастичность или автокорреляцию. Первый подход (общий для обоих ситуаций) состоял в преобразовании уравнения регрессии таким образом, чтобы новая модель удовлетворяла условиям теоремы Гаусса – Маркова. Тогда оценки, полученные в преобразованной модели, «точнее» OLS-оценок в исходной модели и к ним применимы все статистические выводы стандартной модели регрессии. Основной недостаток этого подхода состоит в том, что он требовал априорной информации о структуре гетероскедастичности или автокорреляции ошибок регрессии. Также большинство тестов на гетероскедастичность и автокорреляцию нацелены на выявление гетероскедастичности и автокорреляции определенной структуры. Кроме того, в изложенных тестах на автокорреляцию было требование гомоскедастичности ошибок регрессии.

Второй подход, реализованный в рамках коррекции модели на гетероскедастичность **произвольной структуры**, состоял в применении стандартных ошибок в форме White'a. В рамках этого подхода предлагается для вычисления предсказанных значений зависимой переменной использовать OLS-оценки коэффициентов в исходной модели (мы знаем, что они вполне для этого пригодны), но для исследования статистических свойств модели регрессии необходимо использовать скорректированные $t$- и $F$-статистики, имеющие (асимптотические) нужные распределения.

**2.** Приведем обобщение второго подхода на случай корректировки модели **одновременно на гетероскедастичность и автокорреляцию**. Предположим сначала, что для ошибок модели регрессии выполнено

$$\operatorname{cov}(\varepsilon_t, \varepsilon_{t-j}) = 0 \text{ при } j > Q \tag{4.10}$$

(или $\mathsf{E}(\varepsilon_t \varepsilon_{t-s}) = 0$ при $j > Q$). Это верно, например, если для описания автокорреляции ошибок регрессии используется модель скользящего среднего. Тогда состоятельная оценка матрицы $\operatorname{Var}(\widehat{\beta}_{OLS})$ вариации вектора OLS-оценок коэффи-

циентов регрессии вычисляется по формуле

$$\widehat{\mathrm{Var}}\left(\widehat{\beta}_{OLS}\right) = n\left(\mathbf{X}'\mathbf{X}\right)^{-1}\widehat{S}_n(\mathbf{X}'\mathbf{X})^{-1}, \tag{4.11}$$

где матрица

$$\widehat{S}_n = \frac{1}{n}\sum_{t=1}^{n} e_t^2\mathbf{x}_t\mathbf{x}_t' + \frac{1}{n}\sum_{j=1}^{Q} w_j\left(\sum_{j+1}^{n} e_t e_{t-j}(\mathbf{x}_t\mathbf{x}_{t-j}' + \mathbf{x}_{t-j}\mathbf{x}_t')\right)$$

с некоторыми положительными весами $w_j$ (подробнее о весах ниже). Стандартные ошибки коэффициентов регрессии есть корни из диагональных элементов матрицы $\widehat{\mathrm{Var}}\left(\widehat{\beta}_{OLS}\right)$.

**Определение.** Оценка (4.11) матрицы вариации вектора OLS-оценок коэффициентов называется *оценкой Newey – West*, а вычисленные по ней стандартные ошибки коэффициентов регрессии называются *стандартными ошибками (в форме) Newey – West.*

Таким образом, оценка Newey – West (при подходящем выборе весов $\{w_j\}$!) является состоятельной оценкой матрицы вариации $\mathrm{Var}(\widehat{\beta})$, устойчивой **одновременно и к гетероскедастичности, и к автокорреляции** вида (4.10) (HAC standard errors, heteroskedasticity and autocorrelation consistent standard errors).

В оригинальной работе [25] доказана состоятельность оценки (4.11) для т.н. весов Бартлетта

$$w_j = 1 - \frac{j}{Q} \quad j = 1, \ldots, Q.$$

Часто предпочтительней использовать т.н. веса Парзена

$$w_j = \begin{cases} 1 - \frac{6j^2}{(1+Q)^2} + \frac{6j^3}{(1+Q)^3}, & 1 \leqslant j \leqslant \frac{1+Q}{2} \\ 2\left(1 - \frac{j}{1+Q}\right)^2, & \frac{1+Q}{2} \leqslant j \leqslant Q \end{cases}$$

*Замечание.* Очевидно, что в случае $w_j = 0$ получаем оценку White'а, устойчивую к гетероскедастичности.

Таким образом, при выполнении условия (4.10) на ошибки регрессии, асимптотически для проверки простых гипотез о коэффициентах регрессии и вычисления (асимптотических) доверительных интервалов следует использовать стандартные формулы и стандартные ошибки Newey–West.

Для проверки сложной гипотезы о линейных ограничениях на коэффициенты регрессии

$$H_0 \ : \ R\beta = \mathbf{r}$$

используется статистика

$$F = \frac{1}{q}\left(R\widehat{\beta}_{OLS} - \mathbf{r}\right)'\left[R \cdot \widehat{\mathrm{Var}}\left(\hat{\beta}_{OLS}\right) \cdot R'\right]^{-1}\left(R\widehat{\beta}_{OLS} - \mathbf{r}\right).$$

где $q$ – число линейных ограничений на коэффициенты регрессии. При справедливости нулевой гипотезы эта статистика асимптотически имеет $F$-распределение

$$F \underset{H_0}{\approx} F_{q,n-m},$$

При заданном уровне значимости $\alpha$ гипотеза $H_0$ асимтотически отвергается при $F > F_{\text{кр}}$, где критическое значение $F_{\text{кр}} = F(\alpha; q, n-m)$.

Статистика $\chi^2 = qF$ при справедливости нулевой гипотезы асимптотически имеет распределение хи-квадрат $\chi^2_q$. При заданном уровне значимости $\alpha$ гипотеза $H_0$ асимтотически отвергается при $\chi^2 > \chi^2_{\text{кр}}$, где критическое значение $\chi^2_{\text{кр}} = \chi^2(\alpha; q)$.

**3.** Если для описания автокорреляции ошибок регрессии используется модель (4.8) авторегрессии порядка $p$, то условие (4.10) не выполнено. В работе [25] доказано, что в этом случае оценка (4.11) будет состоятельной, если $Q = Q(n)$ выбрано так, что $Q(n)$ «достаточно велико», а отношение $Q(n)/\sqrt[4]{n}$ «достаточно мало»[5].

[5]Точнее, состоятельность оценки доказана при условии $Q = Q(n) \to +\infty$ и $Q(n)/\sqrt[4]{n} \to 0$ при $n \to +\infty$.

## 4.4. Задачи

1. Доказать, что ковариационная матрица симметрична и положительно определена. Для преобразованного вектора $Y = CX$ проверить равенство

$$\operatorname{cov}(CX) = C \operatorname{cov}(Y) C^T.$$

2. Рассмотрим ОЛММР $y_i = \theta_0 + \theta_1 x_{1i} + \ldots + \theta_k x_{ki} + \varepsilon_i$, где $i = 1, \ldots, n$, $E(\varepsilon_i) = 0$, $\operatorname{cov}(\varepsilon_i, \varepsilon_j) = \sigma_i^2 \delta_{ij}$, $i, j = 1, \ldots, n$.

   В матричном виде

   а) найти ОМНК-оценки параметров;

   б) доказать несмещенность;

   в) найти ковариационную матрицу оценок;

   г)выпишите, из какого условия находится ОМНК-оценка;

   д)выпишите формулу для прогноза зависимой переменной;

   е) докажите равенство $cov(\widehat{\Theta}_{\text{МНК}}, \widehat{\Theta}_{\text{ОМНК}}) = D(\widehat{\Theta}_{\text{ОМНК}})$.

3. Рассмотрим модель регрессии $y_i = \theta x_i + \varepsilon_i$, для ошибки которой выполнены условия $E(\varepsilon_i) = 0$, $D(\varepsilon_i) = \sigma^2 x_i^2$, $\operatorname{cov}(\varepsilon_i, \varepsilon_j) = 0$ для $i \neq j$.

   а) Будет ли МНК-оценка коэффициента несмещенной? Состоятельной? Оптимальной?

   б) Предложите несмещенную оценку коэффициента $\theta$ с наименьшей дисперсией. Как это соотносится с теоремой Гаусса-Маркова?

   в) Что минимизируется при поиске ОМНК-оценки?

4. Рассмотрим модель регрессии $y_i = \theta_0 + \theta_1 x_i + \varepsilon_i$ для ошибок выполнено $E(\varepsilon_i) = 0$, $D(\varepsilon_i) = \sigma^2 e^{10x_i}$, $cov(\varepsilon_i, \varepsilon_j) = 0$ для $i \neq j$.

   Пусть $\widehat{\theta}_1$ есть МНК-оценка коэффициента $\theta_1$.

   а) Будет ли оценка $\widehat{\theta}_1$ несмещенной оценкой $\theta_1$? Состоятельной оценкой?

   б) Будет ли оценка $\widehat{\theta}_1$ наилучшей оценкой коэффициента $\theta_1$? Если нет, то предложите оценку коэффициента с меньшей дисперсией.

5. Исследуется зависимость цена коттеджа $Price$ от удаленности от МКАД $Dist$, площади $Hsize$ и площади участка $Lsize$. Эконометрист полагает, что дисперсия ошибки в линейной модели регрессии зависит от площади дома и площади участка.

   а) Опишите процедуру теста Breusch-Pagan'а и укажите проверяемую статистическую гипотезу.

   б) Можно ли применять этот тест при небольших выборках?

   в) Если тест отвергает нулевую гипотезу, то какие могут быть причины?

6. *

В Вашем распоряжении имеются результаты обследования $n$ семей $\{(x_i, y_i)\}$, $i = 1, \ldots, n$, где $x_i$ и $y_i$ — удельные величины соответственно доходов и расходов (на продукты питания) в $i$-й обследованной семье.

Известно, что $\ln x_i$ и $\ln y_i$ связаны между собой линейной стохастической зависимостью $\ln y_i = a \ln x_i + \varepsilon_i$, где $\varepsilon_i$ — остаточные случайные компоненты (отражающие влияние на $y_i$ других факторов) статистически взаимно независимы, имеют нулевые средние значения и дисперсии, пропорциональные $(\ln x_i)^2$ (т.е. $\operatorname{Var} \varepsilon_i = \sigma^2 (\ln x_i)^2$ — неизвестный параметр).

1) Выписать явный вид оценок $\hat{a}_{\text{ОМНК}}$ и $\hat{\sigma}_{\text{ОМНК}}$ по обобщенному (взвешенному) методу наименьших квадратов для неизвестных значений параметров $a$ и $\sigma^2$.

2) Дать общее определение эластичности анализируемого результирующего показателя по объясняющей переменной и выписать явный вид оценки эластичности семейных расходов на продукты питания по семейному доходу в данном примере.

3) Будет ли оценка по обычному методу наименьших квадратов

$$\hat{a} = \frac{\sum (\ln x_i)(\ln y_i)}{\sum (\ln x_i)^2}$$

состоятельной оценкой параметра $a$ анализируемой модели? Почему оценка $\hat{a}_{\text{ОМНК}}$ по обобщенному методу наименьших квадратов предпочтительнее оценки $\hat{a}$?

4) Выписать явный вид наилучшего (в смысле среднего квадрата ошибки) прогноза для ожидаемой величины расходов на питание в семье, имеющей доходы $x_0$. Какова среднеквадратическая ошибка этого прогноза?

7. *

Имеются результаты произведенных в последовательные "моменты времени" $t$ $(t = 1, 2, \ldots, n)$ наблюдений $(x_t, y_t)$, описывающих динамику основных фондов $x_t$ и объема произведенной продукции $y_t$ на некотором предприятии на протяжении $n$ "тактов"времени (кварталов). Известно, что величины $x_t$ и $y_t$ связаны стохастическим соотношением

$$y_t = \theta x_t + \varepsilon_t, \tag{1}$$

в котором $\varepsilon_t$ $(t = 1, 2, \ldots, n)$ - взаимно независимые случайные остатки, отражающие влияние на $y_t$ неучтенных факторов, так что: $E\varepsilon_t = 0$, $E(\varepsilon_{t_1}\varepsilon_{t_2}) = 0$ при $t_1 \neq t_2$ $(t_1, t_2 = 1, 2, \ldots, n)$.

Однако предварительный анализ остатков $\varepsilon_t$ свидетельствует о возможности их гетероскедастичности в форме $D\varepsilon_t = \sigma^2 h(\beta_0 + \beta_1 x_t)$, где $h(\beta_0 + \beta_1 x_t) = (\beta_0 + \beta_1 x_t)^2$. В ходе реализации процедуры Бреуша-Пагана, предназначенной

для проверки гипотезы $H_0 : \beta_1 = 0$ (или, что то же, $H_0 : \ D\varepsilon_t = const$), то есть гипотезы о гомоскедастичности остатков, были вычислены значения:

$\hat{\theta}_{\text{МНК}} = 1,05$; $\hat{\varepsilon}_t = y_t - \hat{\theta}_{\text{МНК}} x_t$, $\hat{\sigma}^2 = 22,2$; $\hat{\beta}_0 = 0,2$; $\hat{\beta}_1 = 0,005$; $h_t = (\hat{\beta}_0 + \hat{\beta}_1 x_t)^2$; $\bar{h} = \frac{1}{n}\sum\limits_{t=1}^{n} h_t$; $\gamma = \frac{1}{2}\sum\limits_{t=1}^{n}(h_t - \bar{h})^2 = 6,80$; $\hat{\theta}_{\text{ОМНК}} = 1,15$; $\widehat{D\hat{\theta}}_{\text{ОМНК}} = 0,0225$.

Планируется установить объем основных фондов на этом предприятии в следующем $(n+1)$-м квартале на уровне $x_{n+1} = 60$ млн.руб. Требуется:

1) Описать ковариационную матрицу $\Sigma_\varepsilon$ случайных остатков $\varepsilon_t$ (т.е. - ее элементы $E(\varepsilon_t \varepsilon_{t+\tau})$ при $t = 1, 2, \ldots, n;\ t+\tau = 1, 2, \ldots, n$) в терминах параметров условий задачи и наблюдений.

2) Выписать явный вид наилучшей (в смысле среднего квадрата ошибки) оценки для параметра $\theta$. В чём состоит преимущество оценки $\hat{\theta}_{\text{ОМНК}}$ перед оценкой $\hat{\theta}_{\text{МНК}}$.

3) Выписать явный вид наилучшего (в смысле среднего квадрата ошибки) прогноза для ожидаемой величины объема продукции, произведенной в $(n+1)$-м квартале. Какова среднеквадратическая ошибка этого прогноза?

4) Пояснить, о чем свидетельствует значение статистики $\gamma$ и почему?

5) Пояснить, как получены оценки $\hat{\beta}_0$ и $\hat{\beta}_1$?

6) Построить интервальную оценку для параметра $\theta$ (при доверительной вероятности $P_0 = 0,95$).

8. По методу наименьших квадратов была оценена регрессионная модель зависимости недельной занятости $hours$ от почасовой оплаты труда $wage$ (в \$100) и уровня образования $educ$

Зависимая переменная: $hours$; объем выборки 1321

| | const | $wage$ | $wage^2$ | $educ$ |
|---|---|---|---|---|
| coefficient | 23.4 | 0.84 | -0.008 | -0.04 |
| Std. Error | 1.02 | 0.013 | 0.0027 | 0.023 |

Статистика теста White'а равна 6.34.

a) Можно ли использовать полученные оценки коэффициентов для прогнозирования занятости ? Ответ обоснуйте.

b) Можно ли проверять значимость влияния уровня образования на занятость с использование стандартной $t$-статистики? Ответ поясните.

9. Для количественного описания зависимости зарплаты $CEO(salary)$ от объема продаж фирмы $sales$, ее рыночной стоимости $mktval$ и стажа работы в должности в фирме $ceoten$ была оценена модель регрессии. Зависимая переменная: $ln(salary)$; объем выборки 177.

| $Y$ | $const$ | $\ln(sales)$ | $\ln(mktval)$ | $ceoten$ |
|---|---|---|---|---|
| $coeff$ | 4.5038 | 0.1629 | 0.1092 | 0.0117 |
| $OLSStd.Error$ | 0.2572 | 0.0392 | 0.0496 | 0.0053 |
| $White'sStd.Error$ | 0.2801 | 0.0382 | 0.0499 | 0.0073 |

a) Как можно объяснить выбор такой спецификации модели?

b) Проверьте значимость коэффициентов модели с использованием стандартных ошибок коэффициентов метода наименьших квадртов и стандартных ошибок, устойчивых к гетероскедастичности (стандартных ошибок White'a). Какой можно сделать вывод?

10. Для модели регрессии из предыдущей задачи была оценена вспомогательная регрессия. Зависимая переменная: $e^2$; объем выборки 177.

| $Y$ | $const$ | $\ln(sales)$ | $\ln(mktval)$ | $ceoten$ |
|---|---|---|---|---|
| $coeff$ | $-0.0521$ | $-0.0805$ | 0.0992 | 0.0188 |
| $Std.Error$ | 0.3052 | 0.0466 | 0.0588 | 0.0063 |
| $Prob$ | 0.8647 | 0.0856 | 0.0938 | 0.0033 |

R-squared=0.0657; F-statistic=4.0537; Prob(F-statistic)=0.0082.

a) Какие выводы можно сделать на основе этой регрессии?

b) Можно ли (экономически) интерпретировать эту модель регрессии?

11. Рассмотрим модель регрессии $y_t = \theta_0 + \theta_1 x_t + \varepsilon_t$, где для ошибок которой выполнены условия

$$E\varepsilon_i = 0, Var(\varepsilon_i) = \sigma^2, \varepsilon_i = \rho\varepsilon_{i-1} + \delta_i,$$

$$E\delta_i = 0, \text{ при } i \neq j \ cov(\delta_i, \delta_j) = 0, \ D(\delta_i) = \sigma_0^2$$

Найдите дисперсию остатков, ковариации остатков, выпишите ковариационную матрицу остатков и обратную к ней. Какой вероятностный смысл имеет параметр $\rho$ в ОЛММР с автокоррелированными остатками? Выпишите формулу для прогноза зависимой переменной.

