Введение в Эконометрику. Курс лекций

Артамонов Н.В.

6 июня 2010 г.

Оглавление

Введение

Эконометрика является одной из важнейших составляющих современного экономического образования и в ведущих мировых университетах в программах подготовки экономистов ей уделяется большое внимание. Применение эконометрических методов постепенно становится стандартом современных экономических исследований (наряду с теоретико-игровыми методами в микроэкономике). Но не только. Эконометрические методы находят свое применение в политологии, в медицинских исследованиях и других сферах научной деятельности.

Что такое эконометрика? Как и в любой другой области научного знания на этот вопрос нельзя дать однозначного ответа. В редакторской статье в первом выпуске журнала Econometrica (1933) нобелевский лауреат Ranger Frisch пишет [14] (цитируется по переводу в [1]):

> «*Эконометрика это ни в коем случае не тоже самое, что экономическая статистика. Она отнюдь не идентична тому, что мы называем общей экономической теорией, хотя значительная доля этой теории носит определенно количественный характер. Также эконометрика не должна восприниматься как синоним применения математики в экономике. Опыт показывает, что и статистика, и экономическая теория, и математика, взятые по отдельности, являются необходимыми, но не достаточными для действительного понимания количественных отношений в современной экономике. Именно объединение всех трех частей дает мощный эффект. И именно это объединение и составляет эконометрику*»

Lawrence R. Klein, Нобелевский лауреат (1980) видит основную цель эконометрики как (цитируется по [3]):

> «*Основная задача эконометрики – наполнить эмпирическим содержание априорные экономические рассуждения*»

L.R. Klein также пишет, что:

> «*Econometrics had its origin in the recognition of empirical regularities and the systematic attempt to generalize these regularities into "laws" of economics. In a broad sense, the use of such "laws" is to make predictions – about what might have or what will come to pass. Econometrics should give a base for economic prediction beyond experience if it is to be useful. In this broad sense it may be called the science of economic prediction*»

В. Леонтьев характеризует работу эконометриста как:

> «*an attempt to compensate for the glaring weakness of the data base available to us by the widest possible use of more and more sophisticated techniques. Alongside the mounting pile of elaborate theoretical models we see a fast growing stock of equally intricate statistical tools. These are intended to stretch to the limit the meager supply of facts.*»

Приведем еще несколько высказываний:

> «*The method of econometric research aims, essentially, at a conjunction of economic theory and actual measurements, using the theory and technique of statistical inference as a bridge pier*» (Trygve Haavelmo)
>
> «*Econometrics may be defined as the quantitative analysis of actual economic phenomena based on the concurrent development of theory and observation, related by appropriate methods of inference.*» (Samuelson, Koopmans and Stone)
>
> «*Econometrics is concerned with the systematic study of economic phenomena using observed data*». (Aris Spanos)
>
> «*Broadly speaking, econometrics aims to give empirical content to economic relations for testing economic theories, forecasting, decision making, and for ex post decision/policy evaluation*». (J. Geweke, J. Horowitz, and M.H. Pesaran)

В книге В.П. Носко [4] эконометрика определяется как:

> «*Эконометрика – совокупность методов анализа связей между различными экономическими показателями (факторами) на основании реальных статистических данных*

с использованием аппарата теории вероятностей и математической статистики. При помощи этих методов можно выявлять новые, ранее не известные связи, уточнять или отвергать гипотезы о существовании определенных связей между экономическими показателями, предлагаемые экономической теорией.»

О взглядах на предмет и задачи эконометрики других экономистов и эконометристов можно прочитать, например, в эссе G. Tintner [27].

Современную эконометрику можно разделить на два направления: теоретическую и прикладную. Теоретическая эконометрика ориентирована на изучение специальных (абстрактных) вероятностных моделей (как правило т.н. регрессионных моделей) и в этом отношении близка к теории вероятностей и математической статистике и использует их аппарат. В основе прикладной эконометрики лежит применение исследованных вероятностных моделей для количественного описания и анализа экономических явлений и процессов. Вполне естественно, что между этими направления существует глубокая двусторонняя взаимосвязь. Так, новые результаты теоретической эконометрики (например, статистические тесты и новые классы вероятностных моделей) постепенно находят свое применение при решении прикладных задач. С другой стороны, в прикладной эконометрики в процессе исследования экономических явлений возникают ситуации или наблюдаются эффекты, которые не описываются существующими вероятностными моделями. И это способствует и стимулирует дальнейшее развитие теоретического аппарата, рассмотрению и теоретическому исследованию новых вероятностных моделей.

В настоящее время на русском языке издано много хороших современных учебников по эконометрике, как отечественных, так и переводных. В первую очередь стоит отметить учебники Я. Магнуса, П.К. Катышева, А.А. Пересецкого [3] и М. Вербика [2], в которых достаточно полно изложены теоретические основы эконометрики, причем в [3] приведены подробные доказательства вероятностных и статистических свойств эконометрических линейных регрессионных моделей. В книге Э. Берндта [1] обсуждаются вопросы прикладной эконометрики и подробно рассматриваются ставшие уже классическими работы по применению теоретических моделей к различным задачам экономики: нпример, кривая Филлипса, описывающая взаимосвязь безработицы и инфляции в краткосрочном и долгосрочном периодах, и др. Также сто-

ит отметить учебник В.П. Носко [4] (доступный и в электронном виде), прекрасно подходящий для первоначального знакомства с предметом.

В зарубежных учебниках (на английском языке) наиболее полное и последовательное изложение современных эконометрических методов можно найти в книге W.J. Greene [21]. Среди учебников, рассчитанных на первоначальное ознакомление с эконометрикой, стоит в первую очередь отметить книгу J.M. Wooldridge [29], которую характеризует широта охвата, подбор материала, доступный язык изложения и большое количество подробно разбираемых примеров. Для первоначального знакомства также прекрасно подходит книга J.H. Stock и M.W. Watson [26]. В книге J.M. Wooldridge [30] подробно рассматриваются регрессионным модели для пространственных выборок и панельных данных (классы выборочных данных в эконометрике определяются ниже). В книге R. Davidson, J.G. MacKinnon [9] подробно рассматриваются специальные вопросы регрессионных моделей. В книгах J. D. Hamilton [15] (хотя и несколько устаревшей) и W. Enders [13] излагаются эконометрические модели временных рядов.

Настоящий предлагаемый учебник основан на курсе лекций по базовому курсу «Эконометрика» (часто называемому «Эконометрика–1»), читаемых в Московском Государственном Институте Международных Отношений (Университете) МИД России в течение осеннего семестра для студентов третьего курса факультета Международных Экономических Отношений.

Книга рассчитана на студентов, обучающихся по специальности «Экономика» и прослушавших следующие дисциплины: математический анализ, линейная алгебра, теория вероятностей и математическая статистика (включая оценивание параметров распределения, построение доверительных интервалов, проверка статистических гипотез), курс экономической теории (микро- и макроэкономики).

Структура книги

В учебнике отражены следующие разделы, обычно включаемые в начальный курс «Эконометрика–1»:

Линейная однофакторная (парная) модель регрессии. Для простоты изложения все вероятностные и статистические свойства линейной модели регрессии в условиях Гаусса–Маркова (более и менее

строго) доказаны и продемонстрированы на однофакторной линейной модели регрессии, уделено внимание построению доверительных интервалов и проверке статистических гипотез (при разных альтернативах) для коэффициентов регрессии. При этом рассматриваются две вероятностные модели регрессии: с детерминированной и со стохастической влияющей переменной. Основное различие между ними состоит в «регулярности» поведения оценок модели со стохастической влияющей переменной при больших выборках, а именно оценки коэффициентов будут состоятельны и асимптотически нормальны. Обсуждается связь парного коэффициента корреляции с моделью парной регрессии и построение доверительного интервала для парного коэффициента корреляции.

Парная модель регрессии без «свободного члена» или «без константы»: обычно в книгах по эконометрике этой модели не уделяется время и автор решил восполнить этот пробел и посвятить раздел обсуждению этой модели.

Нелинейные однофакторные модели регрессии: особое внимание уделено содержательной экономический интерпретации и экономическому обоснованию применения таких моделей.

Многофакторная линейная модель регрессии. Изложение материала построено так, что эту главу можно читать независимо от парной модели регрессии. В этом разделе строгие полные доказательства вероятностных свойств модели как правило пропущены, так как они требуют использования дополнительного аппарата линейной алгебры и теории вероятностей и при первом чтении могут быть пропущены. Подробные доказательства можно найти в [3, 21]. Также рассматриваются две вероятностные модели: с детерминированными и стохастическими влияющими переменными. Подробно обсуждаются статистические свойства коэффициентов регрессии: эффективность оценок наименьших квадратов, построение доверительных интервалов для коэффициентов, проверка простых статистических гипотез (с двусторонними и односторонними альтернативами), проверка сложных гипотез о коэффициентах регрессии, прогнозирование в рамках модели регрессии, фиктивные (бинарные) переменные. Также рассмотрены асимптотические (при больших выборках) свойства оценок коэффициентов регрессии в модели стохастических влияющих переменных.

Отдельное внимание уделено нелинейным моделям и их содержа-

тельной экономической интерпретации. Как и в случае парной регрессии отдельно рассматривается модель регрессии «без константы».

Отклонения от стандартных условий Гаусса–Маркова. Подробно рассматриваются два наиболее часто встречающихся в приложениях отклонений от стандартных допущений регрессионной модели: неоднородность (гетероскедастичность) и автокоррелируемость ошибок регрессии. Обсуждаются статистические следствия этих отклонений, тесты на выявление этих отклонений и возможные корректировки регрессионной модели.

Спецификация модели. Рассматриваются вопросы, связанные с выбором спецификации модели регрессии. При этом приводятся как и экономические аргументы в пользу той или иной спецификации, так и формальные тесты на спецификацию. Обсуждаются статистические следствия неправильной спецификации модели регрессии.

Введение в регрессионные модели временны́х рядов. В этом разделе кратко рассматриваются особенности построения регрессионных моделей для временных рядов, обобщения условий Гаусса–Маркова для таких моделей, вероятностный и статистические свойства оценок параметров моделей, применимость стандартных тестов, вводится понятие стационарного временного ряда. Рассмотрены статическая регрессионная модель, модель тренда и сезонности, модель распределенных лагов (FDL), модель авторегрессии (AR) стационарных временных рядов, модель распределенных лагов (ADL).

В силу ограничения по времени в курс не включены следующие разделы, иногда включаемые в базовый курс «Эконометрика-1»: модели с бинарной зависимой переменной (Probit- и Logit-модели, линейная вероятностная модель LPM), метод инструментальных переменных (проблема эндогенности), метод максимального правдоподобия оценки параметров линейной модели регрессии и проверки статистических гипотез, системы одновременных уравнений, модели MA (скользящего среднего) и ARMA стационарных временных рядов, модели панельных данных.

В конце каждой главы приведены упражнения по соответствующей тематике. По ряду причин мало упражнений, связанных с непосредственной оценкой модели регрессии по выборочным данным. Бóльшую

часть составляют задачи на анализ уже оцененных регрессионных моделей и теоретические задачи. Задаче по оценке регрессионных моделей по статистическим данным можно найти, например, в [2, 29].

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.

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 простран-

ственных 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 объема 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.

Список обозначений

На протяжении всей книги будут использовать следующие обозначения:

$\mathsf{M}(\xi)$	математическое ожидание случайной величины ξ
$\mathrm{Var}(\xi)$	дисперсия случайной величины ξ
$\xi \sim F$	случайная величина ξ имеет распределение $F(x)$
$\mathrm{cov}(\xi, \eta)$	коэффициент ковариации между случайными величинами ξ и η;
$\mathrm{corr}(\xi, \eta)$	коэффициент корреляции между случайными величинами ξ и η;
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 σ^2
$\Phi(x)$	функция стандартного нормального распределения
χ^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

Благодарности

Автор благодарит д.э.н., проф. Пересецкого А.А. (ЦЭМИ РАН и Российская Экономическая Школа) за полезные обсуждения в процессе подготовки учебника.

Автор благодарит фонд МГИМО–БиПи и лично советника ректора МГИМО (У) МИД России к.полит.н. Мальгина А.В. за помощь в издании книги.

Автор также благодарит д.э.н., проф. Буторину О.В. за помощь и поддержку в процессе работы над книгой.

Глава 1

Парная регрессия

1.1. Парный коэффициент корреляции

1.1.1. Коэффициент корреляции

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

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

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 ± 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

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

т.е. знак коэффициента β_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.

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

Решение этой задачи задается равенствами

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

и наименьшее среднеквадратичное отклонение равно

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

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

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

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 $\mathsf{M}(Y|X = x)$ 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 $\mathsf{M}(Y|X = x)$ случайной величины y уменьшается. В этом случае говорят об **обратной линейной зависимости** между величинами.

3. $\rho = 0$. Тогда $\beta_1^* = 0$, $\mathsf{M}(Y|X = x) = \beta_0^*$ 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* $\operatorname{corr}(X, Y) = 0$.

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: $y = \beta_0^* + \beta_1^* x$.

1.1.2. 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

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

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

Выборочный коэффициент ковариации определяется как[1]

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

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

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.

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 **некоррелируют**. 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 α гипотеза H_0 отвергается в пользу альтернативы H_1 при $|t| > t_{\text{кр}}$,

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

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.

Доверительный интервал для коэффициента корреляции

Задача о построении доверительного интервала для коэффициента корреляции связана с той проблемой, что в общем случае (при $\rho \neq 0$) t-статистика имеет неизвестное распределение. Однако Фишер заметил, что если взять *z-преобразование Фишера*[4] от выборочного коэффициента корреляции

$$z(r) = \frac{1}{2} \ln \frac{1+r}{1-r},$$

то эта статистика при больших объемах выборки (а фактически уже при $n > 6$) имеет распределение, близкое к нормальному

$$z(r) \approx \mathcal{N}\left(z(\rho), \frac{1}{n-3}\right).$$

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

[4] В MS Excel функция ФИШЕР$(\cdot)$

Следовательно, при заданной доверительной вероятности γ приближенный (асимптотический) доверительный интервал для z-преобразования Фишера коэффициента корреляции определяется как

$$\mathsf{P}\left(z(r) - \frac{z_\gamma}{\sqrt{n-3}} < z(\rho) < z(r) + \frac{z_\gamma}{\sqrt{n-3}}\right) \approx \gamma. \qquad (1.1)$$

z_γ есть **двустороннее** критическое значение стандартного нормального распределения при уровне значимости $1-\gamma$ и находится как решение уравнения

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

где $\Phi(x)$ – функция стандартного нормального распределения.

Доверительный интервал для коэффициента корреляции получается применением к интервалу (1.1) *обратного преобразования Фишера*[5]

$$z^{-1}(x) = \mathrm{th}(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}.$$

Таким образом, асимптотический доверительный интервал для коэффициента корреляции имеет вид

$$\mathsf{P}\left(z^{-1}\left(z(r) - \frac{z_\gamma}{\sqrt{n-3}}\right) < \rho < z^{-1}\left(z(r) + \frac{z_\gamma}{\sqrt{n-3}}\right)\right) \approx \gamma.$$

Замечание. Зная доверительный интервал для коэффициента корреляции можно проверить его значимость, т.е. статистическую гипотезу $H_0 : \rho = 0$ при уровне значимости $\alpha = 1-\gamma$. Нулевая гипотеза отвергается тогда и только тогда, когда ноль не принадлежит доверительному интервалу.

Пример. Был рассчитан выборочный коэффициент корреляции $r = 0,68$ между дневными логарифмическими доходностями биржевых индексов NASDAQ и FTSE на основе $n = 100$ выборочных данных. Построим доверительный интервал для коэффициента корреляции с доверительной вероятностью $\gamma = 0.95$. Применим z-преобразование Фишера $z = z(0.68) \approx 0.8291$. Критическое значение z_γ определяется как решение уравнения

$$\Phi(z_\gamma) = \frac{1+0.95}{2} = 0.975,$$

[5] В MS Excel функция ФИШЕРОБР(·)

откуда $z_\gamma = 1.96$. Доверительный интервал для $z(\rho)$ равен

$$\left(0.8291 - \frac{1.96}{\sqrt{100-3}}\ ;\ 0.8291 + \frac{1.96}{\sqrt{100-3}}\right) = (0.6301\ ;\ 1.0281).$$

Применив обратное преобразование Фишера получаем доверительный интервал для коэффициента корреляции

$$\mathsf{P}(0.5581 < \rho < 0.7732) = 0.95$$

($0.5581 = z^{-1}(0.6301)$ и $0.7732 = z^{-1}(1.0281)$).

Проверим значимость коэффициента корреляции, т.е. проверим нулевую гипотезу о **независимости** доходностей обоих биржевых индексов (в предположении их **нормальной распределенности**!). Так как ноль не принадлежит доверительному интервалу, то нулевая гипотеза отвергается при уровне значимости $\alpha = 1 - 0.95 = 0.05$.

1.2. Подгонка прямой. Метод наименьших квадратов

Рассмотрим следующую вспомогательную задачу. 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 β_0 и β_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»? Приведем несколько возможных подходов к определению меры μ отклонения прямой от заданных точек:

1. сумма модулей отклонений в каждой точке x_i:

$$\mu = \sum_{i=1}^{n} |y_i - f(x_i)| = \sum_{i=1}^{n} |y_i - (\beta_0 + \beta_1 x_i)|$$

2. сумма квадратов отклонений в каждой точке x_i:

$$\mu = \sum_{i=1}^{n} (y_i - f(x_i))^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

3. сумма отклонений в каждой точке x_i с заданной весовой функцией $\omega(\cdot) > 0$:

$$\mu = \sum_{i=1}^{n} \omega\big(y_i - f(x_i)\big) = \sum_{i=1}^{n} \omega\big(y_i - (\beta_0 + \beta_1 x_i)\big)$$

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[6]:

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

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

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

[6]Очевидно, $S(\beta_0, \beta_1)$ есть многочлен второго порядка от параметров β_0 и β_1

После простых преобразований приходим к системе линейных уравнений

$$\begin{cases} n\beta_0 + \beta_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \\ \beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i \end{cases} \tag{1.2}$$

называемой системой *нормальных уравнений*. Найдем явные формулы для решения этой системы. Для удобства разделим каждое из уравнение в системе (1.2) 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 β_0 из первого уравнения

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

и подставим во второе уравнение:

$$(\bar{y} - \beta_1 \bar{x})\bar{x} + \beta_1 \left(\overline{x^2}\right) = \overline{xy}.$$

После преобразования получаем (формально)

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

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 (1.2) 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 (1.2) следует, что

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

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

Замечание. Несложно заметить, что система нормальных уравнений (1.2) имеет единственное тогда и только тогда, когда $\widehat{\mathrm{Var}}(x) \neq 0$, т.е. когда не все значения x_i совпадают.

Замечание. 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.

1.3. 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

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n \tag{1.3}$$

где y_i и ε_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 β_1 называется параметром *наклона прямой* (slope), а β_0 – *константой*, *свободным членом* или *параметром сдвига* (intercept).

Уравнение (1.3) называется *уравнение регрессии* или *регрессионным уравнением*, а случайные величины ε_i называются *ошибками* регрессии. Ошибки регрессии удобно представлять себе как «неучтенные факторы», влияющие на y помимо фактора x. Таким образом, уравнение (1.3) отражает наши представления о характере зависимости между факторами.

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

1. $\mathsf{M}\varepsilon_i = 0$, $i = 1, \ldots, n$ (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 $\mathsf{M}\varepsilon_i = 0$ следует, что

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

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) $\mathsf{M}y_i = \beta_0 + \beta_1 x_i$. 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 наблюдения 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.

1.3.1. Теорема Гаусса – Маркова

Итак, мы предполагаем, что зависимость между факторами y и x описывается уравнением регрессии (1.3), но параметры уравнения β_0, β_1 и σ^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[7]. 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 ε_i. Тогда OLS-оценки $\hat{\beta}_0$ и $\hat{\beta}_1$ параметров β_0 и β_1 являются BLUE–оценками, т.е. среди несмещенных линейных (относительно y_i) оценок имеют наименьшую дисперсию.

Доказательство. Докажем несмещенность OLS-оценок. Рассмотрим сначала оценку параметра β_1. Для нее имеем следующее выражение

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

[7] напомним, что оценка параметров вероятностной модели в математической статистике рассматривается как случайная величина

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

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

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

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

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

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

Итак, $\hat{\beta}_0$ и $\hat{\beta}_1$ – несмещенные (unbiased) оценки параметров β_0 и β_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,

$$\mathrm{Var}\left(\hat{\beta}_0\right) = \mathrm{Var}\left(\sum_{i=1}^{n}\left(\frac{1}{n} - \frac{\bar{x}(x_i-\bar{x})}{\sum(x_i-\bar{x})^2}\right)y_i\right) =$$
$$\sum_{i=1}^{n}\left(\frac{1}{n} - \frac{\bar{x}(x_i-\bar{x})}{\sum(x_i-\bar{x})^2}\right)^2\mathrm{Var}(y_i) =$$
$$\sigma^2\sum_{i=1}^{n}\left(\frac{1}{n^2} - 2\frac{\bar{x}(x_i-\bar{x})}{n\sum(x_i-\bar{x})^2} + \frac{(\bar{x})^2(x_i-\bar{x})^2}{\left(\sum(x_i-\bar{x})^2\right)^2}\right) =$$
$$\sigma^2\left(\sum_{i=1}^{n}\frac{1}{n^2} - \frac{2\bar{x}\sum_{i=1}^{n}(x_i-\bar{x})}{n\sum(x_i-\bar{x})^2} + \frac{(\bar{x})^2\sum_{i=1}^{n}(x_i-\bar{x})^2}{\left(\sum(x_i-\bar{x})^2\right)^2}\right) =$$
$$\sigma^2\left(\frac{1}{n} + \frac{(\bar{x})^2}{\sum(x_i-\bar{x})^2}\right) = \frac{\sigma^2\sum x_i^2}{n\sum(x_i-\bar{x})^2} = \frac{\sigma^2\cdot\overline{x^2}}{\sum(x_i-\bar{x})^2}.$$

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 β_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 $\mathsf{M}\widetilde{\beta}_1 = \mathsf{M}\hat{\beta}_1 = \beta_1$, то

$$0 = \mathsf{M}\widetilde{\beta}_1 - \mathsf{M}\hat{\beta}_1 = \mathsf{M}\left(\widetilde{\beta}_1 - \hat{\beta}_1\right)$$
$$= \mathsf{M}\left(\sum\theta_i y_i\right) = \sum\theta_i\mathsf{M}y_i$$
$$= \sum\theta_i(\beta_0 + \beta_1 x_i) = \beta_0\sum\theta_i + \beta_1\sum\theta_i x_i.$$

Так как это равенство должно быть выполнено для произвольных значений β_0 и β_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, $\operatorname{Var}\left(\widetilde{\beta}_1\right) \geqslant \operatorname{Var}\left(\hat{\beta}_1\right)$.

Аналогично можно показать, что для произвольной несмещенной оценки $\widetilde{\beta}_0$ параметра β_0 всегда $\operatorname{Var}\left(\widetilde{\beta}_0\right) \geqslant \operatorname{Var}\left(\hat{\beta}_0\right)$. Теорема доказана. □

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

Замечание. Можно показать, что

$$\operatorname{cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) = -\frac{\sigma^2 \bar{x}}{\sum (x_i - \bar{x})^2}$$

Замечание. Из теоремы Гаусса – Маркова следует, что среди линейных по y несмещенных оценок параметров β_0 и β_1 наилучшими (т.е. с минимальной дисперсией) будут OLS-оценки. Однако могут существовать и нелинейные оценки параметров β_0 и β_1 с дисперсией меньшей, чем у OLS-оценок.

Найдем теперь оценку третьего параметра уравнения регрессии – дисперсии ошибок σ^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 в модели регрессии различать ошибки ε_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 $\sum_{i=1}^{n} e_i = 0$, следующее из первого уравнения системы (1.2), т.е. остатки **всегда зависимы**, в отличие от ошибок регрессии ε_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

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

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

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

1.3.2. Статистические свойства OLS-оценок коэффициентов

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

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

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

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

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 распределение 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*

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

Однако в прикладных задачах дисперсия ошибок регрессии σ^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 χ^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}.

1.3.3. Доверительные интервалы. Проверка гипотез

1. Выведем формулы для доверительных интервалов для коэффициентов регрессии с доверительной вероятностью γ. Для определенности рассмотрим коэффициент наклона β_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 β_1 с доверительной вероятностью γ 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 β_0 с доверительной вероятностью γ 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

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

(θ_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 α. 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 θ_0

- если $|t| < t_{\text{кр}}$, то данные согласуются с нулевой гипотезой при заданном уровне значимости; также говорят, что коэффициент **незначимо** отличается от числа θ_0.

Замечание. Не сложно проверить, что $|t| < t_{\text{кр}}$ тогда и только тогда, когда число θ_0 принадлежит доверительному интервалу для коэффициента β_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 θ_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 β_1 значим (нулевая гипотеза отвергается) при $|t| > t_{\text{кр}}$.
3. Приведем статистический критерий для тестирования гипотезы

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

(θ_0 – заданное значение) против **односторонней** альтернативы

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

Пусть $t'_{\text{кр}} = t(\alpha; n-2)$ – **одностороннее**[8] критическое значение распределения t_{n-2} при заданном уровне значимости α. 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);

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

- если $t < t'_{\text{кр}}$, то данные согласуются с нулевой гипотезой при заданном уровне значимости.

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

1.3.4. 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:

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

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

Определение. *Коэффициент*[9] R^2 определяется как доля объясненной регрессией суммы квадратов в общей сумме квадратов зависимой переменной

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

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

[9] В отечественной литературе коэффициент R^2 иногда называется коэффициентом детерминации

- $R^2 = 0 \Leftrightarrow \mathrm{ESS} = 0 \Leftrightarrow \hat{y}_i = \bar{y} \Leftrightarrow \hat{\beta}_1 = 0$, 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.

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

$$R^2 = \Big(\widehat{\mathrm{corr}}(x, y)\Big)^2,$$

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,

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

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 β_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 α 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

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

1.4. Прогнозирование в модели парной регрессии

Выше мы рассматривали задачу и нахождении 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 β_0, β_1 и σ^2 известны точно. Тогда в качестве прогноза $\hat{y}_{n+1}$ случайной величины y_{n+1} естественно взять ее математическое ожидание

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

Среднеквадратическая ошибка прогноза равна

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

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 γ будем иметь следующее выражение

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

где z_γ 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

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

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 значения является «наилучшим» в смысле среднеквадратичного отклонения.

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

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

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

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

Заменим σ^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 γ 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_γ 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 δ. 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.

1.5. Парная регрессия без константы

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

$$y = \beta x.$$

Найдем оценку параметра β 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 β находится по формуле

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

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.

2. Уравнение прямой без свободного члена используется в некоторых прикладных задачах. Например, эту модель можно использовать в исследовании зависимости дохода от величины налога на доходы.

Итак, рассмотрим модель регрессии

$$y_i = \beta x_i + \varepsilon_i,$$

где значения x_i считаются **неслучайными** (детерминированными) величинами, y_i и ошибки ε_i суть случайные величины. Относительно ошибок регрессии будем предполагать выполнения условий 1. – 4. из парной регрессии с константой. Тогда очевидно

$$\mathsf{M} y_i = \beta x_i, \quad \mathrm{Var}(y_i) = \sigma^2.$$

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

$$y_i = \beta x_i + \varepsilon_i, \tag{1.4}$$

выполнения условий 1. – 3. на ошибки регрессии. Тогда OLS-оценка параметра β является BLUE оценкой, т.е. среди линейных несмещенных оценок имеет минимальную дисперсию (эффективная оценка).

Доказательство. Докажем только несмещенность OLS-оценки:

$$\mathsf{M}\hat{\beta} = \frac{\sum_{i=1}^n x_i \mathsf{M} y_i}{\sum_{i=1}^n x_i^2} = \frac{\sum_{i=1}^n x_i(\beta x_i)}{\sum_{i=1}^n x_i^2} = \beta \frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n x_i^2} = \beta.$$

Эффективность оценки доказывается аналогично модели (1.3). □

Замечание. Из доказательства следует, что для несмещенности OLS-оценки коэффициента в модели регрессии (1.4) достаточно только условия $\mathsf{M}\varepsilon_i = 0$ на ошибки регрессии.

3. Рассмотрим теперь статистические свойства оценки $\hat{\beta}$. Будем предполагать, что ошибки регрессии удовлетворяют условиям 1. – 4. Так как

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

то

$$\operatorname{Var}\left(\hat{\beta}\right) = \frac{\sum_{i=1}^n x_i^2 \operatorname{Var}(y_i)}{\left(\sum_{i=1}^n x_i^2\right)^2} = \frac{\sum_{i=1}^n x_i^2 \sigma^2}{\left(\sum_{i=1}^n x_i^2\right)^2} = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}.$$

Следовательно,

$$\hat{\beta} \sim \mathcal{N}\left(\beta, \frac{\sigma^2}{\sum_{i=1}^n x_i^2}\right).$$

Обозначим через $\hat{y}_i = \hat{\beta} x_i$ *предсказанные* или *подогнанные* (fitted) значения зависимой переменной. *Остатки* регрессии определяются как $e_i = y_i - \hat{y}_i$. Однако, в парной модели регрессии без константы в общем случае $\sum_i e_i \neq 0$.

Обозначим $\mathrm{RSS} = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2$ – *остаточная сумма квадратов* в модели регрессии. Можно показать, что

$$\mathsf{M}(\mathrm{RSS}) = (n-1)\sigma^2.$$

Следовательно, статистика

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

является несмещенной оценкой дисперсии ошибок регрессии σ^2. Выборочная дисперсия оценки $\hat{\beta}$ определяется как

$$\widehat{\mathrm{Var}}(\hat{\beta}) = \frac{s^2}{\sum_{i=1}^{n} x_i^2},$$

а стандартная ошибка коэффициента равна

$$s_1 = \sqrt{\widehat{\mathrm{Var}}(\hat{\beta})} = \sqrt{\frac{s^2}{\sum_{i=1}^{n} x_i^2}}.$$

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

1. *статистики s^2 и $\hat{\beta}$* ***независимы;***
2. *статистика $(n-1)s^2/\sigma^2$ имеет распределение χ^2_{n-1}.*

Далее,

$$\hat{\beta} \sim \mathcal{N}\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}\right) \Longrightarrow \frac{\hat{\beta}-\beta}{\sqrt{\sigma^2/\sum_{i=1}^{n} x_i^2}} \sim \mathcal{N}(0,1).$$

Но тогда статистика

$$t = \frac{\hat{\beta}-\beta}{s_1} = \frac{\hat{\beta}-\beta}{\sqrt{\sigma^2/\sum_{i=1}^{n} x_i^2}} \cdot \frac{1}{s/\sigma}$$

имеет распределение Стьюдента t_{n-1} т.к.

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

Итак, доказана

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

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

имеет распределение Стьюдента t_{n-1}.

4. Выведем формулы для доверительного интервала для коэффициента β с доверительной вероятностью γ. Пусть $t_{\text{кр}} = t_{\text{кр}}(\alpha, n-1)$ – **двустороннее** критическое значение распределения Стьюдента t_{n-1} с уровнем значимости $\alpha = 1 - \gamma$. Тогда из соотношений

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

и

$$|t| < t_{\text{кр}} \Leftrightarrow \left|\frac{\hat{\beta} - \beta}{s_1}\right| < t_{\text{кр}} \Leftrightarrow \hat{\beta} - s_1 \cdot t_{\text{кр}} < \beta < \hat{\beta} + s_1 \cdot t_{\text{кр}}.$$

получаем, что доверительный интервал с доверительной вероятностью γ имеет вид

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

5. Приведем статистический критерия для тестирования гипотезы

$$H_0 \,:\, \beta = \theta_0$$

(θ_0 – заданное значение) против двусторонней альтернативы

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

При справедливости нулевой гипотезы статистика

$$t = \frac{\hat{\beta} - \theta_0}{s_1} \underset{H_0}{\sim} t_{n-1}$$

имеет распределение Стьюдента. Пусть $t_{\text{кр}} = t(\alpha; n-1)$ – **двустороннее** критическое значение распределения Стьюдента t_{n-1} при заданном уровне значимости α. Если верна гипотеза H_0, то вероятность $\mathsf{P}(|t| > t_{\text{кр}}) = \alpha$ мала. Для проверки гипотезы имеем следующее правило:

- 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.

Замечание. Не сложно проверить, что $|t| < t_{\text{кр}}$ тогда и только тогда, когда число θ_0 принадлежит доверительному интервалу для коэффициента β с доверительной вероятностью $1 - \alpha$. Таким образом, мы получаем альтернативный способ проверки нулевой гипотеза:

гипотеза H_0 отвергается при заданном уровне значимости $\Longleftrightarrow$
значение θ_0 не принадлежит доверительному интервалу,
построенному для доверительной вероятности $1 - \alpha$.

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

$$H_0 \,:\, \beta = 0$$

при двусторонней альтернативе t-статистика вычисляется как

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

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

6. Так же как и в случае парной регрессии (1.3) определяются полная (TSS), объясненная (ESS) и остаточная (RSS) суммы квадратов. Однако для модели регрессии (1.4) в общем случае

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

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

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

1.6. 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

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

т.е. среднее значение[10] 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 β_1. Другими словами, маржинальное (предельное) значение y по x в линейной модели регрессии постоянно и равно β_1.

Пример (Wage equation [29]). 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).

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.

[10] под средним значением случайной величины будем подразумевать ее математическое ожидание

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. $\mathsf{M}\varepsilon_i = 0$, то

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

Пусть значение фактора x изменяется на $\triangle_x$:

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

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

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

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 [29]). 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. $\mathsf{M}\varepsilon_i = 0$, то

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

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

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

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

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

Из свойств логарифма следует, что

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

т.е. значение фактора y в среднем изменяется в p^{β_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 β_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 [29]). Была оценена лог-линейная модель зависимости оклада CEO от объема продаж фирмы

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

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

1.7. 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

$$y = \beta_0 + \beta_1 x + u, \tag{1.5}$$

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

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

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

При выполнении условия 1) очевидно

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

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

a) $\mathrm{Var}(u|x) = \sigma^2$,

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

c) $\mathrm{cov}(x, u) = 0$.