12. Рассмотрим модель регрессии $y_t = \theta_0 + \theta_1 x_t + \varepsilon_t$, где для ошибок которой выполнены условия

$$E\varepsilon_t = 0, Var(\varepsilon_t) = \sigma^2, \varepsilon_t = 0,15\varepsilon_{t-1} + \delta_t$$

и $\delta_t$ удовлетворяет условиям Гаусса-Маркова. Пусть $\hat{\theta}_1$ есть МНК-оценка коэффициента $\theta_1$.

a. Сформулируйте условия Гаусса-Маркова для $\delta_t$.

b. Будет ли $\hat{\theta}_1$ несмещенной оценкой $\theta_1$? Состоятельной оценкой?

c. Будет ли $\hat{\theta}_1$ наилучшей оценкой коэффициента $\theta_1$ (с минимальной дисперсией)? Если нет, то предложите оценку коэффициента с меньшей дисперсией (более точную оценку).

13. Оценивается кривая Филлипса с ожиданиями, связывающая инфляцию с безработицей, безработицей с единичным лагом и с инфляцией с лагом один.
a. Опишите тест на автокорреляцию первого порядка ошибок этой модели регрессии и укажите проверяемую статистическую гипотезу.
b. Можно ли применять этот тест при небольших выборках?
c. Если тест отвергает нулевую гипотезу, то какие могут быть причины?

14. По квартальным данным за пять с половиной лет по методу наименьших квадратов была оценена функция спроса с учетом сезонности (факторы Spr, Summ, Fall - весна, лето, осень). Зависимая переменная: $\ln(Q)$

| | $const$ | $\ln(P)$ | $Spr$ | $Summ$ | $Fall$ |
|---|---|---|---|---|---|
| $coefficient$ | 4.4 | 0.64 | $-0.08$ | $-0.13$ | $-0.07$ |
| $Std.Error$ | 1.02 | 0.013 | 0.013 | 0.023 | 0.03 |

$$R^2 = 0.13;\ DW = 0,78$$

a. Можно ли для прогнозирования объема продаж использовать полученные оценки коэффициентов модели регрессии? Ответ обоснуйте.

b. Можно ли для проверки значимости регрессии "в целом"использовать стандартную F-статистику? Ответ обоснуйте.

15. По четырехнедельным историческим данным была оценена регрессионная модель зависимости спроса на мороженное consumption (в пинтах на человека) от среднего недельного дохода домохозяйства income, цены за пинту price и средней температуры temp (по Фаренгейту).Зависимая переменная: consumption; объем выборки 30.

| | $const$ | $income$ | $price$ | $temp$ |
|---|---|---|---|---|
| $coeff$ | 0.1973 | 0.0033 | $-1.0444$ | 0.0034 |
| $Std.Error$ | 0.2702 | 0.0011 | 0.8343 | 0.0004 |

$$R^2 = 0.7890;\ R^2_{adj} = 0.6866\ DW = 1.0211$$

a. Тестируйтесь ошибку модели регрессии на автокорреляцию первого порядка.

b. Какие можно сделать выводы?

16. По предприятиям энергетической отрасли была оценена производственная функция Кобба-Дугласа (K - капитал, L - труд) по данным за февраль 2006 года. Зависимая переменная: $\ln(Q)$; объем выборки 30

| $Y$ | $const$ | $\ln(K)$ | $\ln(L)$ |
|---|---|---|---|
| $coeff$ | 3.45 | 0.47 | 0.32 |
| $Prob$ | 0.0210 | 0.461 | 0.528 |

$$R^2 = 0.45;\ Prob(F - stat) = 0.00031;\ DW = 3.43$$

a) Проверьте значимость влияния капитала, значимость влияния труда, совместную значимость влияния труда и капитала.

b) Как можно объяснить результаты предыдущего пункта? Дайте не менее двух возможных объяснений.

17. По квартальным данным за 2002-2008 года была оценена регрессионная модель, связывающая количество вакансий ($y_t$), уровень безработицы ($u_t$) и объем ВНП ($z_t$) $\widehat{\ln y_t} = 3.4 + 0.3\ln(u_t) + 0.03\ln(z_t)$ $R^2 = 0.69$ $DW = 3.22$ Будут ли ошибки в модели регрессии автокоррелированны (первого порядка)? Если да, то автокорреляция положительная или отрицательная?

18. По квартальным данным за 2002 – 2007 года была оценена регрессионная модель зависимости объема производства $Q$ от объема инвестиций $INV$

$$\widehat{\ln Q}_t = 34.2 + 0.23\ln(INV_t) \quad R^2 = 0.59; \quad \text{DW} = 1.89$$

Будут ли ошибки в модели регрессии автокоррелированны (первого порядка)? Если да, то автокорреляция положительная или отрицательная?

19. * По квартальным данным за 2000 – 2005 года была оценена регрессионная модель, связывающая количество вакансий ($y_t$), уровень безработицы ($u_t$), объем ВНП ($z_t$) и численность населения ($pop_t$):t

$$\widehat{\ln y}_t = 3.4 + 0.3\ln(u_t) + 0,03\ln(z_t) + 0.003\ln(pop_t)$$
$$R^2 = 0.87 \quad \text{DW} = 1.02.$$

Будут ли ошибки в модели регрессии автокоррелированны (первого порядка)? Если да, то автокорреляция положительная или отрицательная?

20. По квартальным данным за 2003 – 2008 года была оценена регрессионная модель зависимости объема производства $Q$ от объема инвестиций $INV$ и цен на сырье $Price$

$$\widehat{\ln Q}_t = 34.2 + 0.23\ln(INV_t) - 0.3\ln(Price_t) \quad R^2 = 0.82 \quad \text{DW} = 0.97.$$

Будут ли ошибки в модели регрессии автокоррелированны (первого порядка)? Если да, то автокорреляция положительная или отрицательная?

21. Рассмотрим статическую кривую Филлипса (static Phillips curve), связывающую темпы инфляции $inf$ и уровень безработицы $unem$:

$$\widehat{inf}_t = 1.34 + 0.34unem_t \quad R^2 = 0.078 \quad n = 50.$$

a. Опишите двухшаговую процедуру тестирования ошибок регрессии на автокорреляцию второго порядка и напишите спецификацию вспомогательной модели регрессии.

b. Пусть во вспомогательной регрессии $R_0^2 = 0.32$. Можно ли сделать вывод об автокоррелированности (второго порядка) ошибок регрессии в кривой Филлипса? Уровень значимости 1%.

c. Можно ли в данной модели регрессии для тестирования ошибок регрессии на автокорреляцию первого порядка использовать тест Durbin – Watson ?

22. Рассмотрим макроэкономическую кейнсианскую функцию потребления без автономного потребления $C_t = \theta_1 Y_t + \theta_2 T_t + \theta_3 r_t + \varepsilon_t$, где $C_t$ - потребление, $Y_t$ - национальный доход, $T_t$ - налоговые поступления, $r_t$ - ставка рефинансирования. Модель оценивается по историческим данным и предполагается, что объясняющие переменные экзогенны. Дайте описание теста на автокоррелированность первого порядка ошибок этой модели регрессии.

23. * Анализируется следующая модель зависимости объема произведенной на предприятии в течение квартала $t$ продукции $y_t$ (млн.руб.) от оценочной стоимости его основных фондов $x_t$ (млн.руб.):

$$\begin{cases} y_t = \theta x_t + \varepsilon_t \\ \varepsilon_t = \rho \varepsilon_t + \delta_t \\ E\delta_t = 0,\ E(\delta_t \cdot \delta_{t+k}) = \begin{cases} \sigma_0^2 \text{ при } k = 0, \\ 0 \text{ при } k \neq 0. \end{cases} \end{cases}$$

в которой $\theta$, $\rho$ ($|\rho|<1$) и $\sigma_0^2$ – некоторые неизвестные параметры.

Реализация обобщенного МНК, примененного к исходным статистическим данным $(x_1, y_1),\ (x_2, y_2), \ldots (x_n, y_n)$, дала следующие ОМНК-оценки параметров:

$\hat{\rho} = 0,707$; $\hat{\theta} = 1,12$; $\hat{\sigma}^2 = 36,0$ (где $\sigma^2 = D\varepsilon_t$); $\hat{\sigma}_0^2 = 18,0$; $\hat{\sigma}_{\hat{\theta}}^2 = 0,04$ (где $\sigma_{\hat{\theta}}^2 = D\hat{\theta}$); $\hat{\varepsilon}_n = y_n - \hat{\theta} x_n = 3,5$ (млн.руб.).

Оценка параметра $\theta$ по обычному МНК: $\hat{\theta}_{\text{МНК}} = 1,02$.

Требуется:

1) Объяснить, по каким формулам и в какой последовательности получены ОМНК-оценки параметров $\theta$, $\sigma^2$, $\rho$, $\sigma_0^2$, $\sigma_{\hat{\theta}}^2$, а также МНК-оценки $\hat{\theta}_{\text{МНК}}$ параметра $\theta$.

2) Дать вероятностную интерпретацию параметра $\rho$.

3) Ответить, на вопрос, будет ли МНК-оценка $\hat{\theta}_{\text{МНК}}$ состоятельной и несмещенной оценкой параметра $\theta$ и почему ОМНК-оценка $\hat{\theta}$ предпочтительнее оценки $\hat{\theta}_{\text{МНК}}$?

4) Проверить гипотезу о статистически незначимом влиянии величины основных фондов на объем произведенной продукции (при уровне значимости критерия $\alpha = 0,05$).

5) Вычислить наилучший (в смысле среднего квадрата ошибки) прогноз ожидаемой величины объема продукции, произведенной в $(n+1)$-м квартале, если стоимость основных фондов в этом квартале оценивается величиной $x_{n+1} = 60$ млн. руб.

# Глава 5

# Модели с дискретными зависимыми переменными

## 5.1. Дискретная зависимая переменная

В первых четырёх главах изучалась стандартная линейная регрессия, описывающая зависимость непрерывной количественной переменной $y$ от набора объясняющих переменных $\mathbf{x}_i' = \begin{pmatrix} 1 & x_{i1} & x_{i2} & \dots & x_{ik} \end{pmatrix}$, среди которых могли быть и бинарные переменные. Однако ряд практических задач эконометрического анализа приводят к необходимости строить и статистически оценивать модели зависимости **дискретного** результирующего показателя $y$ от набора объясняющих переменных $\mathbf{x}_i'$. То есть приходится рассматривать бинарную зависимую переменную, например, если результирующим фактором является качественный признак (гендерный фактор), а в задаче принятия решений может ставиться вопрос о выдаче кредита. Отметим, что дискретные переменные могут быть *порядковыми* (*ординальными*) и *номинальными*.

**Дискретная порядковая** (**ординальная**) **переменная** может принимать только дискретные значения, поддающиеся естественному упорядочению. Например, *уровень образования индивидуума* (высшее, среднее специальное, среднее и т.п.). Обратите внимание, что мы могли бы учитывать образование в качестве *количественной* переменной (число лет обучения), однако для нас решающее значение имеет дискретная природа признака.

**Дискретная номинальная переменная** может принимать только дискретные «значения» (метки), *не поддающиеся упорядочению*. Например, *тип места проживания* (мегаполис, город, поселок городского типа, сельская местность), *профессия опрашиваемого* и т.п.

Ниже мы рассмотрим две регрессии с бинарной зависимой переменной: линейную и (“сильно”) нелинейную.

## 5.2. LPM-модель, спецификация и свойства

Итак, представим себе, что результирующий показатель $y$, «поведение» которого существенно зависит от объясняющих переменных $\mathbf{X} = (1, x_1, x_2, \ldots, x_k)$, является бинарной переменной. В этом случае вектор $Y = (y_1, y_2, \ldots, y_n)'$ исходных статистических данных зависимой переменной будет содержать только два значения: 0 или 1. Как построить регрессионную модель, описывающую зависимость $y$ от $\mathbf{X}$ в данном случае?

Для исследования такой статистической связи между $y$ и $\mathbf{X}$ строят некоторую специальную регрессионную модель зависимости *вероятности* $P\{y = 1 \mid \mathbf{X}\}$ от линейной формы $\mathbf{x}_i'\boldsymbol{\beta}$.

Одним из вариантом такой модели является линейная вероятностная модель LPM (Linear Probability Model). Она представляет собой линейную регрессию с бинарной зависимой переменной $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$, где $y_i$ принимает значения 1 или 0. А на ошибки накладываются следующие условия

- $E(\varepsilon_i|\mathbf{X}) = 0$ — условие экзогенности регрессоров;
- $\mathrm{cov}(\varepsilon_i, \varepsilon_j|\mathbf{X}) = 0$ при $i \neq j$ – отсутствие серийной корреляции.

Из первого условия вытекает, что $E(y_i|\mathbf{X}) = \mathbf{x}_i'\boldsymbol{\beta}$. Так как $y_i$ есть бинарный фактор, то

$$y_i|\mathbf{X} = \begin{cases} 1, & P(y_i = 1|\mathbf{X}), \\ 0, & P(y_i = 0|\mathbf{X}). \end{cases} \quad \Rightarrow E(y_i|\mathbf{X}) = P(y_i = 1|\mathbf{X})$$

Следовательно, в LPM-модели $P(y_i = 1|\mathbf{X}) = \mathbf{x}_i'\boldsymbol{\beta}$

### 5.2.1. LPM-модель: прогноз

В стандартной линейной регрессии оптимальный (в смысле среднего квадратического) прогноз есть среднее значение

$$\hat{y} = \hat{y}|\mathbf{x} = \mathbf{x}'\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

т.е. прогнозируем значение зависимой переменной.

LPM-модель: оптимальный прогноз также среднее значение

$$\hat{y} = \hat{P}(y = 1|\mathbf{x}) = \hat{y}|\mathbf{x} = \mathbf{x}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k,$$

но прогноз даётся для вероятности того, что зависимая переменная примет значение 1.

### 5.2.2. LPM-модель: интерпретация коэффициентов

Чтобы проинтерпретировать коэффициенты такой модели, вычислим частную производную по переменной $x_j$:

$$P(y=1|\mathbf{X}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \Longrightarrow \frac{\partial P(y_i = 1|\mathbf{X})}{\partial x_j} = \beta_j.$$

Таким образом, $\beta_j$ есть (среднее) предельное значение $P(y_i = 1|\mathbf{X})$ по $x_j$. Грубо говоря, коэффициент $\beta_j$ показывает, насколько изменится вероятность того, что зависимая переменная примет значение 1, при увеличении фактора $x_j$ на единицу.

### 5.2.3. LPM-модель: гетероскедастичность ошибки

Так как $y_i$ есть бинарный фактор, то это накладывает сильное условие на ошибку регрессии $\varepsilon_i (= y_i - \mathbf{x}_i'\boldsymbol{\beta})$

$$y_i|\mathbf{X} = \begin{cases} 1, & P(y_i = 1|\mathbf{X}), \\ 0, & P(y_i = 0|\mathbf{X}). \end{cases} \quad \Leftrightarrow \quad \varepsilon_i|\mathbf{X} = \begin{cases} 1 - \mathbf{x}_i'\boldsymbol{\beta}, & P(y_i = 1|\mathbf{X}), \\ -\mathbf{x}_i'\boldsymbol{\beta}, & P(y_i = 0|\mathbf{X}). \end{cases}$$

Откуда получается, что $\mathrm{Var}(\varepsilon_i|\mathbf{X}) = E(\varepsilon_i^2|\mathbf{X}) = \mathbf{x}_i'\boldsymbol{\beta}(1 - \mathbf{x}_i'\boldsymbol{\beta})$. То есть в LPM-модели ошибка имеет бинарное (негауссово) распределение и гетероскедастична.

### 5.2.4. Особенности модели

Подводя итог, из особенностей модели стоит отметить следующие:

- модель является линейной;
- распределение ошибки не является нормальным;
- ошибки гетероскедастичны.

К счастью, состоятельным методом оценивания по-прежнему остаётся метод наименьших квадратов. При тестировании гипотез необходимо использовать робастные тестовые статистики, устойчивые к гетероскедастичности.

*Замечание.* Статистические выводы применимы только при больших выборках! Это следствие негауссовости и гетероскедастичности. А также есть проблема с интерпретацией прогноза как вероятности, если прогноз больше 1 или меньше 0. Тогда нельзя интерпретировать прогноз как вероятность! Другими словами, есть проблемы с “адекватностью” модели.

## 5.3. Нелинейные модели

Так как описание вероятности *непосредственно линейной функцией* $\mathbf{x}'\beta$ не всегда корректно, потому что в этом случае значения *предсказанной* вероятности могут не находиться в пределах от 0 до 1, то для моделирования значений $P\{y = 1 \mid \mathbf{X}\}$ подбирают функции, область значений которых определяется отрезком $[0; 1]$, а линейная форма $\mathbf{x}'\boldsymbol{\beta}$ играет роль аргумента этой функции, то есть

$$P\{y = 1 \mid \mathbf{X}\} = F(\mathbf{x}'\boldsymbol{\beta}), \tag{1}$$

причем функции $F(\cdot)$ должны удовлетворять свойствам фукнции распределения.

Такие модели называют обычно *моделями бинарного выбора.*

### 5.3.1. Латентная переменная

Будем считать, что мы рассматриваем задачу принятия решений, в которой значения зависимой переменной $y_i$ соответствует выбору лица, принимающего решение, между двумя состояниями $i$-го объекта, характеризуемого значениями $x_{i1}, x_{i2}, \ldots x_{ik}$ объясняющих переменных. Выбор же производится, в зависимости от величины некоторой **латентной** (то есть не поддающейся непосредственному измерению) **характеристики** $y_i^*$, которую можно считать линейной функцией от $\mathbf{X}$ с точностью до случайной остаточной компоненты $\varepsilon_i$, то есть

$$y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \tag{2}$$

где, как и прежде, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$, $\mathbf{x}_i' = (1, x_{i1}, \ldots, x_{ik})'$, а $\mathbf{E}\varepsilon_i = 0$ (предположения о ковариационной матрице вектора остатков $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$ определяются видом распределения вектора остатков $\boldsymbol{\varepsilon}$).