Доказательство. По определению условной дисперсии

$$\mathrm{Var}(u|x) = \mathsf{M}(u^2|x) - \big(\mathsf{M}(u|x)\big)^2 = \sigma^2.$$

Из свойств условного математического ожидания

$$\mathsf{M}u = \mathsf{M}_x\big(\mathsf{M}(u|x)\big) = 0$$
$$\mathrm{Var}(u) = \mathsf{M}_x\big(\mathrm{Var}(u|x)\big) = \sigma^2.$$

Далее, так как $\mathsf{M}(xu|x) = x\mathsf{M}(u|x) = 0$, то $\mathsf{M}(xu) = \mathsf{M}_x(\mathsf{M}(xu|x)) = 0$. Следовательно,

$$\mathrm{cov}(x, u) = \mathsf{M}(xu) - \mathsf{M}u \cdot \mathsf{M}x = 0.$$

□

Предложение. *Для модели регрессии* (1.5) *при выполнении условий 1) и 2)*

$$\beta_0 = \mathsf{M}y - \beta_1 \mathsf{M}x, \quad \beta_1 = \frac{\mathrm{cov}(y, x)}{\mathrm{Var}(x)} \quad (\mathrm{Var}(x) \neq 0).$$

Доказательство. Так как $\mathsf{M}u = 0$, то

$$\mathsf{M}y = \mathsf{M}(\beta_0 + \beta_1 x + u) = \beta_0 + \beta_1 \mathsf{M}x + \mathsf{M}u = \beta_0 + \beta_1 \mathsf{M}x$$

и получаем первую формулу. Далее, так как $\mathrm{cov}(x, u) = 0$, то

$$\mathrm{cov}(y, x) = \mathrm{cov}(\beta_0 + \beta_1 x + u, x) = \beta_1 \,\mathrm{cov}(x, x) + \mathrm{cov}(x, u) = \beta_1 \,\mathrm{Var}(x).$$

□

Замечание. Далее будем предполагать, что в модели со стохастическими регрессии выполнено условие $\mathrm{Var}(x) \neq 0$.

Следствие. *Для модели регрессии* (1.5) *при выполнении условий 1) и 2)*

$$\beta_1 = \mathrm{corr}(x, y)\frac{\sigma_y}{\sigma_x},$$

где $\sigma_x = \sqrt{\mathrm{Var}(x)}$ *и* $\sigma_y = \sqrt{\mathrm{Var}(y)}$ *суть стандартные отклонения факторов* x *и* y *соответственно.*

Рассмотрим задачу оценивания параметров β_0, β_1 и σ^2 на основе выборочных данных (x_i, y_i). Основной результат дается следующей теоремой.

Теорема (Гаусс – Марков)**.** *Пусть для линейной модели* (1.5) *выполнены условия 1) и 2) и* (x_i, y_i) *–* ***случайная выборка****. Тогда OLS-оценки* $\hat\beta_0$ *и* $\hat\beta_1$ *параметров* β_0 *и* β_1 *будут линейными несмещенными оценками с минимальной дисперсией*[11]*, т.е. BLUE оценками. Кроме того, эти оценки состоятельны*[12]*, т.е.*

$$\hat\beta_0 \xrightarrow[\mathsf{P}]{} \beta_0, \quad \hat\beta_1 \xrightarrow[\mathsf{P}]{} \beta_1 \ (n \longrightarrow +\infty).$$

Доказательство. **1.** Докажем несмещенность OLS-оценок. Так как

$$\hat\beta_1 = \frac{\sum_{i=1}^n (x_i - \bar x) y_i}{\sum_{i=1}^n (x_i - \bar x)^2} = \frac{\sum_{i=1}^n (x_i - \bar x)(\beta_0 + \beta_1 x_i + u_i)}{\sum_{i=1}^n (x_i - \bar x)^2} = \frac{\beta_0 \sum_{i=1}^n (x_i - \bar x) + \beta_1 \sum_{i=1}^n (x_i - \bar x) x_i + \sum_{i=1}^n (x_i - \bar x) u_i}{\sum_{i=1}^n (x_i - \bar x)^2} = \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar x) u_i}{\sum_{i=1}^n (x_i - \bar x)^2}$$

то

$$\mathsf{M}(\hat\beta_1 | x_1, \ldots, x_n) = \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar x)\mathsf{M}(u_i | x_1, \ldots, x_n)}{\sum_{i=1}^n (x_i - \bar x)^2} = \beta_1.$$

Следовательно, $\mathsf{M}(\hat\beta_1) = \mathsf{M}(\mathsf{M}(\hat\beta_1 | x_1, \ldots, x_n)) = \beta_1$.

[11] имеется ввиду условная дисперсия $\mathrm{Var}(\cdot | x_1, \ldots, x_n)$

[12] напомним, что состоятельность означает сходимость по вероятности: $\hat\beta_j \xrightarrow[\mathsf{P}]{} \beta_j \Leftrightarrow$ для всех $c > 0$ вероятность $\mathsf{P}\left(\left|\hat\beta_j - \beta_j\right| > c\right) \to 0$ при $n \to +\infty$

Далее, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x}$, откуда

$$\begin{aligned}\mathsf{M}(\hat{\beta}_0|x_1,\ldots,x_n) = \mathsf{M}(\bar{y}|x_1,\ldots,x_n) - \bar{x}\mathsf{M}(\hat{\beta}_1|x_1,\ldots,x_n) = \\ \frac{1}{n}\sum \mathsf{M}(y_i|x_1,\ldots,x_n) - \beta_1\bar{x} = \\ \frac{1}{n}\sum(\beta_0 + \beta_1 x_i) - \beta_1\bar{x} = \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} = \beta_0\end{aligned}$$

и $\mathsf{M}(\hat{\beta}_0) = \mathsf{M}(\mathsf{M}(\hat{\beta}_0|x_1,\ldots,x_n)) = \beta_0$.

2. Так же как и в случае детерминированных значений влияющей переменной x доказывается, что среди всех линейных по y оценок OLS-оценки имеют минимальную дисперсию

$$\begin{aligned}\mathrm{Var}\left(\hat{\beta}_0 \middle| x_1,\ldots,x_n\right) = \frac{\sigma^2\bar{x}}{\sum_{i=1}^n (x_i-\bar{x})^2}, \\ \mathrm{Var}\left(\hat{\beta}_1 \middle| x_1,\ldots,x_n\right) = \frac{\sigma^2}{\sum_{i=1}^n (x_i-\bar{x})^2}.\end{aligned}$$

3. Докажем состоятельность OLS-оценок. Имеем

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^n (x_i-\bar{x})^2} = \frac{\frac{1}{n}\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{n}\sum_{i=1}^n (x_i-\bar{x})^2} = \frac{\widehat{\mathrm{cov}}(x,y)}{\widehat{\mathrm{Var}}(x)}.$$

Так как $\widehat{\mathrm{cov}}(x,y) \xrightarrow[\mathsf{P}]{} \mathrm{cov}(x,y)$ (выборочная ковариация – состоятельная оценка ковариации) и $\widehat{\mathrm{Var}}(x) \xrightarrow[\mathsf{P}]{} \mathrm{Var}(x)$ при $n \to +\infty$, то по теореме Слуцкого

$$\hat{\beta}_1 = \frac{\widehat{\mathrm{cov}}(x,y)}{\widehat{\mathrm{Var}}(x)} \xrightarrow[\mathsf{P}]{} \frac{\mathrm{cov}(x,y)}{\mathrm{Var}(x)} = \beta_1.$$

Так как $\bar{y} \xrightarrow[\mathsf{P}]{} \mathsf{M}y$ и $\bar{x} \xrightarrow[\mathsf{P}]{} \mathsf{M}x$, то по теореме Слуцкого

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x} \xrightarrow[\mathsf{P}]{} \mathsf{M}y - \beta_1\mathsf{M}x = \beta_0$$

□

Также 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.

1.8. Задачи

Упражнение 1. На основе опроса 27 семей был вычислен коэффициент корреляции между доходами и расходами на питание: $r = 0.26$. Значимо ли рост доходов влияет на рост расходов на питание семьи (при уровне значимости 2%)? Постройте доверительный интервал для коэффициента корреляции с доверительной вероятностью 98%.

Упражнение 2. По данным 24 магазинов был вычислен коэффициент корреляции между ценой и объемом продаж некоторого товара: $r = -0.37$. Значимо ли увеличение цены влияет на уменьшение объема продаж (при уровне значимости 0.1%)? Постройте доверительный интервал для коэффициента корреляции с доверительной вероятностью 99.9%.

Упражнение 3. На основе опроса 27 семей был вычислен коэффициент корреляции между доходами и накоплениями: $r = 0.62$. Значимо ли рост доходов влияет на рост накоплений (при уровне значимости 10%)? Постройте доверительный интервал для коэффициента корреляции с доверительной вероятностью 90%.

Упражнение 4. По 25 предприятиям был вычислен коэффициент корреляции между объемом продаж и затратами на рекламу: $r = 0.42$. Значимо ли рост затрат на рекламу влияет на рост продаж (при уровне

значимости 1%)? Постройте доверительный интервал для коэффициента корреляции с доверительной вероятностью 99%.

Упражнение 5. На основе выборочных данных за год вычислите выборочный коэффициент корреляции между дневными логарифмическими доходностями биржевых индексов NASDAQ и DAX . Постройте доверительный интервал для коэффициента корреляции с доверительной вероятностью 99%. Проверьте значимость коэффициента корреляции при уровне значимости 1%. Расчеты проведите в MS Excel.

Упражнение 6. На основе выборочных данных за год вычислите выборочный коэффициент корреляции между дневными логарифмическими доходностями биржевых индексов Dow Jones и Nikkei. Постройте доверительный интервал для коэффициента корреляции с доверительной вероятностью 98%. Проверьте значимость коэффициента корреляции при уровне значимости 2%. Расчеты проведите в MS Excel.

Упражнение 7. На основе выборочных данных за год вычислите выборочный коэффициент корреляции между дневными логарифмическими доходностями биржевых индексов Dow Jones и FTSE. Постройте доверительный интервал для коэффициента корреляции с доверительной вероятностью 90%. Проверьте значимость коэффициента корреляции при уровне значимости 10%. Расчеты проведите в MS Excel.

Упражнение 8. Два сотрудника нефтяной компании изучали зависимость объема добычи нефти и мировой цены на нефть. Каждый из них вычислил показатель ковариации и коэффициент корреляции. Первый сотрудник объем добычи считал в баррелях и цену в долларах, а второй – в тоннах и рублях соответственно. Потом они сравнили результаты. Одинаковыми или различными были получены у них результаты? Ответ поясните.

Упражнение 9. Финансовая ситуация вынудила фирму резко сократить расходы на рекламу. В скором времени упали объемы продаж, но в меньшей степени, чем ожидалось. Какому выборочному значению коэффициента корреляции между затратами на рекламу и объемом продаж может соответствовать данная ситуация:

1. $\hat{\rho} = -0.6$;
2. $\hat{\rho} = 0.9$;
3. $\hat{\rho} = 0.5$;

4. $\hat{\rho} = -0.3$?

Ответ обосновать.

Упражнение 10. Какое наименьшее (по абсолютной величине) значение выборочного коэффициента корреляции следует считать значимым на 5% уровне значимости, если объем выборки $n = 38$?

Упражнение 11. Покажите, что $S(\beta_0, \beta_1)$ – выпуклая функция.

Упражнение 12. Докажите равенства

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

Упражнение 13. Для модели регрессии (1.3) докажите равенство

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

Упражнение 14. По 10 наблюдениям показателей x и y были получены следующие данные:

$$\sum x_i = 1700, \quad \sum y_i = 1100, \quad \sum x_i y_i = 204400$$
$$\sum x_i^2 = 316000, \quad \sum y_i^2 = 135000$$

Для модели регрессии (1.3)

1. найдите OLS-оценки коэффициентов регрессии;
2. вычислите стандартную ошибку регрессии $\mathrm{SER} = \sqrt{s^2}$;
3. найдите выборочные стандартные ошибки коэффициентов регрессии;
4. проверьте значимость коэффициентов регрессии;
5. значимо ли коэффициент β_1 отличается от 1?

Упражнение 15. В условиях предыдущей задачи для модели регрессии без константы (1.4)

1. найдите OLS-оценку коэффициента регрессии;
2. вычислите стандартную ошибку регрессии $\mathrm{SER} = \sqrt{s^2}$;

3. найдите выборочную стандартную ошибку коэффициента;

4. проверьте значимость коэффициента регрессии;

5. значимо ли коэффициент регрессии отличается от 1?

Упражнение 16. По $n = 18$ магазинам была оценена модель регрессии зависимости объема продаж от цены (в \$100)

$$\widehat{Sales} = 32.2 - 2.2Price, \quad s_1 = 0.02.$$

- Дайте интерпретацию коэффициентов модели.

- Значимо ли коэффициент β_1 отличается от (-2) при уровне значимости 1%? 2%? 5%? 10%?

- Какой ожидаемый уровень продаж при цене \$96, \$107, \$102, \$92?

Упражнение 17. По 20 выборочным данным была оценена модель регрессии

$$\widehat{y} = 2.3 + 0.7x, \quad s_0 = 0.02, \ s_1 = 0.2.$$

Постройте доверительные интервалы для коэффициентов регрессии доверительной вероятностью 95%, 98%, 90%, 99%.

Упражнение 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)

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

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 одинадцатилетним образованием? С девятилетним образование?

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

$$\widehat{\ln(Sales)} = \underset{(0.07)}{0.91} - \underset{(0,2)}{1.21}\ln(Price) \quad n = 25$$

(в скобках указаны стандартные ошибки коэффициентов).

- Дайте интерпретацию коэффициентов модели. 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%?

Упражнение 20. По 20 наблюдениям было получено следующее уравнение регрессии:

$$\hat{y} = 3 + 2x, \quad t = \frac{\hat{\beta}_1}{s_1} = 6.48.$$

Найдите коэффициент R^2.

Упражнение 21. В таблице приведены данные промежуточного среза ($midterm$) и финального экзамена ($exam$) 12-ти случайно отобранных студентов

№	1	2	3	4	5	6	7	8	9	10	11	12
$midterm$	62	36	82	97	77	55	93	48	72	83	75	96
$exam$	70	30	79	99	76	47	95	51	76	90	67	99

- Найдите OLS-оценки параметров линейной регрессии $exam$ на $midterm$. Дайте интерпретацию коэффициентов регрессии.
- Найдите стандартную ошибку регрессии SER.
- Найдите стандартные ошибки коэффициентов регрессии.
- Проверьте значимость коэффициента наклона прямой (уровень значимости 1%, 5%, 10%).

- Значимо ли коэффициент наклона отличается от 1? Рассмотрите уровни значимости 1%, 5%, 10%.
- Вычислите коэффициент R^2 и дайте его интерпретацию.
- Какая ожидаемая оценка студента за финальный экзамен, если на промежуточном срезе он получил оценку 92? 75?
- Постройте доверительный интервал для оценки студента за финальный экзамен, если на промежуточном срезе он получил 80, 75, 96. Рассмотрите случаи доверительных вероятностей 90%, 95%, 99%.

Расчеты проведите в MS Excel, EViews или STATA.

Упражнение 22. В таблице приведены данные промежуточного среза ($midterm$) и финального экзамена ($exam$) 15-ти случайным образом отобранных студентов

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$midterm$	62	36	82	97	77	55	93	48	72	83	75	96	21	9	18
$exam$	70	30	79	99	76	47	95	51	76	90	67	99	25	11	16

- Оцените регрессию $exam$ на $midtrem$ без константы

$$exam_i = \beta \cdot midterm_i + \varepsilon_i$$

  Дайте интерпретацию коэффициента регрессии.
- Постройте доверительный интервал для коэффициента наклона с доверительной вероятностью 90%. Значим ли коэффициент при уровне значимости 10%?
- Значимо ли коэффициент наклона отличается от 1 при уровне значимости 5%.
- Вычислите $R^2_{\text{нецентр}}$
- На промежуточном срезе студен получил 85 баллов. Какая ожидаемая оценка студента за финальный экзамен?

Для расчетов используйте MS Excel, EViews или STATA.

Упражнение 23. Дилер автосалона продал 10 подержанных автомобилей VW Golf по следующим ценам (в \$1000)

Age	1	1	2	3	3	4	4	5	6	6
Price	18.0	16.5	15.0	15.6	16.0	14.0	13.9	11.0	11.3	10.8

Для описания зависимости цена автомобиля от его возраста была выбрана линейная модель регрессии.

- Найдите OLS-оценки коэффициентов этой модели и дайте их интерпретацию.
- Какая ожидаемая цена семилетнего автомобиля? Десятилетнего автомобиля?
- Постройте доверительный интервал для цены семилетнего и десятилетнего автомобиля (с доверительной вероятностью 90%, 95%, 99%).

Для расчетов используйте MS Excel, EViews или STATA.

Упражнение 24. В условиях предыдущей задачи

- оцените полулогарифмическую регрессионную модель зависимости и интерпретируйте полученные значения.
- Какая ожидаемая цена семилетнего автомобиля? Десятилетнего автомобиля?
- Постройте доверительный интервал для цены семилетнего и десятилетнего автомобиля (с доверительной вероятностью 90%, 95%, 99%).

Для расчетов используйте MS Excel, EViews или STATA.

Упражнение 25. Пусть $\hat{\beta}$ есть OLS-оценка коэффициента наклона в линейной регрессии без константы y на x, а $\hat{\gamma}$ – OLS-оценка коэффициента наклона в линейной регрессии без константы x на y. Верно ли для этих оценок равенство

$$\hat{\gamma} = \frac{1}{\hat{\beta}}?$$

Упражнение 26. Пусть $\widehat{\beta}_1$ есть OLS-оценка коэффициента наклона в линейной регрессии с константой y на x, а $\widehat{\gamma}_1$ – OLS-оценка коэффициента наклона в линейной регрессии с константой x на y. Покажите, что

$$\widehat{\gamma}_1 = \frac{1}{\widehat{\beta}_1} \Longleftrightarrow R^2 = 1.$$

Упражнение 27. Пусть $\widehat{\beta}_0, \widehat{\beta}_1$ – OLS-оценки коэффициентов в регрессии y на x, а $\widetilde{\beta}_0, \widetilde{\beta}_1$ – OLS-оценки коэффициентов в регрессии $(c_1 y)$ на $(c_2 x)$ $(c_1, c_2 \neq 0)$. Покажите, что

$$\widetilde{\beta}_1 = \frac{c_1}{c_2} \cdot \widehat{\beta}_1, \quad \widetilde{\beta}_0 = c_1 \widehat{\beta}_0.$$

Упражнение 28. Пусть $\widehat{\beta}_0, \widehat{\beta}_1$ – OLS-оценки коэффициентов в регрессии y на x, а $\widetilde{\beta}_0, \widetilde{\beta}_1$ – OLS-оценки коэффициентов в регрессии $(y + c_1)$ на $(x + c_2)$. Покажите, что

$$\widetilde{\beta}_1 = \widehat{\beta}_1, \quad \widetilde{\beta}_0 = \widehat{\beta}_0 + c_1 - c_2 \widehat{\beta}_1.$$

Упражнение 29. Пусть $\widehat{\beta}_0, \widehat{\beta}_1$ – OLS-оценки коэффициентов в регрессии $\ln(y)$ на x, а $\widetilde{\beta}_0, \widetilde{\beta}_1$ – OLS-оценки коэффициентов в регрессии $\ln(cy)$ на x $(c > 0)$. Найдите соотношения между этими оценками.

Упражнение 30. Пусть $\widehat{\beta}_0, \widehat{\beta}_1$ – OLS-оценки коэффициентов в регрессии y на $\ln x$, а $\widetilde{\beta}_0, \widetilde{\beta}_1$ – OLS-оценки в регрессии y на $\ln(cx)$ $(c > 0)$. Найдите соотношения между этими оценками.

Упражнение 31. В линейной модели регрессии с константой (1.3) рассмотрим оценку коэффициента наклона β_1

$$\widehat{\beta}_1' = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i - \bar{y}}{x_i - \bar{x}}.$$

Будет ли эта оценка

1. линейной по y;
2. несмещенной;
3. наилучшей (с наименьшей дисперсией)?

Найдите $\mathrm{Var}\left(\widehat{\beta}_1'\right)$.

Упражнение 32. В линейной модели регрессии с константой (1.3) рассмотрим оценку коэффициента наклона β_1

$$\widehat{\beta}_1'' = \frac{\sum (\omega_i - \overline{\omega})\, y_i}{\sum (\omega_i - \overline{\omega})\, x_i}.$$

($\{\omega_i\}_1^n$ – произвольный набор чисел). Будет ли эта оценка

1. линейной по y;
2. несмещенной;
3. наилучшей (с наименьшей дисперсией)?

Найдите $\operatorname{Var}\left(\widehat{\beta}_1''\right)$.

Упражнение 33. Для линейной модели регрессии без константы (1.4) рассмотрим оценку коэффициента наклона β

$$\hat{\beta}' = \frac{\sum y_i}{\sum x_i}.$$

будет ли эта оценка

1. линейной по y;
2. несмещенной;
3. наилучшей (с минимальной дисперсией)?

Найдите $\operatorname{Var}\left(\widehat{\beta}'\right)$.

Упражнение 34. Для линейной модели регрессии без константы (1.4) рассмотрим оценку коэффициента наклона β

$$\hat{\beta}'' = \frac{\sum \omega_i y_i}{\sum \omega_i x_i}$$

($\{\omega_i\}_1^n$ – произвольный набор чисел). Будет ли эта оценка

1. линейной по y;
2. несмещенной;

3. наилучшей (с минимальной дисперсией)?

Найдите $\mathrm{Var}\left(\widehat{\beta}''\right)$.

Упражнение 35. Пусть $\hat{y}_i$ – подогнанные значения в модели регрессии (1.3). Найдите OLS-оценки коэффициентов в модели регрессии

$$y_i = \gamma_0 + \gamma_1 \hat{y}_i + \mathrm{error}\,.$$

Упражнение 36. Может ли в регрессии $y = \beta x + u$ OLS-оценка $\widehat{\beta}$ быть положительной, а оценка коэффициента $\widehat{\beta}$ в модели регрессии $y = \alpha + \beta x + w$ отрицательной (на одной и той же выборке)?

Упражнение 37. На основе 100 данных была оценена функция спроса

$$\widehat{\ln(Q)} = \underset{(0.04)}{0.87} - \underset{(0.2)}{1.23}\ln(Price)$$

Значимо ли коэффициент эластичности отличается от (-1)? Рассмотрите уровень значимости 10%, 5%, 1%.

Упражнение 38. На основе 100 данных была оценена функция спроса

$$\widehat{\ln(Q)} = \underset{(0.04)}{2.87} - \underset{(0.2)}{1.12}\ln(P)$$

Проверьте гипотезу

$$H_0 \ : \ \beta_1 = -1$$

против альтернативы

$$H_1 \ : \ \beta_1 < -1.$$

Рассмотрите уровень значимости 10%, 5%, 1%. Дайте интерпретацию проверяемой гипотезе и альтернативе.

Упражнение 39 ([29]). По данным с 1960 по 2005 гг. была оценена (статическая) кривая Филлипса, связывающая уровень инфляции inf_t и уровень безработицы $unem_t$:

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

Тестируйте гипотезу

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

против альтернативы

$$H_1 \ : \ \beta_1 < 0$$

при уровне значимости 1%. Дайте объяснение проверяемой нулевой гипотезе и альтернативе.

Упражнение 40. Может ли в парной модели регрессии $\hat{y} = \beta_0 + \beta_1 x$ коэффициента R^2 быть «малым», а t-статистика $t_1 = \hat{\beta}_1 / s_1$ «большой»?

Упражнение 41. Модель парной регрессии $\hat{y} = 6.7 + 0.6x$ была оценена по выборке объема 27. Выборочные стандартные отклонения регрессора и зависимой переменной равны $\hat{\sigma}_x = 6$ и $\hat{\sigma}_y = 9$. Проверьте значимость коэффициента регрессии (уровень значимость 1%) и сформулируйте проверяемую статистическую гипотезу.

Упражнение 42. По выборке объема 22 был вычислен выборочный парный коэффициент корреляции $\widehat{\text{corr}}(x, y) = -0.5$. Выборочные стандартные отклонения регрессора и зависимой переменной равны $\hat{\sigma}_x = 4$ и $\hat{\sigma}_y = 10$. Чему равен выборочный коэффициент наклона $\hat{\beta}_1$ в модели парной регрессии? Проверьте значимость коэффициента β_1 при уровне значимости 10% и сформулируйте проверяемую статистическую гипотезу.

Упражнение 43. Была оценена регрессионная модель

$$\widehat{AWE} = 696.7 + 9.6Age \quad R^2 = 0.023 \quad s = 624.1$$

зависимости средней недельной зарплаты (AWE, в \$) от возраста (Age, в годах) для случайной выборки 25 – 65-летних рабочих с полным средним образованием.

- Дайте интерпретацию коэффициентов модели.
- Какая, согласно модели, ожидаемая зарплата 25-летнего рабочего? 45-летнего?
- Дает ли эта модель регрессии удовлетворительный прогноз для 99-летнего человека?
- Средний возраст по выборке равен 41.6 лет. Какое среднее значение для AWE?
- Какие единицы измерения (или безразмерные) величин s и R^2?

Упражнение 44. Был проведен эксперимент по измерению влияния ограничения по времени (отведенного на выполнения экзаменационного задания) на оценки за финальный экзамен. 400 студентам было дано одно и тоже задание, но одним студентам было дано 90 минут на

выполнение задания, а другим – 120 минут. Для каждого студента время выполнения задания назначалось случайным образом (бросанием монетки). Пусть $Score$ – оценка за финальный экзамен, $Time$ – время выполнения экзамена ($Time = 90$ или 120) и была выбрана линейная регрессионная модель

$$\widehat{Score} = 49 + 0.24Time$$

- Объясните почему $\mathsf{M}\varepsilon_i = 0$ для этой модели регрессии.
- Какая средняя экзаменационная оценка студента, если на выполнение экзаменационного задания ему дано 90 минут, 120 минут, 150 минут?
- На сколько в среднем изменится оценка за экзамен, если студенту дополнительно дали 10 минут на выполнение задания?

Глава 2

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*, ε – *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 ε 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

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

2.1. Метод наименьших квадратов

Найдем оценку наименьших квадратов для вектора параметров β при известных значениях зависимой переменной y и регрессоров $x_1, \ldots, x_k$. Для простоты изложения рассмотрим сначала случай двухфакторной модели регрессии. Итак, по заданным значениям $\{y_i, x_{i1}, x_{i2}\}$ необходимо найти плоскость

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

«меньше всего отклоняющуюся» от заданных точек. Оценки параметров уравнения находятся как решение экстремальной задачи

$$S = S(\beta_0, \beta_1, \beta_2) = \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2})\big)^2 \longrightarrow \min.$$

Согласно необходимым условиям существования экстремума значения параметров оптимального уравнения являются решением системы уравнений

$$\frac{\partial S}{\partial \beta_0} = (-2) \sum_i (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2}) = 0$$

$$\frac{\partial S}{\partial \beta_1} = (-2) \sum_i (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2}) x_{i1} = 0$$

$$\frac{\partial S}{\partial \beta_2} = (-2) \sum_i (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2}) x_{i2} = 0$$

После преобразования получаем систему линейных уравнений, называемую *системой нормальных уравнений*:

$$\begin{cases} n\beta_0 + \beta_1 \sum_i x_{i1} + \beta_2 \sum_i x_{i2} = \sum_i y_i \\ \beta_0 \sum_i x_{i1} + \beta_1 \sum_i x_{i1}^2 + \beta_2 \sum_i x_{i1}x_{i2} = \sum_i y_i x_{i1} \\ \beta_0 \sum_i x_{i2} + \beta_1 \sum_i x_{i1}x_{i2} + \beta_2 \sum_i x_{i2}^2 = \sum_i y_i x_{i2} \end{cases}$$

Несложно показать, что функция $S(\cdot)$ выпукла (как функция многих переменных). Следовательно, решение системы нормальных уравнений будет решением экстремальной задачи.

Рассмотрим теперь общий случай. Сумму квадратов отклонений запишем в матричном 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 $S(\cdot)$.

2.2. Основные предположения. Теорема Гаусса – Маркова

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

1. $\mathsf{M}\varepsilon_i = 0$ 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

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

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

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)

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

При изменении значения фактора x_j на величину Δ_j среднее значение зависимой переменной изменяется на величину $\Delta\mathsf{M}y = \beta_j \Delta_j$. Коэффициент β_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 теперь, 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,

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

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

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

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

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

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

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

Так как $\widetilde{\beta}$ – несмещенная оценка ($\mathsf{M}\widetilde{\beta} = \beta$), то $Q\mathbf{X} = 0$. Используя это наблюдение можно показать [3], что

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

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

Замечание. Из доказательства видно, что для несмещенности OLS-оценок достаточно выполнения условия $\mathsf{M}\varepsilon_i = 0$ 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 β_j имеет нормальное распределение

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

Укажем теперь несмещенную оценку дисперсии ошибок регрессии σ^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 $e_i = y_i - \hat{y}_i$.

Важно в модели регрессии различать ошибки ε_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 равенство $\sum_{i=1}^{n} e_i = 0$, (т.к. в модель включена константа β_0), т.е. остатки **всегда зависимы**, в отличие от ошибок регрессии ε_i. Но, 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.

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

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

Следовательно, статистика

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

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

2.3. 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 σ^2 – дисперсия ошибок регрессии. Матрица ковариации вектора OLS-оценок коэффициентов модели регрессии равна

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

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

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

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

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

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

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

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. *Пустить 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 χ^2_{n-m},

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

Теорема доказывается также как и в случае парной регрессии.

2. Приведем формулы доверительных интервалов для коэффициентов регрессии с доверительной вероятностью γ. 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 β_j с доверительной вероятностью γ:

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

3. Приведем статистический критерий для проверки гипотезы

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

(θ_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 α. 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 θ_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 θ_0.

Замечание. Несложно проверить, что $|t| < t_{\text{кр}}$ тогда и только тогда, когда число θ_0 принадлежит доверительному интервалу для коэффициента β_j с доверительной вероятностью $1 - \alpha$. Таким образом, мы получаем альтернативный способ проверки нулевой гипотеза:

гипотеза H_0 отвергается при заданном уровне значимости $\Longleftrightarrow$ значение θ_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 β_j **значим** (нулевая гипотеза отвергается) при $|t_j| > t_{\text{кр}}$.

4. Приведем статистический критерий для тестирования гипотезы

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

(θ_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 α. 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 $\mathsf{P}(t_{n-m} > t'_{\text{кр}}) = \alpha$

2.4. Коэффициент R^2. Проверка сложных гипотез о коэффициентах регрессии

1. Как и в случае парной регрессии введем обозначения

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

Для модели регрессии с включенной константой β_0 верно равенство

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

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

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

Из определения непосредственно следует, что всегда $0 \leqslant R^2 \leqslant 1$. Кроме того, для крайних 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 по коэффициенту детерминации необходимо учитывать два важных обстоятельства:

[2]Иногда коэффициент R^2 называется коэффициентом детерминации

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 β_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

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

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 α. 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.

Замечание. На первый взгляд может показаться, что для проверки значимости **всех** коэффициентов модели регрессии достаточно последовательно проверить значимость каждого коэффициента в отдельности (проверить серию статистических гипотез). При проверки одной статистической гипотезы вероятность неверно отвергнуть нулевую гипотезу (вероятность ошибки первого рода или уровень значимости) мала. При проверке нескольких статистических гипотез вероятности ошибок как правило суммируются и суммарная вероятность ошибки первого рода при таком подходе уже может быть большой. Может случиться так, что каждый коэффициент регрессии в отдельности незначим, но регрессия «в целом» значима.

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

$$H_1 \ : \ \beta_{k-q+1}^2 + \cdots + \beta_k^2 > 0$$

(не все коэффициенты равны нулю).

Обозначим R_{ur}^2 – коэффициент R^2 и 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_r^2 – коэффициент R^2 и 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

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

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 α. 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.

Пример (Wage-equation). Рассмотрим регрессионное уравнение зависимости логарифма зарплаты $\ln(Wage)$ от уровня школьного образования Edu, возраста Age и уровня школьного образования родителей $Fedu$, $Medu$ (модель без ограничений)

$$\widehat{\ln(Wage)} = \beta_0 + \beta_1 Edu + \beta_2 Age + \beta_3 Age^2 + \beta_4 Fedu + \beta_5 Medu.$$

Тогда для проверки статистической гипотезы

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

(проверки значимости влияния образования родителей на зарплату) необходимо вычислить коэффициент R_r^2 или остаточную сумму квадратов RSS_r в модели регрессии с ограничениями, налагаемыми нулевой гипотезой, а именно в регрессии **без факторов** $Fedu$ и $Medu$ (без учета уровня образования родителей)

$$\widehat{\ln(Wage)} = \beta_0 + \beta_1 Edu + \beta_2 Age + \beta_3 Age^2.$$

В этом примере $q = 2$ (на коэффициенты исходной модели наложено два ограничения, а именно приравниваем к нулю два коэффициента) и $m = 6$ (в исходной модели шесть коэффициентов).

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 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 RSS_r – остаточную сумму квадратов в модели регрессии с ограничениями (r = restricted), налагаемыми проверяемой нулевой гипотезой.

Легко видеть, что имеют место неравенства

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

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

$$F = \frac{R^2_{ur} - R^2_r}{1 - R^2_{ur}} \cdot \frac{n-m}{q} = \frac{\mathrm{RSS}_r - \mathrm{RSS}_{ur}}{\mathrm{RSS}_{ur}} \cdot \frac{n-m}{q}. \qquad (2.2)$$

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 $F_{q,n-m}$ с уровнем значимости α. Получаем 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.

Пример (Wage-equation). Рассмотрим регрессионное уравнение зависимости логарифма зарплаты $\ln(Wage)$ от уровня школьного образования Edu, возраста Age и уровня школьного образования родителей $Fedu$, $Medu$ (модель без ограничений)

$$\widehat{\ln(Wage)} = \beta_0 + \beta_1 Edu + \beta_2 Age + \beta_3 Age^2 + \beta_4 Fedu + \beta_5 Medu.$$

Для проверки статистической гипотезы

$$H_0 : \beta_4 = \beta_5$$

необходимо вычислить коэффициент R^2_r или остаточную сумму квадратов RSS_r в модели регрессии с ограничением, налагаемым нулевой гипотезой

$$\widehat{\ln(Wage)} = \beta_0 + \beta_1 Edu + \beta_2 Age + \beta_3 Age^2 + \beta_4 (Fedu + Medu).$$

В этом примере $q = 1$ (на коэффициенты исходной модели наложено одно линейное ограничение) и $m = 6$ (в исходной модели шесть коэффициентов).