"Положительное" решение ($y = 1$) принимается, если значение латентной переменной $y^*$ больше "порогового" значения $a$.

Здесь мы будем считать, что распределение ошибки $F(\varepsilon)$ симметрично, т.е. $F(-\varepsilon_i) = 1 - F(\varepsilon_i)$. Также можно считать, что $a = 0$, включив $a$ в $\beta_0$.

Тогда

$$\begin{aligned} P(y_i = 1|\mathbf{X}) = P(y_i^* > 0|\mathbf{X}) = P(\mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i > 0|\mathbf{X}) = \\ = P(\varepsilon_i > -\mathbf{x}_i'\boldsymbol{\beta}|\mathbf{X}) = P(\varepsilon_i < \mathbf{x}_i'\boldsymbol{\beta}|\mathbf{X}) = F(\mathbf{x}_i'\boldsymbol{\beta}) \end{aligned} \tag{5.1}$$

Получаем нелинейную модель

$$P(y_i = 1|\mathbf{X}) = F(\mathbf{x}_i'\boldsymbol{\beta}), \tag{3}$$

где $F(\cdot)$ – некоторая функция распределения.

В прикладных задачах, как правило, рассматриваются два распределения: гауссово и логистическое. Полагая распределение $F(\cdot)$ стандартным нормальным, мы получаем *пробит-модель* бинарного выбора, а при логистическом распределении $F(\cdot)$ мы имеем *логит-модель.*

### 5.3.2. Интерпретация коэффициентов

Для того чтобы проинтерпретировать коэффициенты, посмотрим на предельные значения в модели (3)

$$\frac{\partial P(y=1|\mathbf{X})}{\partial x_j} = p(\mathbf{x}_i'\boldsymbol{\beta})\beta_j$$

где $p(\cdot) = F'(\cdot) \geq 0$ – плотность распределения.
Из этого уравнения можно сделать следующие выводы:

- предельные значения непостоянны (нелинейно зависят от $\mathbf{x}$);
- коэффициенты не имеют количественной интерпретации;
- имеет интерпретацию знак коэффициента;
- можно легко интерпретировать отношение коэффициентов как отношение предельных значений.

### 5.3.3. Logit- и Probit-модель

Линейная регрессия для латентной переменной:

$$y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$$

- Пусть ошибка имеет гауссово распределение и гомоскедастична:

  $$\varepsilon_i|\mathbf{X} \sim N(0;\sigma^2) \Leftrightarrow F(\varepsilon) = \Phi\left(\frac{\varepsilon}{\sigma}\right),$$

  где $\Phi(\varepsilon)$ – функция стандартного распределения.

  Обратная функция $\Phi^{-1}(p)$, $0 < p < 1$ называется probit-функция и обозначается

  $$\text{probit}(p) = \Phi^{-1}(p).$$

  Probit-модель:

  $$P(y_i = 1|\mathbf{X}) = \Phi(\mathbf{x}_i'\boldsymbol{\beta}) = \Phi(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)$$

  $$\text{probit}(P(y_i = 1|\mathbf{X})) = \mathbf{x}_i'\boldsymbol{\beta}$$

- Функция логистического распределения имеет вид

  $$\Lambda(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$$

Если в модели бинарного выбора рассмотреть функцию $F(\cdot) = \Lambda(\cdot)$, то модель будет называться logit-моделью. Обратная функция $\Lambda^{-1}(p)$, $0 < p < 1$ называется logit-функция, причём

$$\Lambda^{-1}(p) = \ln\left(\frac{p}{1-p}\right)$$

В силу последнего равенства коэффициенты $\beta_j$ получают прозрачную интерпретацию. Изменение на единицу фактора $x_j$ ведёт к изменению отношения шансов $\frac{p}{1-p}$ в $e^{\beta_j}$ раз.

*Замечание.* "Хвосты" логистического распределения "тяжелее" "хвостов" нормального распределения.

### 5.3.4. Проблема оценивания. Метод максимального правдоподобия

Рассмотрим теперь вопрос о том, как (состоятельно!) оценить на статистических данных параметры нелинейных probit – и logit-моделей. Для этого воспользуемся общим методом оценки параметров нелинейных моделей – методом максимального правдоподобия.

Он состоит в следующем:

- выписывается так называемая функция правдоподобия $\mathcal{L} = \mathcal{L}(\boldsymbol{\beta}|y, \mathbf{X})$;
- оценки параметров находятся из условия $\max\limits_{\boldsymbol{\beta}} \mathcal{L}$ или $\max\limits_{\boldsymbol{\beta}} \ln \mathcal{L}$

Отметим, что всегда $0 < \mathcal{L} < 1$ и $\ln \mathcal{L} < 0$. Полученные таким методом оценки будем обозначать $\widehat{\boldsymbol{\beta}}_{\text{мп}}$ или $\widehat{\boldsymbol{\beta}}_{ML}$, а значение функции $\ln \hat{\mathcal{L}} = \ln \mathcal{L}(\widehat{\boldsymbol{\beta}}_{\text{мп}}) = \max\limits_{\boldsymbol{\beta}} \ln \mathcal{L}(\boldsymbol{\beta})$

Приведём общий вид функции правдоподобия и необходимых условий для нахождения её максимального значения. Учитывая взаимную статистическую независимость наблюдений $y_1, y_2, \ldots, y_n$ и факт, что $P\{y_i = 1|\mathbf{X}\} = F(\mathbf{x}_i'\boldsymbol{\beta})$ и, соответственно, $P\{y_i = 0|\mathbf{X}\} = 1 - F(\mathbf{x}_i'\boldsymbol{\beta})$, имеем:

$$\mathcal{L}(\boldsymbol{\beta}|y; \mathbf{X}) = \prod_{i:y_i=1} F(\mathbf{x}_i'\boldsymbol{\beta}) \times \prod_{i:y_i=0} (1 - F(\mathbf{x}_i'\boldsymbol{\beta})) =$$
$$= \prod_{i=1}^{n} (F(\mathbf{x}_i'\boldsymbol{\beta}))^{y_i} (1 - F(\mathbf{x}_i'\boldsymbol{\beta}))^{1-y_i}.$$

Перейдём к *логарифмической* функции правдоподобия $l = \ln \mathcal{L}$

$$\ln \mathcal{L}(\boldsymbol{\beta}|y; \mathbf{X}) = \sum_{i:y_i=1} \ln F(\mathbf{x}_i'\boldsymbol{\beta}) + \sum_{i:y_i=0} \ln(1 - F(\mathbf{x}_i'\boldsymbol{\beta})).$$

Дифференцируя ее по $\boldsymbol{\beta}$, получаем систему уравнений (условий 1-го порядка существования экстремума у функции $l$) для определения оценок $\widehat{\boldsymbol{\beta}}_{\text{ммп}}$:

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left[ \frac{y_i - F(\mathbf{x}_i'\boldsymbol{\beta})}{F(\mathbf{x}_i'\boldsymbol{\beta})(1 - F(\mathbf{x}_i'\boldsymbol{\beta}))} p(\mathbf{x}_i'\boldsymbol{\beta}) \right] \mathbf{x}_i' = \mathbf{0}_{k+1}, \tag{4}$$

где $p(\cdot) = F'(\cdot)$ – функция плотности распределения, задаваемого функцией распределения $F(\cdot)$.

Таким образом, приходится решать нелинейную задачи оптимизации, которую явно решить, как правило, невозможно. Эта задача максимизации функции правдоподобия численно реализована в эконометрических программах.

Отметим, что условия 1-го порядка являются лишь *необходимыми* условиями существования экстремума функции $l = \ln \mathcal{L}$, но так как матрица Гессе вторых производных функции $l$ по $\boldsymbol{\beta}$ является отрицательно определенной, если объясняющие переменные в матрице $\mathbf{X}$ неколлинеарны, а значит, решение системы (4) действительно определяют точку локального максимума по $\boldsymbol{\beta}$ функции $l$.

### 5.3.5. Общее качество модели

Для введения коэффициента, характеризующего качество модели «в целом», рассмотрим регрессию без объясняющих переменных (с условием $\beta_1 = \ldots = \beta_k = 0$, то есть фактически схему Бернулли). В этом случае

$$P(y_i = 1|\mathbf{X}) = F(\beta_0) = p;\ p \in (0; 1)$$

Логарифмическая функция правдоподобия

$$\ln \mathcal{L}_0(p|y) = \sum_{y_i=1} \ln p + \sum_{y_i=0} \ln(1-p) = n_1 \ln p + (n - n_1) \ln(1-p)$$

($n_1$ – число единиц в выборке $y$). Тогда

$$\max(\ln \mathcal{L}_0(p|y)) \Leftrightarrow \hat{p} = \frac{n_1}{n}$$

Обозначив $\ln \hat{\mathcal{L}}_0 = \ln \mathcal{L}_0(\hat{p})$ и учитывая неравенство $\ln \hat{\mathcal{L}}_0 \leq \ln \hat{\mathcal{L}} < 0$, можем ввести

**Определение.** Коэффициент псевдо-$R^2$ или $R^2$ МакФаддена определяется как

$$pseudo - R^2 = R^2_{McFadden} = 1 - \frac{\ln \hat{\mathcal{L}}}{\ln \hat{\mathcal{L}}_0}$$

Псевдо-$R^2$ имеет следующие свойства:

1. $0 \leq R^2_{McFadden} \leq 1$

2. $R^2_{McFadden} = 0 \Leftrightarrow \ln\hat{\mathcal{L}} = \ln\hat{\mathcal{L}}_0 = \ln\mathcal{L}, \beta_1 = \ldots = \beta_k = 0$, т.е. регрессоры "не улучшают" прогноз зависимой переменной.

Поэтому он может быть использован как показатель "качества подгонки" модели регрессии под данные, а также для сравнения моделей.

*Замечание.* Псевдо-$R^2$ можно использовать как показатель "качества подгонки" в любой регрессии, оцениваемой методом максимального правдоподобия (фактически так можно оценить любую нелинейную регрессию).

### 5.3.6. Проверка гипотез

Говоря о тестирование "простых" гипотез, будем учитывать, что все рассматриваемые ниже статистические выводы (тесты) обоснованы только при больших выборках (асимптотически).

#### Значимость коэффициента

Метод максимального правдоподобия вычисляет оценки коэффициентов $\hat{\beta}_j$ и стандартные ошибки коэффициентов $s_j$. Для проверки значимости коэффициента тестируется гипотеза

$$H_0 : \beta_j = 0$$

против альтернативы

$$H_1 : \beta_j \neq 0.$$

Тестовая статистика $z = \frac{\hat{\beta}_j}{s_j}$, где $z_{cr}(\alpha)$ – это критическое значение стандартного нормального распределения. Вывод делается стандартно: при $|z| > z_{cr}$ отвергаем $H_0$, говорим, что коэффициент статистически значим (влияние фактора $x_j$ значимо).

#### Значимость регрессии "в целом"

Для проверки значимости регрессии "в целом" тестируется гипотеза

$$H_0 : \ \beta_1 = \ldots = \beta_k = 0$$

Смысл гипотезы заключается в том, что все включённые в модель регрессоры в совокупности не влияют на зависимую переменную ("ничего не объясняют").

Тестовой статистикой является $LR$-статистика:

$$LR = 2(\ln\hat{\mathcal{L}} - \ln\hat{\mathcal{L}}_0) > 0,$$

которая асимптотически имеет $\chi^2$-распределение. При $LR > \chi^2_{cr}$ отвергаем $H_0$ и говорим, что регрессия "в целом" значима. Альтернативно с помощью $p$-значения.

**Совместная значимость**

Тестирование совместной значимости нескольких коэффициетов: тестируем, например, гипотезу, что первые $J$ коэффициентов равны нулю

$$H_0: \ \beta_1 = \ldots = \beta_J = 0,$$

то есть проверяем гипотезу о том, что первые $J$ факторов не влияют на зависимую переменную.

Для этого оценивается регрессия с ограничением (без первых $J$ факторов), $\ln \hat{\mathcal{L}}_r$ – её логарифмическая функция правдоподобия. Тестовая $LR$-статистика: $LR = 2(\ln \hat{\mathcal{L}} - \ln \hat{\mathcal{L}}_r) > 0$, имеющая асимптотическое распределение $\chi^2_J$. При $LR > \chi^2_{cr}$ отвергаем $H_0$ и делаем вывод о совместной значимости этих коэффициентов.

*Замечание.* Альтернативным способом тестирования гипотез о (линейных) ограничениях на коэффициенты является статистка Вальда $W$. Для неё нет простой формулы, но её значения автоматически вычисляются в программах.

Ещё раз подчеркнём, что все рассмотренные статистические выводы (тесты) обоснованы только при больших выборках, потому что точные распределения тестовых статистик $z$, $LR$ и $W$ неизвестны.

### 5.3.7. Прогнозирование

Обычный прогноз даёт прогнозное значение вероятности, что $y = 1$:

$$\hat{P}(y = 1|x) = F(\mathbf{x}_i' \hat{\boldsymbol{\beta}}), \ F(u) = \Phi(u); \Lambda(u)$$

Чтобы получить прогноз $\hat{y}$ (0 или 1), выбирается "пороговое" значение $c \in (0; 1)$, и прогноз имеет вид

$$\hat{y} = \begin{cases} 1, \ \hat{P}(y_i = 1|\mathbf{X}) \geq c, \\ 0, \ \hat{P}(y_i = 0|\mathbf{X}) < c. \end{cases}$$

Часто выбирают $c = 0.5$, что также даёт показатель "качества подгонки": доля правильных прогнозов на выборочных данных при $c = 0.5$.

### 5.3.8. Информационные критерии. Выбор модели

При выборе наиболее подходящей модели часто используют основные информационные критерии.

Выделим три основных показателя информационных критериев:

1. Akaike

$$AIC = -\frac{2\ln\hat{\mathcal{L}}}{n} + \frac{2k}{n}$$

2. Schwarz или Bayesian

$$SIC = BIC = -\frac{2\ln\hat{\mathcal{L}}}{n} + \frac{k\ln n}{n}$$

3. Hannan–Quinn

$$HQ = -\frac{2\ln\hat{\mathcal{L}}}{n} + \frac{2k\ln(\ln n)}{n}$$

Все они автоматически вычисляются в программах. Для сравнения нескольких альтернативных моделей с одинаковой зависимой переменной (probit/logit), но с разным набором регрессоров, вычисляется показатель инф.критерия для каждой модели и выбирается модель с минимальным показателем.

*Замечание.* Разные информационные критерии могут выбирать разные модели! Для выбора модели в этом случае можно использовать критерий $\max(pseudo{-}R^2)$.

В заключение отметим, что игнорирование гетероскедастичности (неоднородности данных) приводит к несостоятельности оценок коэффициентов, неверному тестированию гипотез и неправильным прогнозам. Это следствие неверной спецификации функции правдоподобия. Ещё одной причиной несостоятельности оценок является избыточность набора объясняющих переменных.

*Замечание.* **Модели множественного выбора** обобщают модели бинарного выбора на случай *нескольких* возможных значений дискретной зависимой переменной $y$ и, соответственно, позволяют строить и статистически оценивать зависимости вида $P(y_i = j|\mathbf{X}) = \varphi_j(\mathbf{x}_i'\boldsymbol{\beta}(j))$.

## 5.4. Задачи

1. Для случайной величины с распределением

| значения | 0 | 1 | 2 | 4 |
|---|---|---|---|---|
| вероятности | $0,5+\boldsymbol{\beta}$ | $0,1-\boldsymbol{\beta}$ | $0,2$ | $0,2$ |

получена выборка: 1, 4, 2, 2, 0, 1.

а) Выпишите функцию правдоподобия и ее логарифм.

б) Вычислите оценку максимального правдоподобия.

2. Дана выборка из пуассоновского распределения с параметром $\boldsymbol{\beta}$: 1, 4, 3, 2, 3, 0, 1, 1, 0, 5.

   а) Выпишите функцию правдоподобия и ее логарифм.

   б) Вычислите оценку максимального правдоподобия.

3. Методом максимального правдоподобия найдите вероятность успеха $\boldsymbol{\beta}$ в схеме испытаний Бернулли. Докажите, что эта оценка является несмещённой и состоятельной.

4. Для выборки $x_1;\ \ldots x_n$ из равномерного распределения на отрезке $[a;b]$ найдите оценки максимального правдоподобия для $a$ и $b$.

5. Для выборки размера $n$ из нормального распределения $N(\mu, \sigma^2)$ найдите оценки максимального правдоподобия для параметров $\mu$ и $\sigma^2$.

6. (задача 10.3 из [8]) По выборке $x_1, x_2 \ldots, x_n$ в случае равномерного распределения на отрезке $[0;\beta]$ методом максимального правдоподобия найти оценку параметра $\beta$. Найдите её смещение. Является ли оценка асимптотически несмещённой? А состоятельной?

7. (задача 10.6 из [8]) По выборке $x_1, x_2 \ldots, x_n$ в случае равномерного распределения на отрезке $[\beta; 2\beta]$ методом максимального правдоподобия найти оценку параметра $\beta$. Найдите её смещение. Является ли оценка асимптотически несмещённой? А состоятельной?

8. Рассмотрим LPM модель

$$\widehat{Male} = 0.37 + \underset{[0.011]}{0.02}\,Educ - \underset{[0.046]}{0.1}\,City \quad n = 700$$

   зависимости пола от уровня образования $Educ$ и места жительства $City$ (в квадратных скобках указаны стандартные ошибки в форме Уайта). Почему при оценке модели были использованы стандартные ошибки в форме Уайта? Кратко поясните. Проверьте значимость влияния каждого из факторов (уровень значимости 1%). Вычислите предсказанное значение для человека с образованием 15 лет и живущего в городе. Как следует интерпретировать предсказанное значение?

9. Рассмотрим probit-модель

$$\mathsf{P}(y = 1) = \Phi(0.45 + \underset{(0.2)}{0.3}x_1 + \underset{(1.2)}{2.1}x_2 - \underset{(0.2)}{0.3}x_3 + \underset{(0.9)}{1.2}x_4)$$
$$Log.likelihood = -1202.3 \quad Restr.likelihood = -1382.2 \quad n = 1050$$

   Проверте значимость коэффициентов регрессии. Проверьте значимость регрессии «в целом».