Таким образом, для вычисления F-статистики для проверки сложной нулевой гипотезы необходимо оценить две модели регрессии: с ограничения, 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

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

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

$$\begin{aligned} y_i &= \mathbf{x}_i'\beta + \varepsilon_i, \qquad & \mathrm{Var}(\varepsilon_i) &= \sigma_1^2 \qquad & i &= 1 \ldots n_1, \\ y_i &= \mathbf{x}_i'\gamma + \nu_i, \qquad & \mathrm{Var}(\nu_i) &= \sigma_2^2 \qquad & i &= n_1 + 1 \ldots n_1 + n_2. \end{aligned}$$

Пусть RSS_1 и 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 α нулевая гипотеза отвергается при $F > F_{\text{кр}}$, где $F_{\text{кр}} = F(\alpha; m, n_1 + n_2 - 2m)$

2.5. 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 σ^2 известны точно. Тогда естественно в качестве прогноза $\hat{y}_{n+1}$ величины y_{n+1} взять ее математическое ожидание

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

среднеквадратическая ошибка этого прогноза равна

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

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 γ будем иметь следующее выражение

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

где z_γ 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

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

Кроме того, так как OLS-оценки коэффициентов регрессии линейны по y, то и прогноз $\hat{y}_{n+1}$ также линеен по y. Также 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 $\mathsf{M}\widetilde{y}_{n+1} = \mathsf{M}y_{n+1} = \mathbf{x}'_{n+1}\beta$. Тогда*

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

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

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

Заменив σ^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 γ 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_γ 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 δ, характеризующего длину доверительного интервала.

2.6. Множественная регрессия без константы

Рассмотрим теперь линейную модель многофакторной регрессии без константы:

$$y_i = \beta_1 x_{i1} + \cdots + \beta_k x_{im} + \varepsilon_i, \quad i = 1, \ldots, n \tag{2.3}$$

y – *зависимая переменная*, $x_1, \ldots, x_m$ – *регрессоры* или *объясняющие переменные*, ε – ошибки модели регрессии, $\beta_1, \ldots, \beta_m$ – параметры или коэффициенты в модели регрессии, i – номер наблюдения. В этой модели число коэффициентов m совпадает с числом объясняющих переменных. Значения регрессоров будем считать неслучайными (детерминироваными).

Запишем уравнение регрессии (2.3) в матричном виде. Для этого введем следующие обозначения

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

Тогда уравнение (2.3) можно записать в виде

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

Обозначим

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

Из определения видно, что $\mathbf{X}$ есть 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 (2.3) могут быть записаны в виде одного матричного уравнения

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

1. Найдем оценки коэффициентов линейной модели регрессии по методу наименьших квадратов. По аналогии с моделью (2.1) сумму квадратов отклонений запишем в матричном виде

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

Тогда условия первого порядка (равенство нулю первых производных по переменным β_j) приводят к системе нормальных уравнений, записываемой в матричном виде как

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

Несложно показать, что $S(\cdot)$ – выпуклая функция нескольких переменных. Следовательно, решение системы нормальных уравнений будет решением экстремальной задачи.

При $\det(\mathbf{X}'\mathbf{X}) \neq 0$ система нормальных уравнений имеет единственное решение и OLS-оценка коэффициентов модели регрессии (2.3) равна

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

2. Рассмотрим теперь вероятностные свойства многофакторной линейной модели регрессии. Покажем теперь, что OLS-оценки параметров $\widehat{\beta}_{OLS}$ являются «наилучшими» среди несмещенных линейных оценок, а именно несмещенными линейными оценками с наименьшей дисперсией, т.е. 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. Аналогично модели регрессии (2.1) с константой. □

Если кроме условий 1. – 3. выполнено условие 4. нормальной распределенности ошибок регрессии, то 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

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

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

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} b_1 & \cdots & \cdots & \star \\ \vdots & b_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ \star & \cdots & \cdots & b_m \end{pmatrix}.$$

Тогда оценка $\hat{\beta}_j$ коэффициента регрессии β_j имеет нормальное распределение

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

Укажем теперь несмещенную оценку дисперсии ошибок регрессии σ^2. Обозначим

$$\hat{y}_i = \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_m x_{im}$$

предсказанные или подогнанные (fitted) значения зависимой переменной. В матричной записи

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

Определение. *Остатки* в модели регрессии (2.3) определяются как $e_i = y_i - \hat{y}_i$.

Замечание. Важно отметить, что для модели регрессии (2.3) в общем случае $\sum e_i \neq 0$, так как в модель регрессии не включена константа β_0.

Также как в случае модели регрессии (2.1) 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

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

Следовательно, статистика

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

является несмещенной оценкой дисперсии ошибок линейной регрессии.

3. Рассмотрим теперь статистические свойства OLS-оценок параметров регрессии. Оценка дисперсии OLS-оценок коэффициентов регрессии равна

$$\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right) = s^2 b_j \quad (j = 1, \ldots, m),$$

а выборочное стандартное отклонение или стандартная ошибка коэффициента регрессии определяется как

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

Оценка ковариационной матрицы вектора OLS-оценок коэффициентов регрессии определяется как

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

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

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

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

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

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

Следовательно, для построения доверительных интервалов для коэффициентов регрессии и проверки статистических гипотез

$$H_0 : \beta_j = \theta_0$$

надо использовать формулы модели регрессии с включенной константой β_0.

4. Для модели регрессии без константы (2.3) в общем случае

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

и коэффициент детерминации уже **не имеет смысла**.

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

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

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

$$H_0 \ : \ \beta_{m-q+1} = \cdots = \beta_m = 0.$$

при альтернативе

$$H_1 \ : \ \beta^2_{m-q+1} + \cdots + \beta^2_m > 0.$$

(не все коэффициенты равны нулю).

Обозначим через RSS_{ur} остаточную сумму квадратов в модели регрессии без ограничений (ur = unrestricted или «длинная регрессия»)

$$\hat{y} = \beta_1 x_1 + \cdots + \beta_m x_m,$$

а через RSS_r остаточную сумму квадратов в модели регрессии с ограничениями (r = restricted или «короткая регрессия»), налагаемыми проверяемой нулевой гипотезой, а именно в модели регрессии без учета факторов $x_{m-q+1}, \ldots, x_m$:

$$\hat{y} = \beta_1 x_1 + \cdots + \beta_{m-q} x_{m-q}.$$

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

$$\mathrm{RSS}_{ur} \leqslant \mathrm{RSS}_r \, .$$

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

$$F = \frac{\mathrm{RSS}_r - \mathrm{RSS}_{ur}}{\mathrm{RSS}_{ur}} \cdot \frac{n-m}{q}.$$

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 α. 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.

Аналогично проверяется статистическая гипотеза о линейных ограничениях на коэффициенты регрессии

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

при альтернативе

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

В этом случае q – число ограничений на коэффициенты модели.

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

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

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 β_j. Другими словами, маржинальные (предельные) величины в линейной модели регрессии постоянны и равны β_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 $\mathsf{M}\varepsilon_i = 0$, то

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

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

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

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.

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 $\mathsf{M}\varepsilon_i = 0$, то

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

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

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

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 β_j есть не что иное как коэффициент **эластичности** y по переменной x_j. В самом деле,

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

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 снижать. 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:

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

Подробнее о спецификации см. раздел «Спецификация модели» в Главе 3.

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

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u \tag{2.4}$$

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 (2.4) можно записать в матричном виде

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

Относительно ошибки u будем предполагать выполнения следующих условий:

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

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

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

При выполнении условия 1) очевидно

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

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

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

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

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

Доказательство. Аналогично случаю парной регрессии. □

Обозначим через Σ_x симметричную матрицу ковариаций объясняющих переменных размера $k \times k$:

$$\Sigma_x = \big(\mathrm{cov}(x_i,x_j)\big)_{i,j=1}^{k} = \begin{pmatrix} \mathrm{Var}(x_1) & \mathrm{cov}(x_1,x_2) & \cdots & \mathrm{cov}(x_1,x_k) \\ \mathrm{cov}(x_2,x_1) & \mathrm{Var}(x_2) & \cdots & \mathrm{cov}(x_2,x_k) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(x_k,x_1) & \mathrm{cov}(x_k,x_2) & \cdots & \mathrm{Var}(x_k) \end{pmatrix}$$

Замечание. Можно показать, что для произвольной корреляционой матрицы $\det \Sigma_x \geqslant 0$.

Предложение. *Для модели регрессии* (2.4) *при выполнении условий 1) и 2)*

$$\beta_0 = \mathsf{M}y - \beta_1 \mathsf{M}x - \dots - \beta_k \mathsf{M}x_k,$$

а коэффициенты $\beta_1,\dots,\beta_k$ *удовлетворяют системе уравнений*

$$\begin{cases} \mathrm{Var}(x_1)\beta_1 + \mathrm{cov}(x_2,x_1)\beta_2 + \dots + \mathrm{cov}(x_k,x_1)\beta_k = \mathrm{cov}(y,x_1) \\ \mathrm{cov}(x_1,x_2)\beta_1 + \mathrm{Var}(x_2)\beta_2 + \dots + \mathrm{cov}(x_k,x_2)\beta_k = \mathrm{cov}(y,x_2) \\ \dots \\ \mathrm{cov}(x_1,x_k)\beta_1 + \mathrm{cov}(x_2,x_k)\beta_2 + \dots + \mathrm{Var}(x_k)\beta_k = \mathrm{cov}(y,x_k) \end{cases}$$

Доказательство. Так как $\mathsf{M}u = 0$, то

$$\mathsf{M}y = \mathsf{M}(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + u) = \\ \beta_0 + \beta_1 \mathsf{M}x_1 + \dots + \beta_k \mathsf{M}x_k + \mathsf{M}u = \beta_0 + \beta_1 \mathsf{M}x_1 + \dots + \beta_k \mathsf{M}x_k$$

и получаем первую формулу. Далее, так как $\operatorname{cov}(x_j, u) = 0$ при $j = 1, \ldots, k$, то по свойству ковариации

$$\operatorname{cov}(y, x_j) = \operatorname{cov}(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u, x_j) = \\ \beta_1 \operatorname{cov}(x_1, x_j) + \beta_2 \operatorname{cov}(x_2, x_j) + \cdots + \beta_k \operatorname{cov}(x_k, x_j).$$

□

Замечание. Система линейных уравнений на коэффициенты $\beta_1, \ldots, \beta_k$ может быть записана в матричном виде

$$\Sigma_x \cdot \beta = \begin{pmatrix} \operatorname{cov}(y, x_1) \\ \operatorname{cov}(y, x_2) \\ \vdots \\ \operatorname{cov}(y, x_k) \end{pmatrix}.$$

Эта система имеет единственное решение тогда и только тогда, когда $\det \Sigma_x \neq 0$ (а значит $\det \Sigma_x > 0$). В этом случае решение системы имеет вид

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} = \Sigma_x^{-1} \cdot \begin{pmatrix} \operatorname{cov}(y, x_1) \\ \operatorname{cov}(y, x_2) \\ \vdots \\ \operatorname{cov}(y, x_k) \end{pmatrix}.$$

Замечание. В случае $\det \Sigma_x = 0$ говорят, что есть (чистая) мультиколлинеарность регрессоров. В этом случаем один из регрессоров линейно выражается через остальные и коэффициенты регрессии определены неоднозначно.

Рассмотрим задачу оценивания параметров $\beta_0, \ldots, \beta_k$ и σ^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:

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

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

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

Основной результат дается следующей теоремой.

Теорема (Гаусс – Марков)**.** *Пусть для линейной модели* (2.4) *выполнены условия 1) и 2),* $\det \Sigma_x \neq 0$ *и* $(y_i, \mathbf{x}_i)$ *– случайная выборка. Тогда OLS-оценки* $\hat{\beta}_0, \ldots, \hat{\beta}_k$ *коэффициентов регрессии* $\beta_0, \ldots, \beta_k$ *будут линейными несмещенными оценками с минимальной дисперсией*[3]*, т.е. BLUE оценками. Кроме того, эти оценки состоятельны, т.е.*

$$\hat{\beta}_j \underset{\mathsf{P}}{\longrightarrow} \beta_j, \quad j = 0 \ldots, k$$

при $n \longrightarrow +\infty$

Доказательство. **1.** Так же как и в случае детерминированных регрессоров показывается, что

$$\mathsf{M}\left(\hat{\beta}_j \middle| \mathbf{x}_1, \ldots, \mathbf{x}_k\right) = \beta_j, \quad j = 0, \ldots, k,$$

откуда

$$\mathsf{M}\left(\hat{\beta}_j\right) = \mathsf{M}\left(\mathsf{M}\left(\hat{\beta}_j \middle| \mathbf{x}_1, \ldots, \mathbf{x}_k\right)\right) = \beta_j.$$

Аналогично случаю детерминированных регрессоров показывается, что OLS–оценки коэффициентов имеют минимальную дисперсию среди всех линейных по y оценок.

2. Докажем состоятельность OLS-оценок коэффициентов регрессии. Так как при $n \to +\infty$ $(j = 1, \ldots, k,)$

$$\widehat{\operatorname{cov}}(y, x_j) \underset{\mathsf{P}}{\longrightarrow} \operatorname{cov}(y, x_j), \quad \widehat{\operatorname{Var}}(x_j) \underset{\mathsf{P}}{\longrightarrow} \operatorname{Var}(x_j),$$

то $\widehat{\Sigma}_x \underset{\mathsf{P}}{\longrightarrow} \Sigma_x$. Следовательно, по теореме Слуцкого

$$\widehat{\beta} = \widehat{\Sigma}_x^{-1} \cdot \begin{pmatrix} \widehat{\operatorname{cov}}(y, x_1) \\ \widehat{\operatorname{cov}}(y, x_2) \\ \vdots \\ \widehat{\operatorname{cov}}(y, x_k) \end{pmatrix} \underset{\mathsf{P}}{\longrightarrow} \Sigma_x^{-1} \begin{pmatrix} \operatorname{cov}(y, x_1) \\ \operatorname{cov}(y, x_2) \\ \vdots \\ \operatorname{cov}(y, x_k) \end{pmatrix} = \beta$$

Так как $\bar{y} \underset{\mathsf{P}}{\longrightarrow} \mathsf{M}y$ и $\bar{x}_j \underset{\mathsf{P}}{\longrightarrow} \mathsf{M}x_j$, то по теореме Слуцкого

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \cdots - \hat{\beta}_k \bar{x}_k \underset{\mathsf{P}}{\longrightarrow} \mathsf{M}y - \beta_1 \mathsf{M}x - \cdots - \beta_k \mathsf{M}x_k = \beta_0$$

□

[3]имеется ввиду условная дисперсия $\operatorname{Var}(\cdot|\mathbf{x}_1, \ldots, \mathbf{x}_n)$

Также 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: $t_j|\mathbf{x}_1 \ldots, \mathbf{x}_n \sim t_{n-m}$ и т.д.

2.8.1. Асимптотические свойства OLS-оценок

При доказательстве оптимальности OLS-оценок коэффициентов в модели со стохастическими регрессорами мы использовали только условия 1) – 2) на ошибки модели регрессии, а условие нормальной распределенности ошибок было нужно для доказательства статистических свойств OLS-оценок коэффициентов (проверки простых и сложных гипотез, построении доверительных интервалов и т.д.).

При больших объемах выборки условие нормальной распределенности ошибок регрессии можно ослабить. А именно, верна следующая теорема.

Теорема. *Пусть распределение ошибок регрессии имеет конечную дисперсию и выполнены условия 1) – 2) на ошибки регрессии. Тогда*

$$t = \frac{\hat{\beta}_j - \beta_j}{s_j} \longrightarrow \mathcal{N}(0,1) \quad (n \to +\infty).$$

1. Таким образом, асимптотически (при большом объеме выборки) доверительный интервал для коэффициента β_j с доверительной вероятностью γ имеет вид:

$$\mathsf{P}\left(\hat{\beta}_j - s_j \cdot z_\gamma < \beta_j < \hat{\beta}_j + s_j \cdot z_\gamma\right) \approx \gamma,$$

где z_γ есть решение уравнения

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

где $\Phi(\cdot)$ – функция стандартного нормального распределения.

2. Для проверки статистической гипотезы

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

против **двусторонней** альтернативы

$$H_1 \ : \ \beta_j \neq \theta_0$$

используется обычная t-статистика

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

При большом объеме выборки для проверки нулевой гипотезы получаем следующий приближенный статистический критерий:

при заданном уровне значимости α гипотеза H_0 отвергается при $|t| > z_{\text{кр}}$,

где $z_{\text{кр}}$ есть **двустороннее** критическое значение стандартного нормального распределения и находится как решение уравнения

$$\Phi(z_{\text{кр}}) = 1 - \frac{\alpha}{2}.$$

3. Для проверки статистической гипотезы

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

против **односторонней** альтернативы

$$H_1 \ : \ \beta_j > \theta_0$$

используется обычная t-статистика

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

При большом объеме выборки получаем следующий приближенный статистический критерий проверки нулевой гипотезы при односторонней альтернативе:

при заданном уровне значимости α гипотеза H_0 отвергается при

$$t > z'_{\text{кр}},$$

где $z'_{\text{кр}}$ есть **одностороннее** критическое значение стандартного нормального распределения и находится как решение уравнения

$$\Phi(z'_{\text{кр}}) = 1 - \alpha.$$

4. Для проверки сложной гипотезы о линейных ограничениях на коэффициенты модели регрессии

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

используется F-статистика (2.2) и следующий приближенный статистический критерий[4], называемый иногда тестом Вальда:

при заданном уровне значимости α гипотеза H_0 отвергается при

$$qF > \chi^2_{\text{кр}},$$

где q – число линейных ограничений на коэффициенты, а $\chi^2_{\text{кр}}$ есть критическое значение распределения χ^2_q. Отметим, что статистика теста Вальда

$$qF = \left(R\widehat{\beta}_{OLS} - \mathbf{r}\right)' \left[R \cdot \widehat{\text{Var}}\left(\widehat{\beta}_{OLS}\right) \cdot R'\right]^{-1} \left(R\widehat{\beta}_{OLS} - \mathbf{r}\right) = \frac{1}{s^2}\left(R\widehat{\beta}_{OLS} - \mathbf{r}\right)' \left[R \cdot (\mathbf{X}'\mathbf{X})^{-1} \cdot R'\right]^{-1} \left(R\widehat{\beta}_{OLS} - \mathbf{r}\right)$$

2.9. 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.

[4] так как для распределения Фишера $qF_{q,N} \approx \chi^2_q$ при $N >> 1$

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*

$$\operatorname{Var}\left(\widehat{\beta}_j\right) = \frac{\sigma^2}{(1 - R_j^2)\,\mathrm{TSS}_j} \quad (j = 1, \ldots, k), \tag{2.5}$$

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*

$$s_j^2 = \widehat{\operatorname{Var}}\left(\widehat{\beta}_j\right) = \frac{s^2}{(1 - R_j^2)\,\mathrm{TSS}_j} \quad (j = 1, \ldots, k)$$

Таким образом, «большие» значения R_j^2 (т.е. «близкие» к единице) приводят (при прочих равных) к «большой» выборочной стандартной ошибке коэффициента β_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 (2.5) видно, что к «большим» стандартным ошибкам коэффициента также приводят и «малые» значения 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

$$x_1 = \gamma_0 + \gamma_1 x_2 + \text{error},$$

а коэффициент R_2^2 – в регрессии

$$x_2 = \delta_0 + \delta_1 x_1 + \text{error}$$

Из соотношений для парной модели регрессии получаем

$$R_1^2 = R_2^2 = \widehat{\text{corr}}^2(x_1, x_2).$$

Следовательно, чем «сильнее» коррелируют регрессоры (т.е. чем «ближе» по абсолютной величине к единице коэффициент $\widehat{\text{corr}}(x_1, x_2)$), 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. [21], Гл. 4.9.1 и ссылки). В книге [29] 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.

Другой подход к мультиколлинеарности основан на состоятельности OLS-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. [29] Гл. 3.4 и ссылки)

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 исключения факторов надо подходить крайне осторожно, т.к. исключение существенного фактора приводит к смещению оценок коэффициентов (omitted variable problem, см. Главу 3).

4. Изменить спецификацию модели.

5. Оставить модель регрессии «как есть» (например, если экономическая теория говорит в пользу именно такой модели регрессии).

2.10. Задачи

Упражнение 1. Рассмотрим двухфакторную модель регрессии со стохастическими регрессорами

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

и будем предполагать, что выполнены все условия на ошибку u. Покажите, что

$$\beta_1 = \frac{\operatorname{cov}(y, x_1)\operatorname{Var}(x_2) - \operatorname{cov}(x_1, x_2)\operatorname{cov}(y, x_2)}{\operatorname{Var}(x_1)\operatorname{Var}(x_2) - \operatorname{cov}^2(x_1, x_2)},$$
$$\beta_2 = \frac{\operatorname{cov}(y, x_2)\operatorname{Var}(x_1) - \operatorname{cov}(x_1, x_2)\operatorname{cov}(y, x_1)}{\operatorname{Var}(x_1)\operatorname{Var}(x_2) - \operatorname{cov}^2(x_1, x_2)}.$$

Найдите формулы для OLS-оценок $\hat{\beta}_1$ и $\hat{\beta}_2$.

Упражнение 2. Рассмотрим двухфакторную модель регрессии со стохастическими регрессорами

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

и будем предполагать, что выполнены все условия на ошибку u. Покажите, что для дисперсии OLS-оценок коэффициентов регрессии

$$\operatorname{Var}\left(\widehat{\beta}_j\right) = \frac{\sigma^2}{(1 - r_{12}^2)\operatorname{TSS}_j} \quad (j = 1, 2)$$

где TSS_j есть общая вариация фактора x_j и $r_{12} = \widehat{\operatorname{corr}}(x_1, x_2)$

Упражнение 3. Рассмотрим двухфакторную модель регрессии со стохастическими регрессорами

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

и предположим, что $\mathsf{M}(u|x_1, x_2) = 0$. Покажите, что OLS-оценка $\hat{\beta}_1$ может быть найдена по следующей двухшаговой процедуре:

1. находим остатки x_1^* в регрессии $\hat{x}_1 = \gamma_0 + \gamma_2 x_2$,
2. находим OLS-оценку $\hat{d}_1$ в регрессии $\hat{y} = \delta_0 + \delta_1 x_1^*$.

Упражнение 4. Покажите, что для линейной модели регрессии при $k > 1$ верно неравенство $R^2_{adj} < R^2$.

Упражнение 5. Рассмотрим однофакторную и двухфакторную модели регрессии

$$(A) : \widehat{y} = \gamma_0 + \gamma_1 x$$
$$(B) : \widehat{y} = \beta_0 + \beta_1 x + \beta_2 z$$

Покажите, что

$$\bar{R}^2_B > \bar{R}^2_A \Longleftrightarrow |t_2| = \left|\frac{\widehat{\beta}_2}{s_2}\right| > 1$$

Упражнение 6. Покажите, что для модели регрессии

$$R^2 = \widehat{\mathrm{corr}}^2(y, \hat{y})$$

Упражнение 7. По 50 выборочным данным была оценена регрессионная модель зависимости рабочего времени $Hours$ (в часах) от уровня оплаты $Wage$ (в \$100):

$$\widehat{Hours} = \underset{(2.3)}{11.3} + \underset{(1.2)}{2.4} Wage - \underset{(0.24)}{1.1}\, Wage^2.$$

Тестируйте гипотезу о линейной зависимости рабочего времени от оплаты. Рассмотрите случаи уровня значимости 10%, 5%, 1%.

Упражнение 8. По 1000 выборочным данным была оценена регрессионная модель зависимости уровня оплаты труда $Wage$ от возраста Age:

$$\widehat{\ln Wage} = 1.2 + \underset{(0.0031)}{0.0214} Age - \underset{(0.0003)}{0.0004} Age^2.$$

Тестируйте гипотезу о линейной зависимости $\ln Wage$ от возраста. Рассмотрите случаи уровня значимости 10%, 5%, 1%.

Упражнение 9. Была оценена модель зависимости логарифма зарплаты w от уровня школьного образования edu и возраста age

$$\widehat{\ln w} = 1.2 + \underset{(0.01)}{0.02} edu + \underset{(0.016)}{0.01}\, age - \underset{(0.0002)}{0.0001} age^2 \quad n = 21.$$

- Дайте интерпретацию коэффициента β_1.
- Постройте доверительный интервал для коэффициента β_3 с доверительной вероятностью 90%.
- Тестируйте гипотезу о линейной зависимости логарифма зарплаты от возраста при уровне значимости 10%.

Упражнение 10. Рассмотрим модель регрессии зависимости оценки за финальный экзамен $exam$ от оценки за промежуточный срез $midterm$ и времени $time$, отведенного на выполнение задания (в мин)

$$\widehat{exam} = \underset{(0.3)}{0.98} midterm + \underset{(0.0006)}{0.001}\, time, \quad n = 29.$$

- Дайте интерпретацию коэффициента β_2.
- Найдите доверительный интервал для коэффициента β_1 с доверительной вероятностью 99%.
- Значим ли коэффициент β_1 при уровне значимости 1%?

Упражнение 11 ([29]). Рассмотрим регрессионную модель зависимости зарплаты CEO ($Salary$) от годового уровня продаж фирмы ($sales$), дохода на собственный капитал (roe, return on equity) и доходности акций (ros, return on stock)

$$\widehat{\ln Salary} = \underset{(0.12)}{5.21} + \underset{(0.016)}{0.27}\ln(sales) + \underset{(0.00023)}{0.0201}roe + \underset{(0.000094)}{0.00052}ros$$
$$n = 316 \quad R^2 = 0.132$$

- Дайте интерпретацию коэффициентам регрессии.
- Проверьте гипотезу
$$H_0 \ : \ \beta_3 = 0$$
против альтернативы
$$H_1 \ : \ \beta_3 > 0$$
уровень значимости 5%. Объясните выбор данной альтернативы.

Упражнение 12. Была оценена модель зависимости зарплаты $Wage$ от уровня школьного образования Edu и возраста Age (предполагаем, что нет проблемы эндогенности)

$$\widehat{\ln Wage} = 1.2 + \underset{(0.01)}{0.02}Edu + \underset{(0.016)}{0.01}\,Age - \underset{(0.0002)}{0.0001}Age^2$$
$$R^2 = 0.23 \quad n = 21.$$

- Дайте интерпретацию коэффициента β_1.
- Тестируйте гипотезу
$$H_0 : \beta_3 = 0$$
против альтернативы
$$H_1 : \beta_3 < 0$$
при уровне значимости 5%. Объясните выбор данной альтернативы.

Упражнение 13. По 32 выборочным данным была оценена модель регрессии (в скобках указаны стандартные ошибки коэффициентов):

$$\widehat{Y} = 0.4 + \underset{(0.06)}{3.2}\, X_1 + \underset{(0.1)}{2.5} X_2 + \underset{(0.001)}{0.1}\, X_3 - \underset{(0.1)}{0.3} X_4.$$

Проверьте значимость коэффициентов регрессии при уровне значимости 10% и сформулируйте проверяемые статистические гипотезы. Значимо ли коэффициент β_2 отличается от 3 (при уровне значимости 10%)? Сформулируйте проверяемую статистическую гипотезу.

Упражнение 14. По выборке объема 19 была оценена модель регрессии (в скобках указаны стандартные ошибки коэффициентов):

$$\widehat{Y} = 1.2 + \underset{(0.02)}{2.2}\, X_1 - \underset{(0.03)}{5.1}\, X_2 + \underset{(0.8)}{1.1} X_3$$
$$\mathrm{ESS} = 110.9 \quad \mathrm{RSS} = 40.5.$$

Чему равен коэффициент R^2? Проверьте значимость регрессии «в целом» (при уровне значимости 5%) и сформулируйте проверяемую статистическую гипотезу. Вычислите $\bar{R}^2$.

Упражнение 15. По выборке объема 25 была оценена модель регрессии (в скобках указаны стандартные ошибки коэффициентов):

$$\widehat{Y} = 10.4 + \underset{(0.6)}{3.2} X_1 + \underset{(0.1)}{2.5} X_2 - \underset{(0.1)}{0.6} X_3 \quad R^2 = 0.55$$

Постройте доверительные интервалы для коэффициентов регрессии с доверительной вероятностью 99%. Какие из факторов значимо, а какие незначимо влияют на Y (при уровне значимости 1%)? Сформулируйте проверяемые статистические гипотезы.

Упражнение 16. По 19 выборочным данным была оценена модель регрессии (в скобках указаны стандартные ошибки коэффициентов):

$$\widehat{Y} = 1.2 + \underset{(0.1)}{3.2} X_1 - \underset{(0.03)}{0.9}\, X_2.$$

Проверьте значимость коэффициентов регрессии при уровне значимости 0.1% и сформулируйте проверяемые статистические гипотезы. Значимо ли коэффициент β_2 отличается от (-1) (при уровне значимости 0.2%)? Сформулируйте проверяемую статистическую гипотезу.

Упражнение 17. На основе опроса 25 человек была оценена модель зависимости логарифма зарплаты ($\ln w$) от уровня образования (edu, в годах) и возраста (age):

$$\widehat{\ln w} = 1.7 + 0.5edu + 0.06age - 0.0004age^2$$
$$\text{ESS} = 90.3 \quad \text{RSS} = 60.4.$$

Когда в модель были введены переменные $fedu$ и $medu$, учитывающие уровень образования родителей, величина ESS увеличилась до 110.3. Напишите спецификацию уравнения регрессии с учетом уровня образования родителей. Сформулируйте и проверьте гипотезу о значимом влиянии уровня образования родителей на зарплату (уровень значимости 5%).

Упражнение 18. На основе опроса 50 человек была оценена модель зависимости логарифма заработной платы ($\ln w$) от уровня образования (edu, в годах), возраста (age):

$$\widehat{\ln w} = 1.7 + 0.5edu + 0.04age - 0.0003age^2$$
$$\text{ESS} = 75.8 \quad \text{RSS} = 61.3.$$

Когда в модель были введены переменные $fedu$ и $medu$, учитывающие уровень образования родителей, величина ESS увеличилась до 82.7. Напишите спецификацию уравнения регрессии с учетом уровня образования родителей. Вычислите скорректированный коэффициент R^2 для обеих моделей регрессии. Улучшило ли включение в модель новых факторов «качество» модели?

Упражнение 19. Была оценена функция Кобба–Дугласа с учетом человеческого капитала H (K – физический капитал, L – труд)

$$\widehat{\ln(Q)} = 1.4 + 0.46\ln(L) + 0.27\ln(H) + 0.23\ln(K)$$
$$\text{ESS} = 170.4 \quad \text{RSS} = 80.3 \quad n = 21.$$

Чему равен коэффициент R^2? Проверьте значимость регрессии «в целом» и сформулируйте проверяемую статистическую гипотезу. Уровень значимости 1%.

Упражнение 20. Случайным образом было выбрано 300 домов. Пусть $Price$ – цена дома (в \$1000), BDR – количество спальных комнат, $Bath$ – число ванных комнат, $Hsize$ – площадь дома, $Lsize$ – площадь

участка вокруг дома, Age – возраст дома (в годах), $Poor$ – бинарная переменная, равная 1, если состояние дома оценивается как «плохое». Была оценена регрессия

$$\widehat{Price} = 134.3 + 0.369BDR + 22.7Bath + 0.134Hsize$$
$$+ 0.00036Lsize + 0.016Age - 23.4Poor, \quad \overline{R}^2 = 0.63, \quad \text{SER} = 6.5$$

- Пусть владелец дома решил переделать одну из жилых комнат (не спальню) в новую ванную комнату. Какое ожидаемое изменение цены дома?
- Пусть владелец дома решил добавить еще одну ванную комнату, увеличив площадь дома на 100м^2. Какое ожидаемое изменение цены дома?
- Найдите коэффициент R^2.

Упражнение 21. В условиях предыдущей задачи были также получены стандартные ошибки коэффициентов

$$\widehat{Price} = \underset{(23.9)}{134.3} + \underset{(2.13)}{0.369}BDR + \underset{(8.94)}{22.7}Bath + \underset{(0.014)}{0.134}Hsize$$
$$+ \underset{(0.000048)}{0.00036}Lsize + \underset{(0.311)}{0.016}Age - \underset{(9.5)}{23.4}Poor, \quad \overline{R}^2 = 0.63, \ \text{SER} = 6.5$$

- Значимо ли фактор BDR влияет на цену дома? Рассмотрите уровень значимости 1%, 2%, 5%, 10%.
- Обычно дома с шестью спальнями стоят гораздо дороже, чем дома с двумя спальнями. Как это соотносится с ответом на предыдущий пункт? Дайте несколько возможных объяснений.
- Владелец дома увеличил площадь участка на 2000 м^2. Постройте 99% доверительный интервал для изменения стоимости дома.
- F-статистика для проверки влияния BDR и Age равна $F = 0.782$. Значимо ли факторы BDR и Age влияют на стоимость дома? Рассмотрите уровни значимости 1%, 5%, 10%.

Упражнение 22. Оценивается модель зависимости школьной оценки по математике ($math$, по 100-бальной системе) от возраста матери при

рождении ребенка ($agebirth$), способностей матери ($mothabil$), уровня школьного образования матери ($school$), дохода матери ($income$, в \$1000) и бинарной переменной $married$, равной 1, если мать ребенка замужем и 0 иначе. Результаты оценивания:

$$\widehat{math} = \underset{(3.439)}{30.741} + \underset{(0.108)}{0.068}agebirth + \underset{(0.571)}{6.775}mothabil + \underset{(0.256)}{1.607}school + \underset{(1.081)}{4.109}married - \underset{(0.005)}{0.005}income \quad R^2 = 0.151, \;\; n = 2524$$

- Значимо ли школьные оценки зависят от семейного положения матери? Рассмотрите уровень значимости 1%, 5%, 10%.
- Значимо ли возраст матери при рождении ребенка влияет на оценку ребенка по математике? Рассмотрите уровень значимости 1%, 5%, 10%.
- Постройте доверительный интервал для эффекта влияния уровня доходов матери на школьную оценку. Рассмотрите доверительные вероятности 90%, 95%, 99%.
- Постройте доверительный интервал для эффекта влияния уровня образования матери на школьную оценку по математике. Рассмотрите доверительные вероятности 90%, 95%, 99%.