10. Рассмотрим *Logit*-модель, оцененную в EViews

Dependent variable: crime86
Sample 1 2725

| Variable | Coefficient | Std. Error | z-Statistic | Prob. |
|---|---|---|---|---|
| C | −0.655477 | 0.119205 | −5.498726 | 0.0000 |
| pcnv | −0.924175 | 0.124452 | −7.425948 | 0.0000 |
| avgsen | 0.018503 | 0.035298 | 0.524185 | 0.6001 |
| totime | −0.010810 | 0.028172 | −0.383710 | 0.7012 |
| ptime86 | −0.123048 | 0.031627 | −3.890563 | 0.0001 |
| qemp86 | 0.066483 | 0.045187 | 1.471275 | 0.1412 |
| inc86 | −0.009057 | 0.001290 | −7.021432 | 0.0000 |
| durat | 0.019617 | 0.010214 | 1.920591 | 0.0548 |
| black | 0.776393 | 0.118454 | 6.554379 | 0.0000 |
| hispan | 0.498160 | 0.109949 | 4.530833 | 0.0000 |
| born60 | 0.018356 | 0.094461 | 0.194319 | 0.8459 |

$$Log.\ likelihood = -1481.458 \quad Restr.\ likelihood = -1608.184$$

- Какие из факторов значимо, а какие незначимо влияют на зависимую переменную? Уровень значимости 10%.
- Вычислите $R^2_{McFadden}$
- Тестируйте значимость регрессии «в целом» при уровне значимости 1%.
- Была оценена *Logit*-модель без факторов black & hispan и получено значение логарифма функции правдоподобия $\ln \widetilde{L} = -1506.727$. Тестируйте совместную значимость факторов black & hispan при уровне значимости 5%.

11. (задача 12.5 из [8]) Рассмотрим модель бинарного выбора $P(y_t = 1) = F(\alpha + \beta d_t)$, где $d$ - фиктивная переменная (принимающая значения 0 или 1). Ниже представлены результаты 100 наблюдений:

| | $y = 0$ | $y = 1$ |
|---|---|---|
| $d = 0$ | 20 | 32 |
| $d = 1$ | 36 | 12 |

А) Оцените параметры $\alpha$, $\beta$, используя логит-модель. Проверьте гипотезу $H_0: \ \beta = 0$.

Б) Повторите п. А для пробит-модели. Изменятся ли ваши выводы?

12. (задача 12.8 из [8]) В таблице ниже приведены 100 наблюдений бинарных

переменных $(x,\ y,\ z)$. Количество наблюдений

| Число наблюдений | $y$ | $x$ | $z$ |
|---|---|---|---|
| 12 | 0 | 0 | 0 |
| 10 | 0 | 0 | 1 |
| 20 | 0 | 1 | 0 |
| 0 | 0 | 1 | 1 |
| 8 | 1 | 0 | 0 |
| 28 | 1 | 0 | 1 |
| 22 | 1 | 1 | 0 |
| 0 | 1 | 1 | 1 |

В рамках логит-модели $P(y=1)=\Lambda(\alpha+\beta x+\gamma z)$ протестируйте значимость влияния $z$ на $y$.

13. (задача 12.11 из [8]) Дано $n_1+n_2+n_3$ наблюдений переменных $x$ и $y$. Известно, что для $n_1$ наблюдений $y=1$ и $x=1$, для $n_2$ наблюдений $y=0$ и $x=1$, для $n_3$ наблюдений $y=0$ и $x=0$. Покажите, что как для $logit-$, так и для $probit-$ модели уравнения правдоподобия не имеют решения.

14. * Рассматривается проблема *дистанционного экспресс-анализа* финансово-экономического состояния банка, а именно, **задача построения и оценки показателя надежности банка на базе значений ограниченного числа его наиболее информативных балансовых показателей**.

Специальный анализ показал, что к таким балансовым переменным могут быть отнесены: $x^{(1)}$ – доля суммы пассивов в валюте баланса банка; $x^{(2)}$ – доля кредитов экономике в сумме работающих активов и $x^{(3)}$ – отношение балансовой прибыли к балансовым убыткам.

С целью информационного обеспечения решения этой задачи проведено обследование 500 наиболее крупных (по сумме пассивов и величине капитала) банков, в результате которого были получены следующие данные:

$$(x_i^{(1)}, x_i^{(2)}, x_i^{(3)}, y_i) \quad i=1,2,\ldots,500, \qquad (*)$$

где $y_i=1$, если $i$-й банк отнесен к категории проблемных, и $y_i=0$, если $i$-й банк отнесен к категории надежных. Требуется:

1) Предложить метод (модель), с помощью которого можно ввести и оценить по значениям $(x^{(1)}, x^{(2)}, x^{(3)})$ показатель надежности банка.

2) Подробно описать процедуру эконометрического анализа предложенной Вами модели на основании исходных данных вида $(*)$, в частности:

– опишите функцию правдоподобия имеющихся в Вашем распоряжении наблюдений $(*)$ в рамках выбранной Вами модели, дайте обоснование предложенного вида функции;

– в результате решения какой оптимизационной задачи вы получите оценки параметров Вашей модели?

3) Ответить на вопрос, существуют ли другие варианты решения поставленной задачи и если да, то какие именно?

15. Рассмотрим probit-модель

$$\mathsf{P}(y=1) = \Phi(0.45 + \underset{(0.2)}{0.3}x_1 + \underset{(1.2)}{2.1}x_2 - \underset{(0.2)}{0.3}x_3 + \underset{(0.9)}{1.2}x_4)$$
$$Log.likelihood = -1202.3 \quad Restr.likelihood = -1382.2 \quad n = 1050$$

Проверте значимость коэффициентов регрессии. Проверьте значимость регрессии «в целом».

16. Пусть $City$ – бинарная переменная, равная 1 если человек живет в городе. Рассмотрим logit-модель ($educ$ – уровень образования, $age$ – возраст, $female$ – пол человека)

Зависимая переменная: $City$; объем выборки 1023

| | $const$ | $educ$ | $age$ | $female$ |
|---|---|---|---|---|
| coefficient | 0.02 | 0.004 | 0.0018 | 0.04 |
| Std. Error | 0.13 | 0.0013 | 0.0002 | 0.023 |

$$Log.\ likelihood = -1281 \quad Restr.\ likelihood = -1508$$

Уровень значимости 5%.

- Запишите уравнение регрессии и формулу для logit-функции.
- Дайте интерпретацию коэффициентов регрессии.
- Значимо ли влияние уровня образования?
- Вычислите $R^2_{McFadden}$
- Тестируйте значимость регрессии «в целом» и сформулируйте проверяемую статическую гипотезу.
- Как интерпретируются предсказанные значения в этой модели регрессии?
- Для мужчины возраста 30 лет и с образованием 15 лет вычислите предсказанное значение и дайте его интерпретацию.
- Укажите формулу для предельных значений по $educ$ и по $age$.
- Для мужчин возраста 25 лет и 35 лет и с образованием 15 лет (каждый) вычислите предельные значения по $age$ и дайте их интерпретацию.
- Пусть $LR$-статистика для проверки совместной значимости факторов $age$ и $female$ равна 13.5. Проверьте совместную значимость этих факторов и сформулируйте проверяемую статистическую гипотезу.

17. Рассмотрим probit-модель (объем выборки 920)

$$\mathsf{P}(Male = 1) = \Phi\left(0.02 + \underset{(0.009)}{0.013}educ - \underset{(0.046)}{0.1}\ city - \underset{(0.011)}{0.029}age - \underset{(0.07)}{0.22}unem\right)$$

зависимости пола от уровня образования $educ$, места жительства $city$, возраста $age$ и статуса работающий/безработный $unem$ (в скобках указаны стандартные ошибки).

- Напишите формулу для функции $\Phi(x)$.
- Проверьте значимость влияния каждого из факторов (уровень значимости 10%).
- Дайте интерпретацию коэффициентов модели.
- Как интерпретируются предсказанные значения в этой модели регрессии?
- Вычислите предельное значение $\mathsf{P}(Male = 1)$ по $educ$.

# Приложение А

# Статистические таблицы

Приведем список функции MS Excel и OpenOffice для вычисления критических значений стандартных распределений с уровнем значимости $\alpha$

| Распределение | MS Excel (Рус) | MS Excel (Eng)<br>OpenOffice |
|---|---|---|
| Двустороннее $\mathcal{N}(0,1)$ | НОРМСТОБР$(1-\alpha/2)$ | NORMSINV$(1-\alpha/2)$ |
| Одностороннее $\mathcal{N}(0,1)$ | НОРМСТОБР$(1-\alpha)$ | NORMSINV$(1-\alpha)$ |
| $\chi^2_k$ (хи-квадрат) | ХИ2ОБР$(\alpha;k)$ | CHIINV$(\alpha;k)$ |
| Двустороннее $t_k$ (Стьюдента) | СТЬЮДРАСПОБР$(\alpha;k)$ | TINV$(\alpha;k)$ |
| Одностороннее $t_k$ (Стьюдента) | СТЬЮДРАСПОБР$(2\alpha;k)$ | TINV$(2\alpha;k)$ |
| Фишера $F_{k_1,k_2}$ | FРАСПОБР$(\alpha;k_1;k_2)$ | FINV $(\alpha;k_1;k_2)$ |

Таблица А.1: Критические значения стандартного нормального распределения

| | Уровень значимости $\alpha$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| two-side | 0.400 | 0.200 | 0.100 | 0.050 | 0.020 | 0.010 | 0.005 | 0.002 | 0.001 |
| one-side | 0.200 | 0.100 | 0.050 | 0.025 | 0.010 | 0.005 | 0.0025 | 0.001 | 0.0005 |
| $z_{\text{кр}}$ | 0.842 | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 | 2.807 | 3.090 | 3.291 |

Таблица А.2: Критические значения распределения $\chi^2_k$

| | Уровень значимости $\alpha$ | | | | | | |
|---|---|---|---|---|---|---|---|
| $k$ | 0.100 | 0.050 | 0.025 | 0.020 | 0.010 | 0.005 | 0.001 |
| 1 | 2.706 | 3.841 | 5.024 | 5.412 | 6.635 | 7.879 | 10,828 |
| 2 | 4.605 | 5.991 | 7.378 | 7.824 | 9.210 | 10.597 | 13.816 |
| 3 | 6.251 | 7.815 | 9.348 | 9.837 | 11.345 | 12.838 | 16.266 |
| 4 | 7.779 | 9.488 | 11.143 | 11.668 | 13.277 | 14.860 | 18.467 |
| 5 | 9.236 | 11.070 | 12.833 | 13.388 | 15.086 | 16.750 | 20.515 |
| 6 | 10.645 | 12.592 | 14.449 | 15.033 | 16.812 | 18.548 | 22.458 |
| 7 | 12.017 | 14.067 | 16.013 | 16.622 | 18.475 | 20.278 | 24.322 |
| 8 | 13.362 | 15.507 | 17.535 | 18.168 | 20.090 | 21.955 | 26.124 |
| 9 | 14.684 | 16.919 | 19.023 | 19.679 | 21.666 | 23.589 | 27.877 |
| 10 | 15.987 | 18.307 | 20.483 | 21.161 | 23.209 | 25.188 | 29.588 |
| 11 | 17.275 | 19.675 | 21.920 | 22.618 | 24.725 | 26.757 | 31.264 |
| 12 | 18.549 | 21.026 | 23.337 | 24.054 | 26.217 | 28.300 | 32.909 |
| 13 | 19.812 | 22.362 | 24.736 | 25.472 | 27.688 | 29.819 | 34.528 |
| 14 | 21.064 | 23.685 | 26.119 | 26.873 | 29.141 | 31.319 | 36.123 |
| 15 | 22.307 | 24.996 | 27.488 | 28.259 | 30.578 | 32.801 | 37.697 |
| 16 | 23.542 | 26.296 | 28.845 | 29.633 | 32.000 | 34.267 | 39.252 |
| 17 | 24.769 | 27.587 | 30.191 | 30.995 | 33.409 | 35.718 | 40.790 |
| 18 | 25.989 | 28.869 | 31.526 | 32.346 | 34.805 | 37.156 | 42.312 |
| 19 | 27.204 | 30.144 | 32.852 | 33.687 | 36.191 | 38.582 | 43.820 |
| 21 | 29.615 | 32.671 | 35.479 | 36.343 | 38.932 | 41.401 | 46.797 |
| 22 | 30.813 | 33.924 | 36.781 | 37.659 | 40.289 | 42.796 | 48.268 |
| 23 | 32.007 | 35.172 | 38.076 | 38.968 | 41.638 | 44.181 | 49.728 |
| 24 | 33.196 | 36.415 | 39.364 | 40.270 | 42.980 | 45.559 | 51.179 |
| 25 | 34.382 | 37.652 | 40.646 | 41.566 | 44.314 | 46.928 | 52.620 |
| 30 | 40.256 | 43.773 | 46.979 | 47.962 | 50.892 | 53.672 | 59.703 |
| 40 | 51.805 | 55.758 | 59.342 | 60.436 | 63.691 | 66.766 | 73.402 |
| 50 | 63.167 | 67.505 | 71.420 | 72.613 | 76.154 | 79.490 | 86.661 |
| 60 | 74.397 | 79.082 | 83.298 | 84.580 | 88.379 | 91.952 | 99.607 |
| 70 | 85.527 | 90.531 | 95.023 | 96.388 | 100.425 | 104.215 | 112.317 |
| 80 | 96.578 | 101.879 | 106.629 | 108.069 | 112.329 | 116.321 | 124.839 |
| 90 | 107.565 | 113.145 | 118.136 | 119.648 | 124.116 | 128.299 | 137.208 |
| 100 | 118.498 | 124.342 | 129.561 | 131.142 | 135.807 | 140.169 | 149.449 |

Таблица А.3: Критические значения распределения $t_k$ (распределения Стьюдента)

| $k$ | Уровень значимости $\alpha$ | | | | | |
|---|---|---|---|---|---|---|
| two-side | 0.100 | 0.050 | 0.025 | 0.010 | 0.005 | 0.001 |
| one-side | 0.050 | 0.0250 | 0.0125 | 0.0050 | 0.0025 | 0.0005 |
| 1 | 6.314 | 12.706 | 25.452 | 63.657 | 127.321 | 636.619 |
| 2 | 2.920 | 4.303 | 6.205 | 9.925 | 14.089 | 31.599 |
| 3 | 2.353 | 3.182 | 4.177 | 5.841 | 7.453 | 12.924 |
| 4 | 2.132 | 2.776 | 3.495 | 4.604 | 5.598 | 8.610 |
| 5 | 2.015 | 2.571 | 3.163 | 4.032 | 4.773 | 6.869 |
| 6 | 1.943 | 2.447 | 2.969 | 3.707 | 4.317 | 5.959 |
| 7 | 1.895 | 2.365 | 2.841 | 3.499 | 4.029 | 5.408 |
| 8 | 1.860 | 2.306 | 2.752 | 3.355 | 3.833 | 5.041 |
| 9 | 1.833 | 2.262 | 2.685 | 3.250 | 3.690 | 4.781 |
| 10 | 1.812 | 2.228 | 2.634 | 3.169 | 3.581 | 4.587 |
| 11 | 1.796 | 2.201 | 2.593 | 3.106 | 3.497 | 4.437 |
| 12 | 1.782 | 2.179 | 2.560 | 3.055 | 3.428 | 4.318 |
| 13 | 1.771 | 2.160 | 2.533 | 3.012 | 3.372 | 4.221 |
| 14 | 1.761 | 2.145 | 2.510 | 2.977 | 3.326 | 4.140 |
| 15 | 1.753 | 2.131 | 2.490 | 2.947 | 3.286 | 4.073 |
| 16 | 1.746 | 2.120 | 2.473 | 2.921 | 3.252 | 4.015 |
| 17 | 1.740 | 2.110 | 2.458 | 2.898 | 3.222 | 3.965 |
| 18 | 1.734 | 2.101 | 2.445 | 2.878 | 3.197 | 3.922 |
| 19 | 1.729 | 2.093 | 2.433 | 2.861 | 3.174 | 3.883 |
| 20 | 1.725 | 2.086 | 2.423 | 2.845 | 3.153 | 3.850 |
| 21 | 1.721 | 2.080 | 2.414 | 2.831 | 3.135 | 3.819 |
| 22 | 1.717 | 2.074 | 2.405 | 2.819 | 3.119 | 3.792 |
| 23 | 1.714 | 2.069 | 2.398 | 2.807 | 3.104 | 3.768 |
| 24 | 1.711 | 2.064 | 2.391 | 2.797 | 3.091 | 3.745 |
| 25 | 1.708 | 2.060 | 2.385 | 2.787 | 3.078 | 3.725 |
| 26 | 1.706 | 2.056 | 2.379 | 2.779 | 3.067 | 3.707 |
| 27 | 1.703 | 2.052 | 2.373 | 2.771 | 3.057 | 3.690 |
| 28 | 1.701 | 2.048 | 2.368 | 2.763 | 3.047 | 3.674 |
| 29 | 1.699 | 2.045 | 2.364 | 2.756 | 3.038 | 3.659 |
| 30 | 1.697 | 2.042 | 2.360 | 2.750 | 3.030 | 3.646 |
| 40 | 1.684 | 2.021 | 2.329 | 2.704 | 2.971 | 3.551 |
| 50 | 1.676 | 2.009 | 2.311 | 2.678 | 2.937 | 3.496 |
| 60 | 1.671 | 2.000 | 2.299 | 2.660 | 2.915 | 3.460 |
| 70 | 1.667 | 1.994 | 2.291 | 2.648 | 2.899 | 3.435 |
| 80 | 1.664 | 1.990 | 2.284 | 2.639 | 2.887 | 3.416 |
| 90 | 1.662 | 1.987 | 2.280 | 2.632 | 2.878 | 3.402 |
| 100 | 1.660 | 1.984 | 2.276 | 2.626 | 2.871 | 3.390 |
| 120 | 1.658 | 1.980 | 2.270 | 2.617 | 2.860 | 3.373 |
| 200 | 1.653 | 1.972 | 2.258 | 2.601 | 2.839 | 3.340 |
| 300 | 1.650 | 1.968 | 2.253 | 2.592 | 2.828 | 3.323 |
| 500 | 1.648 | 1.965 | 2.248 | 2.586 | 2.820 | 3.310 |
| $\infty$ | 1.645 | 1.960 | 2.241 | 2.576 | 2.807 | 3.291 |