Упражнение 23. В условиях предыдущей задачи также была оценена модель

$$\widehat{math} = \underset{(3.299)}{33.019} + \underset{(0.106)}{0.072}agebirth + \underset{(0.569)}{6.936}mothabil + \underset{(0.253)}{1.646}school \quad R^2 = 0.1461 \;\; n = 2524$$

Значимо ли доход матери и ее семейное положение (совместно) влияют на школьную оценку по математике? Сформулируйте проверяемую гипотезу. Рассмотрите уровни значимости 1%, 5%, 10%.

Упражнение 24. Рассмотрим регрессионную модель зависимость логарифма зарплаты $\ln(Wage)$ от уровня образования edu, опыта работы $experience$, $experience^2$ и уровня образования родителей $fedu$ и $medu$

$$\widehat{\ln(Wage)} = \beta_0 + \beta_1 edu + \beta_2 experience + \beta_3 experience^2 + \beta_4 fedu + \beta_5 medu, \quad n = 500 \quad R^2 = 0.273$$

a) Напишите спецификацию модели регрессии с ограничениями для проверки статистической гипотезы $H_0 : \beta_4 = \beta_5$.

b) Дайте интерпретацию нулевой гипотезе из п. a).

c) Для модели регрессии из п. a) был вычислен $R_r^2 = 0.262$. Тестируйте нулевую гипотезу при уровне значимости 5%.

d) Вычислите скорректированный коэффициент R^2 для исходной модели.

Упражнение 25. Была оценена производственная функция Кобба – Дугласа (Cobb–Douglas production function)

$$\widehat{\ln(Q)} = \underset{(0.08)}{2.4} + \underset{(0.21)}{0.35} * \ln(K) + \underset{(0.42)}{0.71} * \ln(L) \quad R^2 = 0.89 \quad n = 20$$

a) Проверьте значимость коэффициентов β_1 и β_2 (уровень значимости 5%).

b) Проверьте значимость регрессии «в целом» (уровень значимости 5%).

c) Также был вычислен $\widehat{\text{corr}}(\ln(K), \ln(L)) = 0.96$. Как это соотносится с ответами на пп. a) и b)?

d) Напишите спецификацию модели регрессии с ограничениями для тестирования гипотезы $H_0 : \beta_1 + \beta_2 = 1$.

e) Дайте интерпретацию нулевой гипотезе.

f) Пусть в модели с ограничениями $R_r^2 = 0.79$. Тестируйте нулевую гипотезу при уровне значимости 1%.

Упражнение 26. Была оценена функция Кобба–Дугласа с учетом человеческого капитала H (K – физический капитал, L – труд)

$$\widehat{\ln(Q)} = 1.4 + \underset{(0.31)}{0.46}\ln(L) + \underset{(0.12)}{0.27}\ln(H) + \underset{(0.14)}{0.23}\ln(K)$$
$$R^2 = 0.86 \quad n = 20$$

a) Проверьте значимость коэффициентов регрессии (уровень значимости 1%);

b) проверьте значимость регрессии «в целом» (уровень значимости 1%);

c) Напишите спецификацию модели регрессии с ограничениями для тестирования гипотезы $H_0 : \beta_1 + \beta_2 + \beta_3 = 1$.

d) Дайте интерпретацию нулевой гипотезе.

e) Пусть в модели с ограничениями $R_r^2 = 0.72$. Тестируйте нулевую гипотезу при уровне значимости 1%.

Упражнение 27. Рассмотрим модель зависимости цены дома $Price$ (в \$1000) от его площади $Hsize$ (в м^2), площади участка $Lsize$ (в м^2), числа ванных комнат $Bath$ и числа спален BDR

$$\widehat{Price} = \beta_0 + \beta_1 Hsize + \beta_2 Lsize + \beta_3 Bath + \beta_4 BDR$$
$$R^2 = 0.218 \quad n = 23$$

Напишите спецификацию регрессии с ограничениями для проверки статистической гипотезы $H_0 : \beta_3 = 20\beta_4$. Дайте интерпретацию проверяемой гипотезе. Для регрессии с ограничениями был вычислен $R_r^2 = 0.136$. Тестируйте нулевую гипотезу при уровне значимости 5%.

Упражнение 28. Рассмотрим модель зависимости почасовой оплаты труда w от уровня образования $educ$, возраста age и уровня образования родителей $fathedu$, $mothedu$

$$\widehat{\ln w} = \beta_0 + \beta_1 educ + \beta_2 age + \beta_3 age^2 + \beta_4 fathedu + \beta_5 mothedu$$
$$n = 27 \quad R^2 = 0.341$$

Напишите спецификацию регрессии с ограничениями для проверки статистической гипотезы $H_0 : \beta_5 = 2\beta_4$. Дайте интерпретацию проверяемой гипотезе. Для регрессии с ограничениями был вычислен $R_r^2 = 0.296$. Тестируйте нулевую гипотезу при уровне значимости 5%.

Упражнение 29. Рассмотрим модель зависимости цены дома P (в \$1000) от его площади $Hsize$ (в м^2), площади участка $Lsize$ (в м^2), числа ванных комнат $Bath$ и числа спален BDR

$$\widehat{P} = \beta_0 + \beta_1 Hsize + \beta_2 Lsize + \beta_3 Bath + \beta_4 BDR \quad R^2 = 0.218 \quad n = 23$$

Напишите спецификацию регрессии с ограничениями для проверки статистической гипотезы $H_0 : \beta_2 = 0, \beta_3 = \beta_4$. Дайте интерпретацию

проверяемой гипотезе. Для регрессии с ограничениями был вычислен $R_r^2 = 0.136$. Тестируйте нулевую гипотезу при уровне значимости 5%.

Упражнение 30. Рассмотрим модель зависимости накоплений домашних хозяйств S (в тыс.) от доходов Inc (в тыс.) и имущества W (в тыс.)

$$\widehat{S} = \beta_0 + \beta_1 Inc + \beta_2 W \quad R^2 = 0.121 \quad n = 26$$

Напишите спецификацию регрессии с ограничениями для проверки статистической гипотезы $H_0 : \beta_1 = \beta_2$. Дайте интерпретацию проверяемой гипотезе. Для регрессии с ограничениями был вычислен $R_r^2 = 0.062$. Тестируйте нулевую гипотезу при уровне значимости 5%.

Упражнение 31. Рассмотрим регрессионную модель зависимость логарифма зарплаты $\ln(Wage)$ от уровня образования edu, опыта работы $experience$, $experience^2$ и уровня образования родителей $fedu$ и $medu$

$$\widehat{\ln(Wage)} = \beta_0 + \beta_1 edu + \beta_2 experience + \beta_3 experience^2 + \beta_4 fedu + \beta_5 medu$$

Модель регрессии была отдельно оценена по выборкам из 35 мужчин и из 23 женщин и были получены остаточные суммы квадратов $\text{RSS}_\text{м} = 34.4$ и $\text{RSS}_\text{ж} = 23.4$. Остаточная сумма квадратов в регрессии, оцененной по объединенной выборке, равна 70.3. Тестируйте гипотезу об отсутствии дискриминации в оплате труда между мужчинами и женщинами. Уровень значимости 5%.

Упражнение 32. Рассмотрим двухфакторную модель регрессии

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u,$$

ошибки которой удовлетворяют условиям Гаусса – Маркова. Выразите F-статистику для проверки нулевой гипотезы

- $H_0 : \beta_1 = \beta_2 = 0$,
- $H_0 : \beta_1 = \beta_2$,
- $H_0 : \beta_1 + \beta_2 = 1$.

через $\widehat{\text{Var}}(\hat{\beta}_1)$, $\widehat{\text{Var}}(\hat{\beta}_2)$ и $\widehat{\text{cov}}(\hat{\beta}_1, \hat{\beta}_2)$.

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 β_3 представляет собой (среднюю) разницу в оплате труде между мужчинами и женщинами при прочих равных условиях. Проверка значимости коэффициента β_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 β_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 β_0 есть (средний) уровень оплаты человека без высшего образования, а коэффициент β_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 β_2. Однако наличие высшего образование дает **одинаковое** (среднее) увеличение уровня почасовой оплаты как для мужчин, так и для женщин на величину β_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 β_2. При наличии высшего образования уровень оплаты женщины увеличивается (в среднем) на величину β_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 β_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 β_2. Аналогичный смысл имеют коэффициенты β_3 и β_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.

Также 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 β_2 и β_3 показывают изменения, соответственно, константы и коэффициента наклона прямой после наступления события, влияющего на структурные изменения количественной зависимости между факторами.

3.2. 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.

3.2.1. 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 β_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 ($\mathrm{cov}(x, u) = 0$)

$$\begin{aligned}\mathrm{cov}(y, x) &= \mathrm{cov}(\beta_0 + \beta_1 x + \beta_2 z + u, x) = \\ &\beta_1 \,\mathrm{cov}(x, x) + \beta_2 \,\mathrm{cov}(z, x) + \mathrm{cov}(u, x) = \beta_1 \,\mathrm{Var}(x) + \beta_2 \,\mathrm{cov}(z, x).\end{aligned}$$

Следовательно,

$$\hat{\beta}_1 \xrightarrow[\mathsf{P}]{} \beta_1 + \beta_2 \frac{\mathrm{cov}(z, x)}{\mathrm{Var}(x)}.$$

Таким образом, если $\beta_2 \,\mathrm{cov}(z, x) \neq 0$, то OLS-оценка коэффициента β_1 в «недооцененной» модели **несостоятельная** и асимптотически **смещена**. Можно также показать, что выборочная стандартная ошибка коэффициента β_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 β_0 в «укороченной» регрессии.

Теорема (Omitted variable problem)**.** *Пусть модель зависимости y от факторов x и z описывается уравнением*

$$y = \beta_0 + \beta_1 x + \beta_2 z + u, \tag{3.1}$$

где ошибка u удовлетворяет всем условиям теоремы Гаусса-Маркова. Пусть $\hat{\beta}_0$ и $\hat{\beta}_1$ – OLS-оценки коэффициентов β_0 и β_1, полученные при рассмотрении «укороченой» модели

$$y = \beta_0 + \beta_1 x + w. \tag{3.2}$$

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 β_0 и β_1, вычисленные в рамках модели* (3.2) ***меньше*** *выборочных стандартных ошибок тех же коэффициентов, но вычисленных в рамках «истинной» модели* (3.1)*.*

3. *$\mathsf{M}s^2 > \sigma^2$, т.е. s^2 – смещенная оценка дисперсии ошибок регрессии* (3.1)*.*

Следствие. *При невключении в модель регрессии влияющей переменной, т.е. при оценивание «укороченной» модели* (3.2)*, 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. [3, 21]).

3.2.2. 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

$$y = \gamma_0 + \gamma_1 x + \gamma_2 z + w, \tag{3.3}$$

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 (3.3) 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 (3.3). 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 (3.3), т.е. оценка $\hat{\gamma}_1$ не будет наилучшей. Как следствие, доверительный интервал для коэффициента β_1 в «истинной» модели будет у́же, чем доверительный интервал для коэффициента $\gamma_1 = \beta_1$ в модели (3.3) 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.*

[1] так как рассматриваем модель со стохастическими регрессорами

Замечание. При включении в модель незначимых объясняющих переменных OLS-оценки становятся «менее точными», т.к. возрастает выборочная стандартная ошибка коэффициентов. Это соответствует общему принципу математической статистики: чем больше коэффициентов в вероятностной модели нужно оценить, тем «хуже» (при одинаковом объеме выборки) статистические свойства выборочных коэффициентов.

3.2.3. 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 (подробнее см. Приложение 2).

3.2.4. 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 [7, 8, 20] 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 θ_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 θ_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).

3.2.5. 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.

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, Ramsey [24])

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 $M = 3, 4$ [29].

Замечание. RESET-тест может отвергать нулевую гипотезу о линейной спецификации в случае невключения в линейную модель значимого регрессора.

y **или** $\ln y$ **?**

Основная модели регрессии – линейная – характеризуется линейной зависимостью $\mathsf{M}y_i$ от значений объясняющих переменных и постоянством маржинальных значений. Коэффициент β_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 другу 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 θ_{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 θ_{Log}. Если коэффициент незначим, то данные согласованы с моделью $(LogLin)$.

3.3. Гетероскедастичность ошибок регрессии. Взвешенный метод наименьших квадратов

Определение. *Гетероскедастичностью*[2] (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 будут несмещенными и состоятельными (в случаем стохастических регрессоров), но* ***не наилучшими*** *(не BLUE оценки).*

[2]Иногда употребляют термин *неоднородность ошибок регрессии*

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.

3.3.1. 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 [17] Это тест применяется в случае, если есть априорные предположения о зависимости дисперсии ошибок от значения одной из объясняющих переменных[3] (например, фактора x_j) вида $\sigma_i^2 = \sigma^2 x_{ij}^2$ с неизвестным параметром σ. 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 RSS_1 и RSS_2 соответствующие остаточные суммы квадратов, а через SER_1 и 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 α гипотезу H_0 следует отвергается при $F > F_{\text{кр}}$ (т.е. при больших значения статистики), где критическое значение $F_{\text{кр}} = F(\alpha; n_1 - m, n_2 - m)$.

[3] В тесте неявно предполагается, что значения этой объясняющей переменной должны быть **положительными**.

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 [17].

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[4] (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

$$F = \frac{\mathrm{SER}_2^2}{\mathrm{SER}_1^2} = \frac{\mathrm{RSS}_2 / (n_2 - m)}{\mathrm{RSS}_1 / (n_1 - m)}.$$

Тест White [28] 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;

[4] Легко понять, что в этой ситуации нет необходимости в отбрасывании значений для «более надежного» разделения на группы

2. Вычисляем коэффициент R_0^2 во вспомогательной регрессии e_i^2 на константу, регрессоры, их квадраты и попарные произведения:

$$e^2 = \delta_0 + \sum_{i=1}^{k} \delta_i x_i + \sum_{i \leqslant j}^{k} \delta_{ij} x_i x_j + \text{error} = \\ \delta_0 + \delta_1 x_1 + \cdots + \delta_{11} x_1^2 + \cdots + \delta_{12} x_1 x_2 + \cdots + \text{error} \quad (3.4)$$

При справедливости нулевой гипотезы о гомоскедастичности ошибок статистика nR_0^2 для проверки значимости вспомогательной регрессии (3.4) «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 (3.4). Таким образом, при заданном уровне значимости α (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 (3.4) не имеет никакой экономической интерпретации, а нужна **только для вычисления** коэффициента R_0^2.

Тест White является наиболее общим тесто 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 (3.4) «в целом» также можно использовать F-статистику, асимптотически имеющую распределение Фишера.

Если во вспомогательной модели (3.4) «много» (относительно объема выборки) коэффициентов, то можно воспользоваться следующей модификацией теста White [29]:

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 $\hat{y}_i^2$:

$$e_i^2 = \delta_0 + \delta_1 \hat{y}_i + \delta_2 \hat{y}_i^2 + \text{error} \tag{3.5}$$

   и находим коэффициента R_0^2.

При справедливости нулевой гипотезы о гомоскедастичности ошибок статистика nR_0^2 для проверки значимости вспомогательной регрессии (3.5) «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 α (асимптотически) нулевая гипотеза о гомоскедастичности отвергается при $nR_0^2 > \chi^2_{\text{кр}}$, где критическое значение $\chi^2_{\text{кр}} = \chi^2(\alpha; 2)$.

Тест Breusch – Pagan [6] Это тест применяется, если есть априорные предположения, что дисперсия ошибок регрессии зависит от факторов $\mathbf{z}' = (z_1, \ldots, z_p)$

$$\sigma_i^2 = f(\gamma_0 + \mathbf{z}_i'\gamma) = f(\gamma_0 + z_{i1}\gamma_1 + \cdots + z_{ip}\gamma_p) \tag{3.6}$$

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 (3.6) 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 (3.6)) LM-стастистика[5]

$$LM = \text{ESS}\,/2$$

асимптотически имеет распределение хи-квадрат

$$LM \underset{H_0}{\approx} \chi_p^2.$$

Следовательно, для больших выборок при заданном уровне значимости α 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

$$nR_0^2 \underset{H_0}{\approx} \chi_p^2.$$

[5] LM = Lagrange Multiplier, статистика множителей Лагранжа

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.

3.3.2. Корректировка на гетероскедастичность

Как отмечалось выше, при гетероскедастичности ошибок регрессии OLS-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

- $\mathsf{M}\varepsilon_i = 0$;
- $\operatorname{cov}(\varepsilon_i, \varepsilon_j) = 0$ при $i \neq j$;
- $\operatorname{Var}(\varepsilon_i) = \sigma_i^2$ и дисперсии σ_i^2 **известны**.

Таким образом, нам известна **полная информация** о структуре неоднородности (гетероскедастичности) ошибок регрессии.