Таблица А.4: 5% критические значения распределения $F_{k_1,k_2}$ (распределения Фишера)

| | $k_1$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $k_2$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 2 | 18.513 | 19.000 | 19.164 | 19.247 | 19.296 | 19.330 | 19.353 | 19.371 | 19.385 | 19.396 |
| 3 | 10.128 | 9.552 | 9.277 | 9.117 | 9.013 | 8.941 | 8.887 | 8.845 | 8.812 | 8.786 |
| 4 | 7.709 | 6.944 | 6.591 | 6.388 | 6.256 | 6.163 | 6.094 | 6.041 | 5.999 | 5.964 |
| 5 | 6.608 | 5.786 | 5.409 | 5.192 | 5.050 | 4.950 | 4.876 | 4.818 | 4.772 | 4.735 |
| 6 | 5.987 | 5.143 | 4.757 | 4.534 | 4.387 | 4.284 | 4.207 | 4.147 | 4.099 | 4.060 |
| 7 | 5.591 | 4.737 | 4.347 | 4.120 | 3.972 | 3.866 | 3.787 | 3.726 | 3.677 | 3.637 |
| 8 | 5.318 | 4.459 | 4.066 | 3.838 | 3.687 | 3.581 | 3.500 | 3.438 | 3.388 | 3.347 |
| 9 | 5.117 | 4.256 | 3.863 | 3.633 | 3.482 | 3.374 | 3.293 | 3.230 | 3.179 | 3.137 |
| 10 | 4.965 | 4.103 | 3.708 | 3.478 | 3.326 | 3.217 | 3.135 | 3.072 | 3.020 | 2.978 |
| 11 | 4.844 | 3.982 | 3.587 | 3.357 | 3.204 | 3.095 | 3.012 | 2.948 | 2.896 | 2.854 |
| 12 | 4.747 | 3.885 | 3.490 | 3.259 | 3.106 | 2.996 | 2.913 | 2.849 | 2.796 | 2.753 |
| 13 | 4.667 | 3.806 | 3.411 | 3.179 | 3.025 | 2.915 | 2.832 | 2.767 | 2.714 | 2.671 |
| 14 | 4.600 | 3.739 | 3.344 | 3.112 | 2.958 | 2.848 | 2.764 | 2.699 | 2.646 | 2,602 |
| 15 | 4.543 | 3.682 | 3.287 | 3.056 | 2.901 | 2.790 | 2.707 | 2.641 | 2.588 | 2.544 |
| 16 | 4.494 | 3.634 | 3.239 | 3.007 | 2.852 | 2.741 | 2.657 | 2.591 | 2.538 | 2.494 |
| 17 | 4.451 | 3.592 | 3.197 | 2.965 | 2.810 | 2.699 | 2.614 | 2.548 | 2.494 | 2.450 |
| 18 | 4.414 | 3.555 | 3.160 | 2.928 | 2.773 | 2.661 | 2.577 | 2.510 | 2.456 | 2.412 |
| 19 | 4.381 | 3.522 | 3.127 | 2.895 | 2.740 | 2.628 | 2.544 | 2.477 | 2.423 | 2.378 |
| 20 | 4.351 | 3.493 | 3.098 | 2.866 | 2.711 | 2.599 | 2.514 | 2.447 | 2.393 | 2.348 |
| 21 | 4.325 | 3.467 | 3.072 | 2.840 | 2.685 | 2.573 | 2.488 | 2.420 | 2.366 | 2.321 |
| 22 | 4.301 | 3.443 | 3.049 | 2.817 | 2.661 | 2.549 | 2.464 | 2.397 | 2.342 | 2.297 |
| 23 | 4.279 | 3.422 | 3.028 | 2.796 | 2.640 | 2.528 | 2.442 | 2.375 | 2.320 | 2.275 |
| 24 | 4.260 | 3.403 | 3.009 | 2.776 | 2.621 | 2.508 | 2.423 | 2.355 | 2.300 | 2.255 |
| 25 | 4.242 | 3.385 | 2.991 | 2.759 | 2.603 | 2.490 | 2.405 | 2.337 | 2.282 | 2.236 |
| 26 | 4.225 | 3.369 | 2.975 | 2.743 | 2.587 | 2.474 | 2.388 | 2.321 | 2.265 | 2.220 |
| 27 | 4.210 | 3.354 | 2,960 | 2.728 | 2.572 | 2.459 | 2.373 | 2.305 | 2.250 | 2.204 |
| 28 | 4.196 | 3.340 | 2.947 | 2.714 | 2.558 | 2.445 | 2.359 | 2.291 | 2.236 | 2.190 |
| 29 | 4.183 | 3.328 | 2.934 | 2.701 | 2.545 | 2.432 | 2.346 | 2.278 | 2.223 | 2.177 |
| 30 | 4.171 | 3.316 | 2.922 | 2.690 | 2.534 | 2.421 | 2.334 | 2.266 | 2.211 | 2.165 |
| 40 | 4.085 | 3.232 | 2.839 | 2.606 | 2.449 | 2.336 | 2.249 | 2.180 | 2.124 | 2.077 |
| 50 | 4.034 | 3.183 | 2.790 | 2.557 | 2.400 | 2.286 | 2.199 | 2.130 | 2.073 | 2.026 |
| 60 | 4.001 | 3.150 | 2.758 | 2.525 | 2.368 | 2.254 | 2.167 | 2.097 | 2.040 | 1.993 |
| 100 | 3.936 | 3.087 | 2.696 | 2.463 | 2.305 | 2.191 | 2.103 | 2.032 | 1.975 | 1.927 |
| 120 | 3.920 | 3.072 | 2.680 | 2.447 | 2.290 | 2.175 | 2.087 | 2.016 | 1.959 | 1.910 |
| 500 | 3.860 | 3.014 | 2.623 | 2.390 | 2.232 | 2.117 | 2.028 | 1.957 | 1.899 | 1.850 |
| $\infty$ | 3.841 | 2.996 | 2.605 | 2.372 | 2.214 | 2.099 | 2.010 | 1.938 | 1.880 | 1.831 |

Таблица А.5: 5% критические значения распределения $F_{k_1,k_2}$ (распределения Фишера)

| | $k_1$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $k_2$ | 15 | 20 | 25 | 30 | 40 | 50 | 60 | 100 | 120 |
| 2 | 19.429 | 19.446 | 19.456 | 19.462 | 19.471 | 19.476 | 19.479 | 19.486 | 19.487 |
| 3 | 8.703 | 8.660 | 8.634 | 8.617 | 8.594 | 8.581 | 8.572 | 8.554 | 8.549 |
| 4 | 5.858 | 5.803 | 5.769 | 5.746 | 5.717 | 5.699 | 5.688 | 5.664 | 5.658 |
| 5 | 4.619 | 4.558 | 4.521 | 4.496 | 4.464 | 4.444 | 4.431 | 4.405 | 4.398 |
| 6 | 3.938 | 3.874 | 3.835 | 3.808 | 3.774 | 3.754 | 3.740 | 3.712 | 3.705 |
| 7 | 3.511 | 3.445 | 3.404 | 3.376 | 3.340 | 3.319 | 3.304 | 3.275 | 3.267 |
| 8 | 3.218 | 3.150 | 3.108 | 3.079 | 3.043 | 3.020 | 3.005 | 2.975 | 2.967 |
| 9 | 3.006 | 2.936 | 2.893 | 2.864 | 2.826 | 2.803 | 2.787 | 2.756 | 2.748 |
| 10 | 2.845 | 2.774 | 2.730 | 2.700 | 2.661 | 2.637 | 2.621 | 2.588 | 2.580 |
| 11 | 2.719 | 2.646 | 2.601 | 2.570 | 2.531 | 2.507 | 2.490 | 2.457 | 2.448 |
| 12 | 2.617 | 2.544 | 2.498 | 2.466 | 2.426 | 2.401 | 2.384 | 2.350 | 2.341 |
| 13 | 2.533 | 2.459 | 2.412 | 2.380 | 2.339 | 2.314 | 2.297 | 2.261 | 2.252 |
| 14 | 2.463 | 2.388 | 2.341 | 2.308 | 2.266 | 2.241 | 2.223 | 2.187 | 2.178 |
| 15 | 2.403 | 2.328 | 2.280 | 2.247 | 2.204 | 2.178 | 2.160 | 2.123 | 2.114 |
| 16 | 2.352 | 2.276 | 2.227 | 2.194 | 2.151 | 2.124 | 2.106 | 2.068 | 2.059 |
| 17 | 2.308 | 2.230 | 2.181 | 2.148 | 2.104 | 2.077 | 2.058 | 2.020 | 2.011 |
| 18 | 2.269 | 2.191 | 2.141 | 2.107 | 2.063 | 2.035 | 2.017 | 1.978 | 1.968 |
| 19 | 2.234 | 2.155 | 2.106 | 2.071 | 2.026 | 1.999 | 1.980 | 1.940 | 1.930 |
| 20 | 2.203 | 2.124 | 2.074 | 2.039 | 1.994 | 1.966 | 1.946 | 1.907 | 1.896 |
| 21 | 2.176 | 2.096 | 2.045 | 2.010 | 1.965 | 1.936 | 1.916 | 1.876 | 1.866 |
| 22 | 2.151 | 2.071 | 2.020 | 1.984 | 1.938 | 1.909 | 1.889 | 1.849 | 1.838 |
| 23 | 2.128 | 2.048 | 1.996 | 1.961 | 1.914 | 1.885 | 1.865 | 1.823 | 1.813 |
| 24 | 2.108 | 2.027 | 1.975 | 1.939 | 1.892 | 1.863 | 1.842 | 1.800 | 1.790 |
| 25 | 2.089 | 2.007 | 1.955 | 1.919 | 1.872 | 1.842 | 1.822 | 1.779 | 1.768 |
| 26 | 2.072 | 1.990 | 1.938 | 1.901 | 1.853 | 1.823 | 1.803 | 1.760 | 1.749 |
| 27 | 2.056 | 1.974 | 1.921 | 1.884 | 1.836 | 1.806 | 1.785 | 1.742 | 1.731 |
| 28 | 2.041 | 1.959 | 1.906 | 1.869 | 1.820 | 1.790 | 1.769 | 1.725 | 1.714 |
| 29 | 2.027 | 1.945 | 1.891 | 1.854 | 1.806 | 1.775 | 1.754 | 1.710 | 1.698 |
| 30 | 2.015 | 1.932 | 1.878 | 1.841 | 1.792 | 1.761 | 1.740 | 1.695 | 1.683 |
| 40 | 1.924 | 1.839 | 1.783 | 1.744 | 1.693 | 1.660 | 1.637 | 1.589 | 1.577 |
| 50 | 1.871 | 1.784 | 1.727 | 1.687 | 1.634 | 1.599 | 1.576 | 1.525 | 1.511 |
| 60 | 1.836 | 1.748 | 1.690 | 1.649 | 1.594 | 1.559 | 1.534 | 1.481 | 1.467 |
| 100 | 1.768 | 1.676 | 1.616 | 1.573 | 1.515 | 1.477 | 1.450 | 1.392 | 1.376 |
| 120 | 1.750 | 1.659 | 1.598 | 1.554 | 1.495 | 1.457 | 1.429 | 1.369 | 1.352 |
| 500 | 1.686 | 1.592 | 1.528 | 1.482 | 1.419 | 1.376 | 1.345 | 1.275 | 1.255 |
| $\infty$ | 1.666 | 1.571 | 1.506 | 1.459 | 1.394 | 1.350 | 1.318 | 1.243 | 1.221 |

Таблица А.6: 10% критические значения распределения $F_{k_1,k_2}$ (распределения Фишера)

| | $k_1$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $k_2$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 2 | 8.526 | 9.000 | 9.162 | 9.243 | 9.293 | 9.326 | 9.349 | 9.367 | 9.381 | 9.392 |
| 3 | 5.538 | 5.462 | 5.391 | 5.343 | 5.309 | 5.285 | 5.266 | 5.252 | 5.240 | 5.230 |
| 4 | 4.545 | 4.325 | 4.191 | 4.107 | 4.051 | 4.010 | 3.979 | 3.955 | 3.936 | 3.920 |
| 5 | 4.060 | 3.780 | 3.619 | 3.520 | 3.453 | 3.405 | 3.368 | 3.339 | 3.316 | 3.297 |
| 6 | 3.776 | 3.463 | 3.289 | 3.181 | 3.108 | 3.055 | 3.014 | 2.983 | 2.958 | 2.937 |
| 7 | 3.589 | 3.257 | 3.074 | 2.961 | 2.883 | 2.827 | 2.785 | 2.752 | 2.725 | 2.703 |
| 8 | 3.458 | 3.113 | 2.924 | 2.806 | 2.726 | 2.668 | 2.624 | 2.589 | 2.561 | 2.538 |
| 9 | 3.360 | 3.006 | 2.813 | 2.693 | 2.611 | 2.551 | 2.505 | 2.469 | 2.440 | 2.416 |
| 10 | 3.285 | 2.924 | 2.728 | 2.605 | 2.522 | 2.461 | 2.414 | 2.377 | 2.347 | 2.323 |
| 11 | 3.225 | 2.860 | 2.660 | 2.536 | 2.451 | 2.389 | 2.342 | 2.304 | 2.274 | 2.248 |
| 12 | 3.177 | 2.807 | 2.606 | 2.480 | 2.394 | 2.331 | 2.283 | 2.245 | 2.214 | 2.188 |
| 13 | 3.136 | 2.763 | 2.560 | 2.434 | 2.347 | 2.283 | 2.234 | 2.195 | 2.164 | 2.138 |
| 14 | 3.102 | 2.726 | 2.522 | 2.395 | 2.307 | 2.243 | 2.193 | 2.154 | 2.122 | 2.095 |
| 15 | 3.073 | 2.695 | 2.490 | 2.361 | 2.273 | 2.208 | 2.158 | 2.119 | 2.086 | 2.059 |
| 16 | 3.048 | 2.668 | 2.462 | 2.333 | 2.244 | 2.178 | 2.128 | 2.088 | 2.055 | 2.028 |
| 17 | 3.026 | 2.645 | 2.437 | 2.308 | 2.218 | 2.152 | 2.102 | 2.061 | 2.028 | 2.001 |
| 18 | 3.007 | 2.624 | 2.416 | 2.286 | 2.196 | 2.130 | 2.079 | 2.038 | 2.005 | 1.977 |
| 19 | 2.990 | 2.606 | 2.397 | 2.266 | 2.176 | 2.109 | 2.058 | 2.017 | 1.984 | 17956 |
| 20 | 2.975 | 2.589 | 2.380 | 2.249 | 2.158 | 2.091 | 2.040 | 1.999 | 1.965 | 1.937 |
| 21 | 2.961 | 2.575 | 2.365 | 2.233 | 2.142 | 2.075 | 2.023 | 1.982 | 1.948 | 1.920 |
| 22 | 2.949 | 2.561 | 2.351 | 2.219 | 2.128 | 2.060 | 2.008 | 1.967 | 1.933 | 1.904 |
| 23 | 2.937 | 2.549 | 2.339 | 2.207 | 2.115 | 2.047 | 1.995 | 1.953 | 1.919 | 1.890 |
| 24 | 2.927 | 2.538 | 2.327 | 2.195 | 2.103 | 2.035 | 1.983 | 1.941 | 1.906 | 1.877 |
| 25 | 2.918 | 2.528 | 2.317 | 2.184 | 2.092 | 2.024 | 1.971 | 1.929 | 1.895 | 1.866 |
| 26 | 2.909 | 2.519 | 2.307 | 2.174 | 2.082 | 2.014 | 1.961 | 1.919 | 1.884 | 1.855 |
| 27 | 2.901 | 2.511 | 2.299 | 2.165 | 2.073 | 2.005 | 1.952 | 1.909 | 1.874 | 1.845 |
| 28 | 2.894 | 2.503 | 2.291 | 2.157 | 2.064 | 1.996 | 1.943 | 1.900 | 1.865 | 1.836 |
| 29 | 2.887 | 2.495 | 2.283 | 2.149 | 2.057 | 1.988 | 1.935 | 1.892 | 1.857 | 1.827 |
| 30 | 2.881 | 2.489 | 2.276 | 2.142 | 2.049 | 1.980 | 1.927 | 1.884 | 1.849 | 1.819 |
| 40 | 2.835 | 2.440 | 2.226 | 2.091 | 1.997 | 1.927 | 1.873 | 1.829 | 1.793 | 1.763 |
| 50 | 2.809 | 2.412 | 2.197 | 2.061 | 1.966 | 1.895 | 1.840 | 1.796 | 1.760 | 1.729 |
| 60 | 2.791 | 2.393 | 2.177 | 2.041 | 1.946 | 1.875 | 1.819 | 1.775 | 1.738 | 1.707 |
| 100 | 2.756 | 2.356 | 2.139 | 2.002 | 1.906 | 1.834 | 1.778 | 1.732 | 1.695 | 1.663 |
| 120 | 2.748 | 2.347 | 2.130 | 1.992 | 1.896 | 1.824 | 1.767 | 1.722 | 1.684 | 1.652 |
| 500 | 2.716 | 2.313 | 2.095 | 1.956 | 1.859 | 1.786 | 1.729 | 1.683 | 1.644 | 1.612 |
| $\infty$ | 2.706 | 2.303 | 2.084 | 1.945 | 1.847 | 1.774 | 1.717 | 1.670 | 1.632 | 1.599 |

Таблица А.7: 10% критические значения распределения $F_{k_1,k_2}$ (распределения Фишера)

| | $k_1$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $k_2$ | 15 | 20 | 25 | 30 | 40 | 50 | 60 | 100 | 120 |
| 2 | 9.425 | 9.441 | 9.451 | 9.458 | 9.466 | 9.471 | 9.475 | 9.481 | 9.483 |
| 3 | 5.200 | 5.184 | 5.175 | 5.168 | 5.160 | 5.155 | 5.151 | 5.144 | 5.143 |
| 4 | 3.870 | 3.844 | 3.828 | 3.817 | 3.804 | 3.795 | 3.790 | 3.778 | 3.775 |
| 5 | 3.238 | 3.207 | 3.187 | 3.174 | 3.157 | 3.147 | 3.140 | 3.126 | 3.123 |
| 6 | 2.871 | 2.836 | 2.815 | 2.800 | 2.781 | 2.770 | 2.762 | 2.746 | 2,742 |
| 7 | 2.632 | 2.595 | 2.571 | 2.555 | 2.535 | 2.523 | 2.514 | 2.497 | 2.493 |
| 8 | 2.464 | 2.425 | 2.400 | 2.383 | 2.361 | 2.348 | 2.339 | 2.321 | 2.316 |
| 9 | 2.340 | 2.298 | 2.272 | 2.255 | 2.232 | 2.218 | 2.208 | 2.189 | 2.184 |
| 10 | 2.244 | 2.201 | 2.174 | 2.155 | 2.132 | 2.117 | 2.107 | 2.087 | 2.082 |
| 11 | 2.167 | 2.123 | 2.095 | 2.076 | 2.052 | 2.036 | 2.026 | 2.005 | 2.000 |
| 12 | 2.105 | 2.060 | 2.031 | 2.011 | 1.986 | 1.970 | 1.960 | 1.938 | 1.932 |
| 13 | 2.053 | 2.007 | 1.978 | 1.958 | 1.931 | 1.915 | 1.904 | 1.882 | 1.876 |
| 14 | 2.010 | 1.962 | 1.933 | 1.912 | 1.885 | 1.869 | 1.857 | 1.834 | 1.828 |
| 15 | 1.972 | 1.924 | 1.894 | 1.873 | 1.845 | 1.828 | 1.817 | 1.793 | 1.787 |
| 16 | 1.940 | 1.891 | 1.860 | 1.839 | 1.811 | 1.793 | 1.782 | 1.757 | 1.751 |
| 17 | 1.912 | 1.862 | 1.831 | 1.809 | 1.781 | 1.763 | 1.751 | 1.726 | 1.719 |
| 18 | 1.887 | 1.837 | 1.805 | 1.783 | 1.754 | 1.736 | 1.723 | 1.698 | 1.691 |
| 19 | 1.865 | 1.814 | 1.782 | 1.759 | 1.730 | 1.711 | 1.699 | 1.673 | 1.666 |
| 20 | 1.845 | 1.794 | 1.761 | 1.738 | 1.708 | 1.690 | 1.677 | 1.650 | 1.643 |
| 21 | 1.827 | 1.776 | 1.742 | 1.719 | 1.689 | 1.670 | 1.657 | 1.630 | 1.623 |
| 22 | 1.811 | 1.759 | 1.726 | 1.702 | 1.671 | 1.652 | 1.639 | 1.611 | 1.604 |
| 23 | 1.796 | 1.744 | 1.710 | 1.686 | 1.655 | 1.636 | 1.622 | 1.594 | 1.587 |
| 24 | 1.783 | 1.730 | 1.696 | 1.672 | 1.641 | 1.621 | 1.607 | 1.579 | 1.571 |
| 25 | 1.771 | 1.718 | 1.683 | 1.659 | 1.627 | 1.607 | 1.593 | 1.565 | 1.557 |
| 26 | 1.760 | 1.706 | 1.671 | 1.647 | 1.615 | 1.594 | 1.581 | 1.551 | 1.544 |
| 27 | 1.749 | 1.695 | 1.660 | 1.636 | 1.603 | 1.583 | 1.569 | 1.539 | 1.531 |
| 28 | 1.740 | 1.685 | 1.650 | 1.625 | 1.592 | 1.572 | 1.558 | 1.528 | 1.520 |
| 29 | 1.731 | 1.676 | 1.640 | 1.616 | 1.583 | 1.562 | 1.547 | 1.517 | 1.509 |
| 30 | 1.722 | 1.667 | 1.632 | 1.606 | 1.573 | 1.552 | 1.538 | 1.507 | 1.499 |
| 40 | 1.662 | 1.605 | 1.568 | 1.541 | 1.506 | 1.483 | 1.467 | 1.434 | 1.425 |
| 50 | 1.627 | 1.568 | 1.529 | 1.502 | 1.465 | 1.441 | 1.424 | 1.388 | 1.379 |
| 60 | 1.603 | 1.543 | 1.504 | 1.476 | 1.437 | 1.413 | 1.395 | 1.358 | 1.348 |
| 100 | 1.557 | 1.494 | 1.453 | 1.423 | 1.382 | 1.355 | 1.336 | 1.293 | 1.282 |
| 120 | 1.545 | 1.482 | 1.440 | 1.409 | 1.368 | 1.340 | 1.320 | 1.277 | 1.265 |
| 500 | 1.501 | 1.435 | 1.391 | 1.358 | 1.313 | 1.282 | 1.260 | 1.209 | 1.194 |
| $\infty$ | 1.487 | 1.421 | 1.375 | 1.342 | 1.295 | 1.263 | 1.240 | 1.185 | 1.169 |

Таблица А.8: 1% критические значения распределения $F_{k_1,k_2}$ (распределения Фишера)

| | $k_1$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $k_2$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 2 | 98.503 | 99.000 | 99.166 | 99.249 | 99.299 | 99.333 | 99.356 | 99.374 | 99.388 | 99.399 |
| 3 | 34.116 | 30.817 | 29.457 | 28.710 | 28.237 | 27.911 | 27.672 | 27.489 | 27.345 | 27.229 |
| 4 | 21.198 | 18.000 | 16.694 | 15.977 | 15.522 | 15.207 | 14.976 | 14.799 | 14.659 | 14.546 |
| 5 | 16.258 | 13.274 | 12.060 | 11.392 | 10.967 | 10.672 | 10.456 | 10.289 | 10.158 | 10.051 |
| 6 | 13.745 | 10.925 | 9.780 | 9.148 | 8.746 | 8.466 | 8.260 | 8.102 | 7.976 | 7.874 |
| 7 | 12.246 | 9.547 | 8.451 | 7.847 | 7.460 | 7.191 | 6.993 | 6.840 | 6.719 | 6.620 |
| 8 | 11.259 | 8.649 | 7.591 | 7.006 | 6.632 | 6.371 | 6.178 | 6.029 | 5.911 | 5.814 |
| 9 | 10.561 | 8.022 | 6.992 | 6.422 | 6.057 | 5.802 | 5.613 | 5.467 | 5.351 | 5.257 |
| 10 | 10.044 | 7.559 | 6.552 | 5.994 | 5.636 | 5.386 | 5.200 | 5.057 | 4.942 | 4.849 |
| 11 | 9.646 | 7.206 | 6.217 | 5.668 | 5.316 | 5.069 | 4.886 | 4.744 | 4.632 | 4.539 |
| 12 | 9.330 | 6.927 | 5.953 | 5.412 | 5.064 | 4.821 | 4.640 | 4.499 | 4.388 | 4.296 |
| 13 | 9.074 | 6.701 | 5.739 | 5.205 | 4.862 | 4.620 | 4.441 | 4.302 | 4.191 | 4.100 |
| 14 | 8.862 | 6.515 | 5.564 | 5.035 | 4.695 | 4.456 | 4.278 | 4.140 | 4.030 | 3.939 |
| 15 | 8.683 | 6.359 | 5.417 | 4.893 | 4.556 | 4.318 | 4.142 | 4.004 | 3.895 | 3.805 |
| 16 | 8.531 | 6.226 | 5.292 | 4.773 | 4.437 | 4.202 | 4.026 | 3.890 | 3.780 | 3.691 |
| 17 | 8.400 | 6.112 | 5.185 | 4.669 | 4.336 | 4.102 | 3.927 | 3.791 | 3.682 | 3.593 |
| 18 | 8.285 | 6.013 | 5.092 | 4.579 | 4.248 | 4.015 | 3.841 | 3.705 | 3.597 | 3.508 |
| 19 | 8.185 | 5.926 | 5.010 | 4.500 | 4.171 | 3.939 | 3.765 | 3.631 | 3.523 | 3.434 |
| 20 | 8.096 | 5.849 | 4.938 | 4.431 | 4.103 | 3.871 | 3.699 | 3.564 | 3.457 | 3.368 |
| 21 | 8.017 | 5.780 | 4.874 | 4.369 | 4.042 | 3.812 | 3.640 | 3.506 | 3.398 | 3.310 |
| 22 | 7.945 | 5.719 | 4.817 | 4.313 | 3.988 | 3.758 | 3.587 | 3.453 | 3.346 | 3.258 |
| 23 | 7.881 | 5.664 | 4.765 | 4.264 | 3.939 | 3.710 | 3.539 | 3.406 | 3.299 | 3.211 |
| 24 | 7.823 | 5.614 | 4.718 | 4.218 | 3.895 | 3.667 | 3.496 | 3.363 | 3.256 | 3.168 |
| 25 | 7.770 | 5.568 | 4.675 | 4.177 | 3.855 | 3.627 | 3.457 | 3.324 | 3.217 | 3.129 |
| 26 | 7.721 | 5.526 | 4.637 | 4.140 | 3.818 | 3.591 | 3.421 | 3.288 | 3.182 | 3.094 |
| 27 | 7.677 | 5.488 | 4.601 | 4.106 | 3.785 | 3.558 | 3.388 | 3.256 | 3.149 | 3.062 |
| 28 | 7.636 | 5.453 | 4.568 | 4.074 | 3.754 | 3.528 | 3.358 | 3.226 | 3.120 | 3.032 |
| 29 | 7.598 | 5.420 | 4.538 | 4.045 | 3.725 | 3.499 | 3.330 | 3.198 | 3.092 | 3.005 |
| 30 | 7.562 | 5.390 | 4.510 | 4.018 | 3.699 | 3.473 | 3.304 | 3.173 | 3.067 | 2.979 |
| 40 | 7.314 | 5.179 | 4.313 | 3.828 | 3.514 | 3.291 | 3.124 | 2.993 | 2.888 | 2.801 |
| 50 | 7.171 | 5.057 | 4.199 | 3.720 | 3.408 | 3.186 | 3.020 | 2.890 | 2.785 | 2.698 |
| 60 | 7.077 | 4.977 | 4.126 | 3.649 | 3.339 | 3.119 | 2.953 | 2.823 | 2.718 | 2.632 |
| 100 | 6.895 | 4.824 | 3.984 | 3.513 | 3.206 | 2.988 | 2.823 | 2.694 | 2.590 | 2.503 |
| 120 | 6.851 | 4.787 | 3.949 | 3.480 | 3.174 | 2.956 | 2.792 | 2.663 | 2.559 | 2.472 |
| 500 | 6.686 | 4.648 | 3.821 | 3.357 | 3.054 | 2.838 | 2.675 | 2.547 | 2.443 | 2.356 |
| $\infty$ | 6.635 | 4.605 | 3.782 | 3.319 | 3.017 | 2.802 | 2.639 | 2.511 | 2.407 | 2.321 |

Таблица А.9: 1% критические значения распределения $F_{k_1,k_2}$ (распределения Фишера)

| | $k_1$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $k_2$ | 15 | 20 | 25 | 30 | 40 | 50 | 60 | 100 | 120 |
| 2 | 99.433 | 99.449 | 99.459 | 99.466 | 99.474 | 99.479 | 99.482 | 99.489 | 99.491 |
| 3 | 26.872 | 26.690 | 26.579 | 26,505 | 26.411 | 26.354 | 26.316 | 26.240 | 26.221 |
| 4 | 14.198 | 14.020 | 13.911 | 13.838 | 13.745 | 13.690 | 13.652 | 13.577 | 13.558 |
| 5 | 9.722 | 9.553 | 9.449 | 9.379 | 9.291 | 9.238 | 9.202 | 9.130 | 9.112 |
| 6 | 7.559 | 7.396 | 7.296 | 7.229 | 7.143 | 7.091 | 7.057 | 6.987 | 6.969 |
| 7 | 6.314 | 6.155 | 6.058 | 5.992 | 5.908 | 5.858 | 5.824 | 5.755 | 5.737 |
| 8 | 5.515 | 5.359 | 5.263 | 5.198 | 5.116 | 5.065 | 5.032 | 4.963 | 4.946 |
| 9 | 4.962 | 4.808 | 4.713 | 4.649 | 4.567 | 4.517 | 4.483 | 4.415 | 4.398 |
| 10 | 4558 | 4.405 | 4.311 | 4.247 | 4.165 | 4.115 | 4.082 | 4.014 | 3.996 |
| 11 | 4.251 | 4.099 | 4.005 | 3.941 | 3.860 | 3.810 | 3.776 | 3.708 | 3.690 |
| 12 | 4.010 | 3.858 | 3.765 | 3.701 | 3.619 | 3.569 | 3.535 | 3.467 | 3.449 |
| 13 | 3.815 | 3.665 | 3.571 | 3.507 | 3.425 | 3.375 | 3.341 | 3.272 | 3.255 |
| 14 | 3.656 | 3.505 | 3.412 | 3.348 | 3.266 | 3.215 | 3.181 | 3,112 | 3.094 |
| 15 | 3.522 | 3.372 | 3.278 | 3.214 | 3.132 | 3.081 | 3.047 | 2.977 | 2.959 |
| 16 | 3.409 | 3.259 | 3.165 | 3.101 | 3.018 | 2.967 | 2.933 | 2.863 | 2.845 |
| 17 | 3.312 | 3.162 | 3.068 | 3.003 | 2.920 | 2.869 | 2.835 | 2.764 | 2.746 |
| 18 | 3.227 | 3.077 | 2.983 | 2.919 | 2.835 | 2.784 | 2.749 | 2.678 | 2.660 |
| 19 | 3.153 | 3.003 | 2.909 | 2.844 | 2.761 | 2.709 | 2.674 | 2.602 | 2.584 |
| 20 | 3.088 | 2.938 | 2.843 | 2.778 | 2.695 | 2.643 | 2.608 | 2.535 | 2.517 |
| 21 | 3.030 | 2.880 | 2.785 | 2.720 | 2.636 | 2.584 | 2.548 | 2.475 | 2.457 |
| 22 | 2.978 | 2.827 | 2.733 | 2.667 | 2.583 | 2.531 | 2.495 | 2.422 | 2.403 |
| 23 | 2.931 | 2.781 | 2.686 | 2.620 | 2.535 | 2.483 | 2.447 | 2.373 | 2.354 |
| 24 | 2.889 | 2.738 | 2.643 | 2.577 | 2.492 | 2.440 | 2.403 | 2.329 | 2.310 |
| 25 | 2.850 | 2.699 | 2.604 | 2.538 | 2.453 | 2.400 | 2.364 | 2.289 | 2.270 |
| 26 | 2.815 | 2.664 | 2.569 | 2.503 | 2.417 | 2.364 | 2.327 | 2.252 | 2.233 |
| 27 | 2.783 | 2.632 | 2.536 | 2.470 | 2.384 | 2.330 | 2.294 | 2.218 | 2.198 |
| 28 | 2.753 | 2.602 | 2.506 | 2.440 | 2.354 | 2.300 | 2.263 | 2.187 | 2.167 |
| 29 | 2.726 | 2.574 | 2.478 | 2,412 | 2,325 | 2,271 | 2.234 | 2.158 | 2,138 |
| 30 | 2.700 | 2.549 | 2.453 | 2.386 | 2.299 | 2.245 | 2.208 | 2.131 | 2.111 |
| 40 | 2.522 | 2.369 | 2.271 | 2.203 | 2.114 | 2.058 | 2.019 | 1.938 | 1.917 |
| 50 | 2.419 | 2.265 | 2.167 | 2.098 | 2.007 | 1.949 | 1.909 | 1.825 | 1.803 |
| 60 | 2.352 | 2.198 | 2.098 | 2.028 | 1.936 | 1.877 | 1.836 | 1.749 | 1.726 |
| 100 | 2.223 | 2.067 | 1.965 | 1.893 | 1.797 | 1.735 | 1.692 | 1.598 | 1.572 |
| 120 | 2.192 | 2.035 | 1.932 | 1.860 | 1.763 | 1.700 | 1.656 | 1.559 | 1.533 |
| 500 | 2.075 | 1.915 | 1.810 | 1.735 | 1.633 | 1.566 | 1.517 | 1.408 | 1.377 |
| $\infty$ | 2.039 | 1.878 | 1.773 | 1.696 | 1.592 | 1.523 | 1.473 | 1.358 | 1.325 |

Таблица A.10: 1% критические значения теста Durbin–Watson

| | $k$=1 | | $k$=2 | | $k$=3 | | $k$=4 | | $k$=5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | dL | dU | dL | dU | dL | dU | dL | dU | dL | dU |
| 6 | 0.390 | 1.142 | — | — | — | — | — | — | — | — |
| 7 | 0.435 | 1.036 | 0.294 | 1.676 | — | — | — | — | — | — |
| 8 | 0.497 | 1.003 | 0.345 | 1.489 | 0.229 | 2.102 | — | — | — | — |
| 9 | 0.554 | 0.998 | 0.408 | 1.389 | 0.279 | 1.875 | 0.183 | 2.433 | — | — |
| 10 | 0.604 | 1.001 | 0.466 | 1.333 | 0.340 | 1.733 | 0.230 | 2.193 | 0.150 | 2.690 |
| 11 | 0.653 | 1.010 | 0.519 | 1.297 | 0.396 | 1.640 | 0.286 | 2.030 | 0.193 | 2.453 |
| 12 | 0.697 | 1.023 | 0.569 | 1.274 | 0.449 | 1.575 | 0.339 | 1.913 | 0.244 | 2.280 |
| 13 | 0.738 | 1.038 | 0.616 | 1.261 | 0.499 | 1.526 | 0.391 | 1.826 | 0.294 | 2.150 |
| 14 | 0.776 | 1.054 | 0.660 | 1.254 | 0.547 | 1.490 | 0.441 | 1.757 | 0.343 | 2.049 |
| 15 | 0.811 | 1.070 | 0.700 | 1.252 | 0.591 | 1.465 | 0.487 | 1.705 | 0.390 | 1.967 |
| 16 | 0.844 | 1.086 | 0.738 | 1.253 | 0.633 | 1.447 | 0.532 | 1.664 | 0.437 | 1.901 |
| 17 | 0.873 | 1.102 | 0.773 | 1.255 | 0.672 | 1.432 | 0.574 | 1.631 | 0.481 | 1.847 |
| 18 | 0.902 | 1.118 | 0.805 | 1.259 | 0.708 | 1.422 | 0.614 | 1.604 | 0.522 | 1.803 |
| 19 | 0.928 | 1.133 | 0.835 | 1.264 | 0.742 | 1.416 | 0.650 | 1.583 | 0.561 | 1.767 |
| 20 | 0.952 | 1.147 | 0.862 | 1.270 | 0.774 | 1.410 | 0.684 | 1.567 | 0.598 | 1.736 |
| 21 | 0.975 | 1.161 | 0.889 | 1.276 | 0.803 | 1.408 | 0.718 | 1.554 | 0.634 | 1.712 |
| 22 | 0.997 | 1.174 | 0.915 | 1.284 | 0.832 | 1.407 | 0.748 | 1.543 | 0.666 | 1.691 |
| 23 | 1.017 | 1.186 | 0.938 | 1.290 | 0.858 | 1.407 | 0.777 | 1.535 | 0.699 | 1.674 |
| 24 | 1.037 | 1.199 | 0.959 | 1.298 | 0.881 | 1.407 | 0.805 | 1.527 | 0.728 | 1.659 |
| 25 | 1.055 | 1.210 | 0.981 | 1.305 | 0.906 | 1.408 | 0.832 | 1.521 | 0.756 | 1.645 |
| 26 | 1.072 | 1.222 | 1.000 | 1.311 | 0.928 | 1.410 | 0.855 | 1.517 | 0.782 | 1.635 |
| 27 | 1.088 | 1.232 | 1.019 | 1.318 | 0.948 | 1.413 | 0.878 | 1.514 | 0.808 | 1.625 |
| 28 | 1.104 | 1.244 | 1.036 | 1.325 | 0.969 | 1.414 | 0.901 | 1.512 | 0.832 | 1.618 |
| 29 | 1.119 | 1.254 | 1.053 | 1.332 | 0.988 | 1.418 | 0.921 | 1.511 | 0.855 | 1.611 |
| 30 | 1.134 | 1.264 | 1.070 | 1.339 | 1.006 | 1.421 | 0.941 | 1.510 | 0.877 | 1.606 |
| 31 | 1.147 | 1.274 | 1.085 | 1.345 | 1.022 | 1.425 | 0.960 | 1.509 | 0.897 | 1.601 |
| 32 | 1.160 | 1.283 | 1.100 | 1.351 | 1.039 | 1.428 | 0.978 | 1.509 | 0.917 | 1.597 |
| 33 | 1.171 | 1.291 | 1.114 | 1.358 | 1.055 | 1.432 | 0.995 | 1.510 | 0.935 | 1.594 |
| 34 | 1.184 | 1.298 | 1.128 | 1.364 | 1.070 | 1.436 | 1.012 | 1.511 | 0.954 | 1.591 |
| 35 | 1.195 | 1.307 | 1.141 | 1.370 | 1.085 | 1.439 | 1.028 | 1.512 | 0.971 | 1.589 |
| 36 | 1.205 | 1.315 | 1.153 | 1.376 | 1.098 | 1.442 | 1.043 | 1.513 | 0.987 | 1.587 |
| 37 | 1.217 | 1.322 | 1.164 | 1.383 | 1.112 | 1.446 | 1.058 | 1.514 | 1.004 | 1.585 |
| 38 | 1.227 | 1.330 | 1.176 | 1.388 | 1.124 | 1.449 | 1.072 | 1.515 | 1.019 | 1.584 |
| 39 | 1.237 | 1.337 | 1.187 | 1.392 | 1.137 | 1.452 | 1.085 | 1.517 | 1.033 | 1.583 |
| 40 | 1.246 | 1.344 | 1.197 | 1.398 | 1.149 | 1.456 | 1.098 | 1.518 | 1.047 | 1.583 |
| 45 | 1.288 | 1.376 | 1.245 | 1.424 | 1.201 | 1.474 | 1.156 | 1.528 | 1.111 | 1.583 |
| 50 | 1.324 | 1.403 | 1.285 | 1.445 | 1.245 | 1.491 | 1.206 | 1.537 | 1.164 | 1.587 |
| 55 | 1.356 | 1.428 | 1.320 | 1.466 | 1.284 | 1.505 | 1.246 | 1.548 | 1.209 | 1.592 |
| 60 | 1.382 | 1.449 | 1.351 | 1.484 | 1.317 | 1.520 | 1.283 | 1.559 | 1.248 | 1.598 |
| 65 | 1.407 | 1.467 | 1.377 | 1.500 | 1.346 | 1.534 | 1.314 | 1.568 | 1.283 | 1.604 |
| 70 | 1.429 | 1.485 | 1.400 | 1.514 | 1.372 | 1.546 | 1.343 | 1.577 | 1.313 | 1.611 |
| 75 | 1.448 | 1.501 | 1.422 | 1.529 | 1.395 | 1.557 | 1.368 | 1.586 | 1.340 | 1.617 |
| 80 | 1.465 | 1.514 | 1.440 | 1.541 | 1.416 | 1.568 | 1.390 | 1.595 | 1.364 | 1.624 |
| 85 | 1.481 | 1.529 | 1.458 | 1.553 | 1.434 | 1.577 | 1.411 | 1.603 | 1.386 | 1.630 |
| 90 | 1.496 | 1.541 | 1.474 | 1.563 | 1.452 | 1.587 | 1.429 | 1.611 | 1.406 | 1.636 |
| 95 | 1.510 | 1.552 | 1.489 | 1.573 | 1.468 | 1.596 | 1.446 | 1.618 | 1.425 | 1.641 |
| 100 | 1.522 | 1.562 | 1.502 | 1.582 | 1.482 | 1.604 | 1.461 | 1.625 | 1.441 | 1.647 |
| 150 | 1.611 | 1.637 | 1.598 | 1.651 | 1.584 | 1.665 | 1.571 | 1.679 | 1.557 | 1.693 |
| 200 | 1.664 | 1.684 | 1.653 | 1.693 | 1.643 | 1.704 | 1.633 | 1.715 | 1.623 | 1.725 |

Таблица A.11: 1% критические значения теста Durbin–Watson

| | $k$=6 | | $k$=7 | | $k$=8 | | $k$=9 | | $k$=10 | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | dL | dU | dL | dU | dL | dU | dL | dU | dL | dU |
| 6 | — | — | — | — | — | — | — | — | — | — |
| 7 | — | — | — | — | — | — | — | — | — | — |
| 8 | — | — | — | — | — | — | — | — | — | — |
| 9 | — | — | — | — | — | — | — | — | — | — |
| 10 | — | — | — | — | — | — | — | — | — | — |
| 11 | 0.124 | 2.892 | — | — | — | — | — | — | — | — |
| 12 | 0.164 | 2.665 | 0.105 | 3.053 | — | — | — | — | — | — |
| 13 | 0.211 | 2.490 | 0.140 | 2.838 | 0.090 | 3.182 | — | — | — | — |
| 14 | 0.257 | 2.354 | 0.183 | 2.667 | 0.122 | 2.981 | 0.078 | 3.287 | — | — |
| 15 | 0.303 | 2.244 | 0.226 | 2.530 | 0.161 | 2.817 | 0.107 | 3.101 | 0.068 | 3.374 |
| 16 | 0.349 | 2.153 | 0.269 | 2.416 | 0.200 | 2.681 | 0.142 | 2.944 | 0.094 | 3.201 |
| 17 | 0.393 | 2.078 | 0.313 | 2.319 | 0.241 | 2.566 | 0.179 | 2.811 | 0.127 | 3.053 |
| 18 | 0.435 | 2.015 | 0.355 | 2.238 | 0.282 | 2.467 | 0.216 | 2.697 | 0.160 | 2.925 |
| 19 | 0.476 | 1.963 | 0.396 | 2.169 | 0.322 | 2.381 | 0.255 | 2.597 | 0.196 | 2.813 |
| 20 | 0.515 | 1.918 | 0.436 | 2.110 | 0.362 | 2.308 | 0.294 | 2.510 | 0.232 | 2.174 |
| 21 | 0.552 | 1.881 | 0.474 | 2.059 | 0.400 | 2.244 | 0.331 | 2.434 | 0.268 | 2.625 |
| 22 | 0.587 | 1.849 | 0.510 | 2.015 | 0.437 | 2.188 | 0.368 | 2.367 | 0.304 | 2.548 |
| 23 | 0.620 | 1.821 | 0.545 | 1.977 | 0.473 | 2.140 | 0.404 | 2.308 | 0.340 | 2.479 |
| 24 | 0.652 | 1.797 | 0.578 | 1.944 | 0.507 | 2.097 | 0.439 | 2.255 | 0.375 | 2.417 |
| 25 | 0.682 | 1.776 | 0.610 | 1.915 | 0.540 | 2.059 | 0.473 | 2.209 | 0.409 | 2.362 |
| 26 | 0.711 | 1.759 | 0.640 | 1.889 | 0.572 | 2.026 | 0.505 | 2.168 | 0.441 | 2.313 |
| 27 | 0.738 | 1.743 | 0.669 | 1.867 | 0.602 | 1.997 | 0.536 | 2.131 | 0.473 | 2.269 |
| 28 | 0.764 | 1.729 | 0.696 | 1.847 | 0.630 | 1.970 | 0.566 | 2.098 | 0.504 | 2.229 |
| 29 | 0.788 | 1.718 | 0.723 | 1.830 | 0.658 | 1.947 | 0.595 | 2.068 | 0.533 | 2.193 |
| 30 | 0.812 | 1.707 | 0.748 | 1.814 | 0.684 | 1.925 | 0.622 | 2.041 | 0.562 | 2.160 |
| 31 | 0.834 | 1.698 | 0.772 | 1.800 | 0.710 | 1.906 | 0.649 | 2.017 | 0.589 | 2.131 |
| 32 | 0.856 | 1.690 | 0.794 | 1.788 | 0.734 | 1.889 | 0.674 | 1.995 | 0.615 | 2.104 |
| 33 | 0.876 | 1.683 | 0.816 | 1.776 | 0.757 | 1.874 | 0.698 | 1.975 | 0.641 | 2.080 |
| 34 | 0.896 | 1.677 | 0.837 | 1.766 | 0.779 | 1.860 | 0.722 | 1.957 | 0.665 | 2.057 |
| 35 | 0.914 | 1.671 | 0.857 | 1.757 | 0.800 | 1.847 | 0.744 | 1.940 | 0.689 | 2.037 |
| 36 | 0.932 | 1.666 | 0.877 | 1.749 | 0.821 | 1.836 | 0.766 | 1.925 | 0.711 | 2.018 |
| 37 | 0.950 | 1.662 | 0.895 | 1.742 | 0.841 | 1.825 | 0.787 | 1.911 | 0.733 | 2.001 |
| 38 | 0.966 | 1.658 | 0.913 | 1.735 | 0.860 | 1.816 | 0.807 | 1.899 | 0.754 | 1.985 |
| 39 | 0.982 | 1.655 | 0.930 | 1.729 | 0.878 | 1.807 | 0.826 | 1.887 | 0.774 | 1.970 |
| 40 | 0.997 | 1.652 | 0.946 | 1.724 | 0.895 | 1.799 | 0.844 | 1.876 | 0.749 | 1.956 |
| 45 | 1.065 | 1.643 | 1.019 | 1.704 | 0.974 | 1.768 | 0.927 | 1.834 | 0.881 | 1.902 |
| 50 | 1.123 | 1.639 | 1.081 | 1.692 | 1.039 | 1.748 | 0.997 | 1.805 | 0.955 | 1.864 |
| 55 | 1.172 | 1.638 | 1.134 | 1.685 | 1.095 | 1.734 | 1.057 | 1.785 | 1.018 | 1.837 |
| 60 | 1.214 | 1.639 | 1.179 | 1.682 | 1.144 | 1.726 | 1.108 | 1.771 | 1.072 | 1.817 |
| 65 | 1.251 | 1.642 | 1.218 | 1.680 | 1.186 | 1.720 | 1.153 | 1.761 | 1.120 | 1.802 |
| 70 | 1.283 | 1.645 | 1.253 | 1.680 | 1.223 | 1.716 | 1.192 | 1.754 | 1.162 | 1.792 |
| 75 | 1.313 | 1.649 | 1.284 | 1.682 | 1.256 | 1.714 | 1.227 | 1.748 | 1.199 | 1.783 |
| 80 | 1.338 | 1.653 | 1.312 | 1.683 | 1.285 | 1.714 | 1.259 | 1.745 | 1.232 | 1.777 |
| 85 | 1.362 | 1.657 | 1.337 | 1.685 | 1.312 | 1.714 | 1.287 | 1.743 | 1.262 | 1.773 |
| 90 | 1.383 | 1.661 | 1.360 | 1.687 | 1.336 | 1.714 | 1.312 | 1.741 | 1.288 | 1.769 |
| 95 | 1.403 | 1.666 | 1.381 | 1.690 | 1.358 | 1.715 | 1.336 | 1.741 | 1.313 | 1.767 |
| 100 | 1.421 | 1.670 | 1.400 | 1.693 | 1.378 | 1.717 | 1.357 | 1.741 | 1.335 | 1.765 |
| 150 | 1.543 | 1.708 | 1.530 | 1.722 | 1.515 | 1.737 | 1.501 | 1.752 | 1.486 | 1.767 |
| 200 | 1.613 | 1.735 | 1.603 | 1.746 | 1.592 | 1.757 | 1.582 | 1.768 | 1.571 | 1.779 |

Таблица A.12: 5% критические значения теста Durbin–Watson

| | $k$=1 | | $k$=2 | | $k$=3 | | $k$=4 | | $k$=5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | dL | dU | dL | dU | dL | dU | dL | dU | dL | dU |
| 6 | 0.610 | 1.400 | — | — | — | — | — | — | — | — |
| 7 | 0.700 | 1.356 | 0.467 | 1.896 | — | — | — | — | — | — |
| 8 | 0.763 | 1.332 | 0.559 | 1.777 | 0.367 | 2.287 | — | — | — | — |
| 9 | 0.824 | 1.320 | 0.629 | 1.699 | 0.455 | 2.128 | 0.296 | 2.588 | — | — |
| 10 | 0.879 | 1.320 | 0.697 | 1.641 | 0.525 | 2.016 | 0.376 | 2.414 | 0.243 | 2.822 |
| 11 | 0.927 | 1.324 | 0.758 | 1.604 | 0.595 | 1.928 | 0.444 | 2.283 | 0.315 | 2.645 |
| 12 | 0.971 | 1.331 | 0.812 | 1.579 | 0.658 | 1.864 | 0.512 | 2.177 | 0.380 | 2.506 |
| 13 | 1.010 | 1.340 | 0.861 | 1.562 | 0.715 | 1.816 | 0.574 | 2.094 | 0.444 | 2.390 |
| 14 | 1.045 | 1.350 | 0.905 | 1.551 | 0.767 | 1.779 | 0.632 | 2.030 | 0.505 | 2.29 |
| 15 | 1.077 | 1.361 | 0.946 | 1.543 | 0.814 | 1.750 | 0.685 | 1.977 | 0.562 | 2.220 |
| 16 | 1.106 | 1.371 | 0.982 | 1.539 | 0.857 | 1.728 | 0.734 | 1.935 | 0.615 | 2.157 |
| 17 | 1.133 | 1.381 | 1.015 | 1.536 | 0.897 | 1.710 | 0.779 | 1.900 | 0.664 | 2.104 |
| 18 | 1.158 | 1.391 | 1.046 | 1.535 | 0.933 | 1.696 | 0.820 | 1.872 | 0.710 | 2.060 |
| 19 | 1.180 | 1.401 | 1.074 | 1.536 | 0.967 | 1.685 | 0.859 | 1.848 | 0.752 | 2.023 |
| 20 | 1.201 | 1.411 | 1.100 | 1.537 | 0.998 | 1.676 | 0.894 | 1.828 | 0.792 | 1.991 |
| 21 | 1.221 | 1.420 | 1.125 | 1.538 | 1.026 | 1.669 | 0.927 | 1.812 | 0.829 | 1.964 |
| 22 | 1.239 | 1.429 | 1.147 | 1.541 | 1.053 | 1.664 | 0.958 | 1.797 | 0.863 | 1.940 |
| 23 | 1.257 | 1.437 | 1.168 | 1.543 | 1.078 | 1.660 | 0.986 | 1.785 | 0.895 | 1.920 |
| 24 | 1.273 | 1.446 | 1.188 | 1.546 | 1.101 | 1.656 | 1.013 | 1.775 | 0.925 | 1.902 |
| 25 | 1.288 | 1.454 | 1.206 | 1.550 | 1.123 | 1.654 | 1.038 | 1.767 | 0.953 | 1.886 |
| 26 | 1.302 | 1.461 | 1.224 | 1.553 | 1.143 | 1.652 | 1.062 | 1.759 | 0.979 | 1.873 |
| 27 | 1.316 | 1.469 | 1.240 | 1.556 | 1.162 | 1.651 | 1.084 | 1.753 | 1.004 | 1.861 |
| 28 | 1.328 | 1.476 | 1.255 | 1.560 | 1.181 | 1.650 | 1.104 | 1.747 | 1.028 | 1.850 |
| 29 | 1.341 | 1.483 | 1.270 | 1.563 | 1.198 | 1.650 | 1.124 | 1.743 | 1.050 | 1.841 |
| 30 | 1.352 | 1.489 | 1.284 | 1.567 | 1.214 | 1.650 | 1.143 | 1.739 | 1.071 | 1.833 |
| 31 | 1.363 | 1.496 | 1.297 | 1.570 | 1.229 | 1.650 | 1.160 | 1.735 | 1.090 | 1.825 |
| 32 | 1.373 | 1.502 | 1.309 | 1.574 | 1.244 | 1.650 | 1.177 | 1.732 | 1.109 | 1.819 |
| 33 | 1.383 | 1.508 | 1.321 | 1.577 | 1.258 | 1.651 | 1.193 | 1.730 | 1.127 | 1.813 |
| 34 | 1.393 | 1.514 | 1.333 | 1.580 | 1.271 | 1.652 | 1.208 | 1.728 | 1.144 | 1.808 |
| 35 | 1.402 | 1.519 | 1.343 | 1.584 | 1.283 | 1.653 | 1.222 | 1.726 | 1.160 | 1.803 |
| 36 | 1.411 | 1.525 | 1.354 | 1.587 | 1.295 | 1.654 | 1.236 | 1.724 | 1.175 | 1.799 |
| 37 | 1.419 | 1.530 | 1.364 | 1.590 | 1.307 | 1.655 | 1.249 | 1.723 | 1.190 | 1.795 |
| 38 | 1.427 | 1.535 | 1.373 | 1.594 | 1.318 | 1.656 | 1.261 | 1.722 | 1.204 | 1.792 |
| 39 | 1.435 | 1.540 | 1.382 | 1.597 | 1.328 | 1.658 | 1.273 | 1.722 | 1.218 | 1.789 |
| 40 | 1.442 | 1.544 | 1.391 | 1.600 | 1.338 | 1.659 | 1.285 | 1.721 | 1.230 | 1.786 |
| 45 | 1.475 | 1.566 | 1.430 | 1.615 | 1.383 | 1.666 | 1.336 | 1.720 | 1.287 | 1.776 |
| 50 | 1.503 | 1.585 | 1.462 | 1.628 | 1.421 | 1.674 | 1.378 | 1.721 | 1.335 | 1.771 |
| 55 | 1.528 | 1.601 | 1.490 | 1.641 | 1.452 | 1.681 | 1.414 | 1.724 | 1.374 | 1.768 |
| 60 | 1.549 | 1.616 | 1.514 | 1.652 | 1.480 | 1.689 | 1.444 | 1.727 | 1.408 | 1.767 |
| 65 | 1.567 | 1.629 | 1.536 | 1.662 | 1.503 | 1.696 | 1.471 | 1.731 | 1.438 | 1.767 |
| 70 | 1.583 | 1.641 | 1.554 | 1.672 | 1.525 | 1.703 | 1.494 | 1.735 | 1.464 | 1.768 |
| 75 | 1.598 | 1.652 | 1.571 | 1.680 | 1.543 | 1.709 | 1.515 | 1.739 | 1.487 | 1.770 |
| 80 | 1.611 | 1.662 | 1.586 | 1.688 | 1.560 | 1.715 | 1.534 | 1.743 | 1.507 | 1.772 |
| 85 | 1.624 | 1.671 | 1.600 | 1.696 | 1.575 | 1.721 | 1.550 | 1.747 | 1.525 | 1.774 |
| 90 | 1.635 | 1.679 | 1.612 | 1.703 | 1.589 | 1.726 | 1.566 | 1.751 | 1.542 | 1.776 |
| 95 | 1.645 | 1.687 | 1.623 | 1.709 | 1.602 | 1.732 | 1.579 | 1.755 | 1.557 | 1.778 |
| 100 | 1.654 | 1.694 | 1.634 | 1.715 | 1.613 | 1.736 | 1.592 | 1.758 | 1.571 | 1.780 |
| 150 | 1.720 | 1.747 | 1.706 | 1.760 | 1.693 | 1.774 | 1.679 | 1.788 | 1.665 | 1.802 |
| 200 | 1.758 | 1.779 | 1.748 | 1.789 | 1.738 | 1.799 | 1.728 | 1.809 | 1.718 | 1.820 |

Таблица A.13: 5% критические значения теста Durbin–Watson

| | $k$=6 | | $k$=7 | | $k$=8 | | $k$=9 | | $k$=10 | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | dL | dU | dL | dU | dL | dU | dL | dU | dL | dU |
| 6 | — | — | — | — | — | — | — | — | — | — |
| 7 | — | — | — | — | — | — | — | — | — | — |
| 8 | — | — | — | — | — | — | — | — | — | — |
| 9 | — | — | — | — | — | — | — | — | — | — |
| 10 | — | — | — | — | — | — | — | — | — | — |
| 11 | 0.203 | 3.004 | — | — | — | — | — | — | — | — |
| 12 | 0.268 | 2.832 | 0.171 | 3.149 | — | — | — | — | — | — |
| 13 | 0.328 | 2.692 | 0.230 | 2.985 | 0.147 | 3.266 | — | — | — | — |
| 14 | 0.389 | 2.572 | 0.286 | 2.848 | 0.200 | 3.111 | 0.127 | 3.360 | — | — |
| 15 | 0.447 | 2.471 | 0.343 | 2.727 | 0.251 | 2.979 | 0.175 | 3.216 | 0.111 | 3.438 |
| 16 | 0.502 | 2.388 | 0.398 | 2.624 | 0.304 | 2.860 | 0.222 | 3.090 | 0.155 | 3.304 |
| 17 | 0.554 | 2.318 | 0.451 | 2.537 | 0.356 | 2.757 | 0.272 | 2.975 | 0.198 | 3.184 |
| 18 | 0.603 | 2.258 | 0.502 | 2.461 | 0.407 | 2.668 | 0.321 | 2.873 | 0.244 | 3.073 |
| 19 | 0.649 | 2.206 | 0.549 | 2.396 | 0.456 | 2.589 | 0.369 | 2.783 | 0.290 | 2.974 |
| 20 | 0.691 | 2.162 | 0.595 | 2.339 | 0.502 | 2.521 | 0.416 | 2.704 | 0.336 | 2.885 |
| 21 | 0.731 | 2.124 | 0.637 | 2.290 | 0.546 | 2.461 | 0.461 | 2.633 | 0.380 | 2.806 |
| 22 | 0.769 | 2.090 | 0.677 | 2.246 | 0.588 | 2.407 | 0.504 | 2.571 | 0.424 | 2.735 |
| 23 | 0.804 | 2.061 | 0.715 | 2.208 | 0.628 | 2.360 | 0.545 | 2.514 | 0.465 | 2.670 |
| 24 | 0.837 | 2.035 | 0.750 | 2.174 | 0.666 | 2.318 | 0.584 | 2.464 | 0.506 | 2.613 |
| 25 | 0.868 | 2.013 | 0.784 | 2.144 | 0.702 | 2.280 | 0.621 | 2.419 | 0.544 | 2.560 |
| 26 | 0.897 | 1.992 | 0.816 | 2.117 | 0.735 | 2.246 | 0.657 | 2.379 | 0.581 | 2.513 |
| 27 | 0.925 | 1.974 | 0.845 | 2.093 | 0.767 | 2.216 | 0.691 | 2.342 | 0.616 | 2.470 |
| 28 | 0.951 | 1.959 | 0.874 | 2.071 | 0.798 | 2.188 | 0.723 | 2.309 | 0.649 | 2.431 |
| 29 | 0.975 | 1.944 | 0.900 | 2.052 | 0.826 | 2.164 | 0.753 | 2.278 | 0.681 | 2.396 |
| 30 | 0.998 | 1.931 | 0.926 | 2.034 | 0.854 | 2.141 | 0.782 | 2.251 | 0.712 | 2.363 |
| 31 | 1.020 | 1.920 | 0.950 | 2.018 | 0.879 | 2.120 | 0.810 | 2.226 | 0.741 | 2.333 |
| 32 | 1.041 | 1.909 | 0.972 | 2.004 | 0.904 | 2.102 | 0.836 | 2.203 | 0.769 | 2.306 |
| 33 | 1.061 | 1.900 | 0.994 | 1.991 | 0.927 | 2.085 | 0.861 | 2.181 | 0.796 | 2.281 |
| 34 | 1.079 | 1.891 | 1.015 | 1.978 | 0.950 | 2.069 | 0.885 | 2.162 | 0.821 | 2.257 |
| 35 | 1.097 | 1.884 | 1.034 | 1.967 | 0.971 | 2.054 | 0.908 | 2.144 | 0.845 | 2.236 |
| 36 | 1.114 | 1.876 | 1.053 | 1.957 | 0.991 | 2.041 | 0.930 | 2.127 | 0.868 | 2.216 |
| 37 | 1.131 | 1.870 | 1.071 | 1.948 | 1.011 | 2.029 | 0.951 | 2.112 | 0.891 | 2.197 |
| 38 | 1.146 | 1.864 | 1.088 | 1.939 | 1.029 | 2.017 | 0.970 | 2.098 | 0.912 | 2.180 |
| 39 | 1.161 | 1.859 | 1.104 | 1.932 | 1.047 | 2.007 | 0.990 | 2.085 | 0.932 | 2.164 |
| 40 | 1.175 | 1.854 | 1.120 | 1.924 | 1.064 | 1.997 | 1.008 | 2.072 | 0.952 | 2.149 |
| 45 | 1.238 | 1.835 | 1.189 | 1.895 | 1.139 | 1.958 | 1.089 | 2.022 | 1.038 | 2.088 |
| 50 | 1.291 | 1.822 | 1.246 | 1.875 | 1.201 | 1.930 | 1.156 | 1.986 | 1.110 | 2.044 |
| 55 | 1.334 | 1.814 | 1.294 | 1.861 | 1.253 | 1.909 | 1.212 | 1.959 | 1.170 | 2.010 |
| 60 | 1.372 | 1.808 | 1.335 | 1.850 | 1.298 | 1.894 | 1.260 | 1.939 | 1.222 | 1.984 |
| 65 | 1.404 | 1.805 | 1.370 | 1.843 | 1.336 | 1.882 | 1.301 | 1.923 | 1.266 | 1.964 |
| 70 | 1.433 | 1.802 | 1.401 | 1.838 | 1.369 | 1.874 | 1.337 | 1.910 | 1.305 | 1.948 |
| 75 | 1.458 | 1.801 | 1.428 | 1.834 | 1.399 | 1.867 | 1.369 | 1.901 | 1.339 | 1.935 |
| 80 | 1.480 | 1.801 | 1.453 | 1.831 | 1.425 | 1.861 | 1.397 | 1.893 | 1.369 | 1.925 |
| 85 | 1.500 | 1.801 | 1.474 | 1.829 | 1.448 | 1.857 | 1.422 | 1.886 | 1.396 | 1.916 |
| 90 | 1.518 | 1.801 | 1.494 | 1.827 | 1.469 | 1.854 | 1.445 | 1.881 | 1.420 | 1.909 |
| 95 | 1.535 | 1.802 | 1.512 | 1.827 | 1.489 | 1.852 | 1.465 | 1.877 | 1.442 | 1.903 |
| 100 | 1.550 | 1.803 | 1.528 | 1.826 | 1.506 | 1.850 | 1.484 | 1.874 | 1.462 | 1.898 |
| 150 | 1.651 | 1.817 | 1.637 | 1.832 | 1.622 | 1.846 | 1.608 | 1.862 | 1.593 | 1.877 |
| 200 | 1.707 | 1.831 | 1.697 | 1.841 | 1.686 | 1.852 | 1.675 | 1.863 | 1.665 | 1.874 |

# Приложение В

# Список обозначений

На протяжении всей книги использованы следующие обозначения:

| | |
|---|---|
| $\mathsf{E}(\xi)$ | математическое ожидание случайной величины $\xi$ |
| $\mathrm{Var}(\xi)$ | дисперсия случайной величины $\xi$ |
| $\xi \sim F$ | случайная величина $\xi$ имеет распределение $F(x)$ |
| $\mathrm{cov}(\xi, \eta)$ | коэффициент ковариации между случайными величинами $\xi$ и $\eta$; |
| $\mathrm{corr}(\xi, \eta)$ | коэффициент корреляции между случайными величинами $\xi$ и $\eta$; |
| символ $\widehat{\cdot}$ | обозначает выборочное значение (например, $\widehat{\mathrm{Var}}(\xi)$ есть выборочная дисперсия) |
| $\mathcal{N}(a, \sigma^2)$ | нормальное распределение с математическим ожиданием $a$ и дисперсией $\sigma^2$ |
| $\Phi(x)$ | функция стандартного нормального распределения |
| $\chi^2_p$ | распределение хи-квадрат с $p$ степенями свободы |
| $t_p$ | распределение Стьюдента с $p$ степенями свободы ($t$-распределение) |
| $F_{p,q}$ | распределение Фишера со степенями свободы $p$ и $q$ ($F$-распределение) |
| $n$ | объем выборки |
| $k$ | число влияющих переменных в модели регрессии |
| $m$ | число коэффициентов в модели регрессии |

# Литература


[1] Айвазян С.А. Методы эконометрики: учебник – М.:МАГИСТР: ИНФРА-М, 2010. - 512 с.

[2] Артамонов Н.В. Введение в эконометрику.– 2-е изд., испр. и доп.– М.:МЦНМО, 2014. – 224 с.

[3] Берндт Э., *Практика эконометрики. Классика и современность*, М.: Юнити, 2005.

[4] Бородич С.А., *Вводный курс эконометрики*, Учебное пособие, Мн.: БГУ, 2000.

[5] Вербик М., *Путеводитель по современной эконометрике*, М.: Научна книга, 2008.

[6] Доугерти К., Введение в эконометрику, М.: ИНФРА-М, 1997

[7] Магнус Я.Р., Катышев П.К., Пересецкий А.А., *Эконометрика. Начальный курс*, М.: Дело, 2007

[8] Магнус Я.Р., Катышев П.К., Пересецкий А.А., Головань С. В., *Сборник задач к начальному курсу эконометрики* М.: «Дело», 2007

[9] Носко В. П., *Эконометрика для начинающих*, М.: Изд. ИЭПП, 2000.

[10] T. S. Breusch, A. R. Pagan, *A simple test for heteroscedasticity and random coefficient variation* Econometrica, vol. 47, No 5, pp. 1287–1294, 1979.

[11] R. Davidson, J. G. MacKinnon, *Several test for model specification in the presence of alternative hypothesis*, Econometrica, vol. 49, No 3, pp. 781–793, 1981

[12] R. Davidson, J. G. MacKinnon, *Some non-nested hypothesis test and the relations among them*, The Review of Economic Studies, vol. 49, No 4, pp.551–565, 1982

[13] R. Davidson, J. G. MacKinnon, *Econometric theory and methods*, Oxford University Press, USA, 2003

[14] J. Durbin, *Testing for serial correlation in least–squares regression when some of the regressors are lagged dependent variables*, Econometrica, vol 38., No 3, pp. 410–421, 1970

[15] J. Durbin, G. S. Watson, *Testing for serial correlation in least squares regression. I*, Biometrica, vol. 37, No 3/4, pp. 409–428, 1950

[16] J. Durbin, G. S. Watson, *Testing for serial correlation in least squares regression. II*, Biometrica, vol. 38, No 1/2, pp. 159–177, 1951

[17] R. Frisch, *Editor's note*, Econometrica, vol. 1, pp. 1–4, 1933

[18] J. D. Hamilton, *Time series analysis*, Princeton University Press, 1994

[19] S. M. Goldfeld, R. M. Quandt, *Some tests for homoscedasticity*, Journal of the American Statistical Association, Vol. 60, No.310, pp. 539-547, 1965

[20] L. G. Godfrey, *Testing for higher order serial correlation in regression equations when the regressors include lagged dependent variables*, Econometrica, vol. 46, No 6, pp. 1303–1310, 1978

[21] L. G. Godfrey, *Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables*, Econometrica, vol. 46, No 6, pp. 1293–1301, 1978

[22] L. G. Godfrey, *On the use of misspecification checks and tests of non-nested hypotheses in empirical econometrics*, The Economic Journal, vol. 94, pp. 69–81, 1984

[23] W. J. Greene, *Econometric Analysis*, Prentice Hall, 6th edition, 2007

[24] G. S. Maddala, *Introduction to Econometrics*, Second Edition, Macmillian Publishing, 1992

[25] W. K. Newey, K. D. West, *A simple, positive, semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix* Econometrica, v.55, No. 3, pp.703–708, 1987

[26] J. B. Ramsey, *Tests for specification errors in classical linear least-squares regression analysis*, Journal of the Royal Statistical Society. Series B, v.31, No 2, pp.350–371, 1969

[27] G. Shwarz, *Estimating the dimension of a model*, Ann. Statist, 6, pp.461–464, 1978

[28] J. H. Stock, M. W. Watson, *Introduction to Econometrics*, Addison Wesley, 2nd edition, 2006

[29] G. Tintner, *The Definition of Econometrics*, Econometrica, vol. 21, No. 1, pp. 31-40, 1953

[30] H. White, *A heteroskedasticity-consisten covariance matrix estimator and a direct test for heteroskedasticity*, Econometrica, v. 48, No 4, pp.817–838, 1980.

[31] J. M. Wooldridge, *Introductory Econometrics. A modern approach*, Forth Edition. 2009

[32] J. M. Wooldridge, *Econometric analysis of cross section and panel data*, The MIT Press, 2001.

Научное издание

Ивин Евгений Александрович
Артамонов Никита Вячеславович
Курбацкий Алексей Николаевич

**МЕТОДИЧЕСКОЕ ПОСОБИЕ ПО ЭКОНОМЕТРИКЕ (сокращённый вариант)**

для социально-экономических специальностей

| | |
|---|---|
| Редакционная подготовка | Т.В. Алешина |
| Оригинал-макет | И.В. Артамонов |
| Корректор | М.В. Чумаченко |

---

Подписано в печать _ . _ .2016.
Формат бумаги _ × _ / _ . Печать цифровая. Бумага офсетная.
Усл. печ. л. _ , _ Печ. л. _ , _ . Тираж _ экз. Заказ № _

---

Институт социально-экономического развития территорий РАН
(ИСЭРТ РАН)
160014, г. Вологда, ул. Горького, 56а
Телефон (8172) 59-78-03, e-mail: common@vscc.ac.ru