Разделим уравнение регрессии на σ_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 ε_i' в преобразованном уравнении регрессии уже **будут гомоскедастичны**, так как

$$\mathrm{Var}(\varepsilon_i') = \mathrm{Var}\left(\frac{\varepsilon_i}{\sigma_i}\right) = \frac{1}{\sigma_i^2}\,\mathrm{Var}(\varepsilon_i) = 1.$$

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:

$$S = \sum_{i=1}^{n}\left(\frac{y_i}{\sigma_i} - \beta_0\frac{1}{\sigma_i} - \beta_1\frac{x_{i1}}{\sigma_i} - \cdots - \beta_k\frac{x_{ik}}{\sigma_i}\right)^2 = \sum_{i=1}^{n}\frac{1}{\sigma_i^2}\left(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik}\right)^2 \longrightarrow \min$$

Определение. Пусть заданы положительные числа $\{w_i\}_{i=1}^n$ (веса). Оценки коэффициентов модели регрессии, полученные минимизацией *взвешенной суммы квадратов отклонений*

$$\sum_{i=1}^{n} w_i\left(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik}\right)^2 \longrightarrow \min,$$

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 линейными оценками коэффициентов (BLUE–оценками) 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 σ_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 точность до множителя

Предположим, что нам известна функциональная форма зависимости дисперсии ошибок от объясняющих переменных **с точностью до неизвестного множителя** σ^2:

$$\mathrm{Var}(\varepsilon_i) = \sigma^2 h(x_1, \ldots, x_k), \tag{3.7}$$

где $h(\mathbf{x}) = h(x_1, \ldots, x_k) > 0$ – **известная функция**. По аналогии с предыдущим разделом преобразуем уравнение регрессии разделив его на $\sqrt{h(\mathbf{x}_i)}$:

$$\frac{y_i}{\sqrt{h(\mathbf{x}_i)}} = \beta_0 \frac{1}{\sqrt{h(\mathbf{x}_i)}} + \beta_1 \frac{x_{i1}}{\sqrt{h(\mathbf{x}_i)}} + \cdots + \beta_k \frac{x_{ik}}{\sqrt{h(\mathbf{x}_i)}} + \varepsilon_i' \quad \left(\varepsilon_i' = \frac{\varepsilon_i}{\sqrt{h(\mathbf{x}_i)}} \right).$$

Тогда ошибки ε_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* (3.7) *в линейной модели регрессии наилучшими линейными оценками коэффициетов (BLUE – оценками) 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

$$\mathrm{Var}(u|x_1, \ldots, x_k) = \sigma^2 h(\mathbf{x}) = \sigma^2 \exp(\delta_1 x_1 + \cdots + \delta_k x_k) \tag{3.8}$$

где параметры $\delta_1, \ldots, \delta_k$ и σ^2 **неизвестны**. Прологарифмировав равенство (3.8) 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

$$\ln\left(e^2\right) = \delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k + \text{error}$$

   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 (3.8) 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 [28] 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 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 α 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 χ^2_q. Следовательно, нулевая гипотеза асимптотически отвергается при $\chi^2 > \chi^2_{\text{кр}}$, где критическое значение $\chi^2_{\text{кр}} = \chi^2(\alpha; q)$.

3.4. 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 моделей регрессии 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.

3.4.1. Автокорреляция первого порядка

Одной из наиболее часто используемых моделей автокорреляции является модель автокорреляции первого порядка

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t, \quad |\rho| < 1 \tag{3.9}$$

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 ρ есть не что иное как коэффициента корреляции между соседними ошибками в модели регрессии.

В зависимости от знака параметра ρ 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 (3.9) 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 (подробнее о моделях авторегрессии см. Главу 4)

Тесты на автокорреляции

Опишем несколько наиболее часто используемых тестов на выявление автокорреляции первого порядка. Во всех тестах в рамках модели автокорреляции (3.9) проверяется гипотеза

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

против альтернативы

$$H_1 \,:\, \rho \neq 0.$$

Тест Durbin – Watson [11, 12] 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

$$\mathrm{DW} \underset{\mathsf{P}}{\longrightarrow} 2(1-\rho).$$

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 $d_l < \text{DW} < d_u$ или $4-d_u < \text{DW} < 4-d_l$ **неопределенность**, нельзя сделать выбор ни в пользу H_0, ни в пользу H_1.

Особенностью тест 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** β_0;
2. регрессоры **не коррелируют** с ошибками регрессии (strictly exogenous), в частности среди объясняющих факторов не должно быть лаговых значений зависимой переменной.

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

$$e_t = \gamma_0 + \gamma_1 e_{t-1} + \sum_{j=1}^{n} \delta_j x_j + \text{error} \tag{3.10}$$

m-тест Durbin основан на t-тесте проверки значимости коэффициента при e_{t-1} во вспомогательной регрессии (3.10): если коэффициент значим, то гипотеза H_0 об отсутствии автокорреляции ошибок регрессии отвергается.

Тест множителей Лагранжа (LM-test) основан на статистике

$$LM = (n-1)R_0^2,$$

где коэффициент R_0^2 вычисляется во вспомогательной регрессии (3.10). 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 χ^2_1.

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** (**гомоскедастичности**) 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 (3.9) рассмотрим на примере однофакторной модели регрессии

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t.$$

Значение ρ известно Умножим уравнение для наблюдения $t-1$ на ρ, вычтем из уравнения для наблюдения t и с учетом (3.9) 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 β_0 и β_1 в преобразованном уравнении будут BLUE-оценками и для них верны стандартные статистические выводы модели без константы.

Описанная процедура преобразования регрессии 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 ρ неизвестно Однако во многих прикладных задачах значение параметра ρ, 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 ρ 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.

3.4.2. Автокорреляция произвольного порядка

Рассмотрим теперь модель автокорреляции ошибок регрессии произвольного порядка p:

$$\varepsilon_t = \rho_1 \varepsilon_{t-1} + \cdots + \rho_p \varepsilon_{t-p} + w_t, \tag{3.11}$$

и предполагается, что u_t удовлетворяют условиям Гаусса – Маркова. Модель (3.11) зависимости ошибок регрессии есть модель $AR(p)$ авторегрессии порядка p и для получения «хороших» статистических свойств оценок регрессии на модель (3.11) необходимо накладывать условия, гарантирующие стационарности временного ряда ε_t (см. Глава 4).

Как и в случае первого порядка рассмотрим два аспекта: тесты на обнаружение автокорреляции и корректировка модели на автокорреляцию.

Тесты на автокорреляцию произвольного порядка

В рамках модели автокорреляции ошибок регрессии (3.11) (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

$$e_t = \gamma_0 + \gamma_1 e_{t-1} + \cdots + \gamma_p e_{t-p} + \sum_{j=1}^{k} \delta_j x_j + \text{error} \qquad (3.12)$$

F-тест основан на обычной проверки значимости совместного влияния лаговых переменных на e_t в регрессии (3.12), 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 (3.12). 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 **постоянства дисперсий** (**гомоскедастичности**) ошибок регрессии. Существуют модификации этих тестов, устойчивые к гетероскедастичности ошибок регрессии.

Неправильная спецификация Как и в модели первого порядка, иногда выводы об автокорреляция ошибок регрессии могут быть связаны с неправильной спецификацией модели.

Корректировка модели на автокорреляцию

Рассмотрим корректировку модели на примере модели автокорреляции второго порядка в однофакторной модели регрессии

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$
$$\varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + u_t$$

и предположим, что значения параметров ρ_1 и ρ_2 известны. Условия стационарности модели автокорреляции ошибок регрессии в данном случае принимают вид

$$\rho_2 > -1, \quad \rho_2 - \rho_1 < 1, \quad \rho_1 + \rho_2 < 1.$$

По аналогии с автокорреляцией первого порядка для $t > 2$ получаем преобразованное уравнение

$$y_t - \rho_1 y_{t-1} - \rho_2 y_{t-2} = \beta_0(1 - \rho_1 - \rho_2) + \beta_1(x_t - \rho_1 x_{t-1} - \rho_2 x_{t-2}) + u_t \quad t = 3, \ldots, n \quad (3.13)$$

Для учета первого и второго наблюдения введем новые переменные

$$\tilde{y}_t = \begin{cases} y_t - \rho_1 y_{t-1} - \rho_2 y_{t-2}, & t = 3, \ldots n \\ \sqrt{1-\rho_2^2} y_2 - \left(\rho_1 \sqrt{1-\rho_1^2}/(1-\rho_2)\right) y_1, & t = 2 \\ \sqrt{(1+\rho_2)[(1-\rho_2)^2 - \rho_1^2]/(1-\rho_2)}\, y_1, & t = 1 \end{cases}$$

$$\tilde{x}_t = \begin{cases} x_t - \rho_1 x_{t-1} - \rho_2 x_{t-2}, & t = 3, \ldots n \\ \sqrt{1-\rho_2^2} x_2 - \left(\rho_1 \sqrt{1-\rho_1^2}/(1-\rho_2)\right) x_1, & t = 2 \\ \sqrt{(1+\rho_2)[(1-\rho_2)^2 - \rho_1^2]/(1-\rho_2)}\, x_1, & t = 1 \end{cases}$$

$$\tilde{z}_t = \begin{cases} 1 - \rho_1 - \rho_2, & t = 3, \ldots n \\ \sqrt{1-\rho_2^2} - \left(\rho_1 \sqrt{1-\rho_1^2}/(1-\rho_2)\right), & t = 2 \\ \sqrt{(1+\rho_2)[(1-\rho_2)^2 - \rho_1^2]/(1-\rho_2)}, & t = 1 \end{cases}$$

Тогда в новых факторах модель регрессии, скорректированная на автокорреляцию второго порядка, принимает вид

$$\tilde{y}_t = \beta_0 \tilde{z}_t + \beta_1 \tilde{x}_t + u_t \quad t = 1, \ldots, n \quad (3.14)$$

Ошибки в преобразованном уравнении (3.14) буду удовлетворять всем условиям теоремы Гаусса –Маркова. Следовательно, OLS-оценки коэффициентов β_0 и β_1 в преобразованном уравнении (3.14) будут BLUE-оценками и для них верны стандартные статистические выводы модели без константы.

Замечание. Так как предпочтительней работать с моделью регрессии с включенной константой, то при большом объеме выборки предпочтительней использовать преобразованную модель (3.13)

Если значения коэффициентов ρ_1 и ρ_2 неизвестны, то вначале они оцениваются в рамках вспомогательной модели (e_t есть OLS-остатки в исходной модели)

$$e_t = \rho_1 e_{t-1} + \rho_2 e_{t-2} + \text{error},$$

а затем проводится преобразование (3.13) или (3.14). Описанная двухшаговая процедура представляет собой доступный метод наименьших квадратов в условиях автокорреляции второго порядка.

Замечание. Аналогично проводится корректировка модели на автокорреляцию произвольного порядка. Многие современные эконометрические пакеты позволяют проводить корректировку модели на автокорреляцию произвольного порядка.

Другие модели зависимости ошибок регрессии

Можно рассматривать другие модели зависимости ошибок регрессии, например модель скользящего среднего MA(Q) порядка Q

$$\varepsilon_t = u_t + \theta_1 u_{t-1} + \cdots + \theta_q u_{t-Q},$$

где $\{u_t\}$ удовлетворяют условиям Гаусса – Маркова. В этой модели

$$\operatorname{cov}(\varepsilon_t, \varepsilon_{t-l}) = 0 \quad l > Q.$$

Основной недостаток этой модели зависимости ошибок регрессии состоит в том, что трудно провести корректировку модели на автокорреляцию.

3.5. Корректировка модели на гетероскедастичность и автокорреляцию

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

$$\operatorname{cov}(\varepsilon_t, \varepsilon_{t-j}) = 0 \text{ при } j > Q \tag{3.15}$$

(или $\mathsf{M}(\varepsilon_t \varepsilon_{t-s}) = 0$ 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 коэффициентов регрессии вычисляется по формуле

$$\widehat{\operatorname{Var}}\left(\widehat{\beta}_{OLS}\right) = n\,(\mathbf{X}'\mathbf{X})^{-1}\,\widehat{S}_n(\mathbf{X}'\mathbf{X})^{-1}, \tag{3.16}$$

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 $\widehat{\operatorname{Var}}\left(\widehat{\beta}_{OLS}\right)$.

Определение. Оценка (3.16) 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 $\operatorname{Var}(\widehat{\beta})$, устойчивой **одновременно и к гетероскедастичности и к автокорреляции** вида (3.15) (heteroskedasticity and autocorrelation consistent standard errors, HAC standard errors)).

В оригинальной работе [23] доказана состоятельность оценки (3.16) 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'a, устойчивую к гетероскедастичности.

Таким образом, при выполнении условия (3.15) 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 α 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 χ^2_q. При заданном уровне значимости α 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 (3.11) авторегрессии порядка p, то условие (3.15) не выполнено. В работе [23] доказано, что в этом случае оценка (3.16) будет состоятельной, если $Q = Q(n)$ выбрано так, что $Q(n)$ «достаточно велико», а отношение $Q(n)/\sqrt[4]{n}$ «достаточно мало»[6].

3.6. Задачи

Упражнение 1. Вывести систему нормальных уравнений для парной линейной модели регрессии для WLS с произвольными весами $\{w_i\}$.

[6] Точнее, состоятельность оценки доказана при условии $Q = Q(n) \to +\infty$ и $Q(n)/\sqrt[4]{n} \to 0$ при $n \to +\infty$.

Упражнение 2. Предположим, что оценивается модель регрессии

$$y = \beta_0 + \beta_1 x + u,$$

для ошибок которой выполнены условия $\mathsf{M}(u|x) = 0$ и $\mathrm{Var}(u|x) = \sigma^2 x^2$. Пусть $\hat{\beta}_1$ есть OLS-оценка коэффициента β_1. Будет ли $\mathsf{M}\hat{\beta}_1$ больше, меньше или равно β_1?

Упражнение 3. Рассмотрим модель регрессии

$$y_i = \beta x_i + u_i$$

для ошибки которой выполнены условия $\mathsf{M}(u_i) = 0$, $\mathrm{cov}(u_i, u_j) = 0$ $(i \neq j)$ и $\mathrm{Var}(u_i) = \sigma^2 x_i^2$.

- Найдите дисперсию OLS-оценки коэффициента β.
- Предложите несмещенную оценку коэффициента β с меньшей дисперсией. Как это соотносится с теоремой Гаусса-Маркова?

Упражнение 4. Вы оцениваете регрессию $y = \beta_0 + \beta_1 x + u$ по методу наименьших квадратов (стохастические регрессоры). Но $u = \gamma z + \nu$ с ν независящей от x и z. Объясните как это может повлиять на смещенность OLS-оценок коэффициентов β_0 и β_1.

Упражнение 5. На основе квартальных данных с 2003 по 2008 год было получено следующее уравнение регрессии, описывающее зависимость цены на товар y_t от нескольких факторов:

$$\widehat{y}_t = 3.5 + \underset{(0.001)}{0.4}\, x_t + \underset{(0.01)}{1.1}\, w_t; \quad \mathrm{ESS} = 70.4; \quad \mathrm{RSS} = 40.5.$$

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%).

Упражнение 6. На основе квартальных данных с 2001 по 2008 год было получено следующее уравнение регрессии, описывающее зависимость цены y_t на товар от нескольких факторов:

$$\widehat{y}_t = 9.5 + \underset{(0.01)}{0.2}\, x_t + \underset{(0.002)}{0.6}\, w_t + \underset{(0.02)}{0.7}\, z_t; \quad \mathrm{ESS} = 62.5 \quad \mathrm{RSS} = 44.7.$$

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?

Упражнение 7. 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%.

Упражнение 8. 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%.

Упражнение 9. 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.

- Пусть $\hat{\beta}_4 = -0.11$. Дайте интерпретацию полученной оценки.
- 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

$$\widehat{\ln(wage)} = \underset{(0.149)}{1.154} + \underset{(0.003)}{0.0493}ed + \underset{(0.007)}{0.069}age - \underset{(0.00008)}{0.0007}age^2 + \underset{(0.0803)}{0.0963}imm - \underset{(0.0073)}{0.0314}(ed \cdot imm) \quad R^2 = 0.209$$

  Какой ожидаемый эффект от дополнительного года образования для иммигранта?

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

Упражнение 10. Исследуется зависимость логарифма цены коттеджа ($\ln Price$) от его площади ($Square$), удаленности от МКАД ($Dist$) и количества этажей ($Floor$). Было высказано предположение, что с увеличением площади коттеджа дисперсия ошибок регрессии возрастает. Для проверки этого предположения отдельно оценили модель регрессии по 19 коттеджам небольшой площади и по 19 коттеджам большой площади (всего в выборке 45 коттеджей) и получили остаточные суммы квадратов $\mathrm{RSS}_1 = 11.7$ и $\mathrm{RSS}_2 = 25.4$. Можно ли сделать вывод о возрастании дисперсий ошибок регрессий? Уровень значимости 5%.

Упражнение 11. 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 предположения отдельно оценили регрессию по 42 коттеджам небольшой площади и по 42 коттеджам большой площади (всего в выборке 100 коттеджей) и получили остаточные суммы квадратов $\mathrm{RSS}_1 = 11.2$ и $\mathrm{RSS}_2 = 27.2$ соответственно.

1. Напишите спецификацию уравнения регрессии.

2. Можно ли сделать вывод о возрастании дисперсий ошибок регрессий при увеличении площади коттеджа? Уровень значимости 10%.

Упражнение 12. 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 White'a.

- Опишите процедуру теста White'a и напишите спецификацию вспомогательной модели регрессии.

- Пусть во вспомогательной модели регрессии $R_0^2 = 0.075$. Можно ли сделать вывод о гетероскедастичности в ошибках регрессии? Уровень значимости 1%.

Упражнение 13. Оценивается эффект от членства в профсоюзе на почасовую оплату труда. Пусть $union$ – фиктивная переменная, отвечающая за членство в профсоюзе, $wage$ – уровень почасовой оплаты труда.

- Предложите спецификацию регрессии для измерения влияния членства в профсоюзе на логарифм зарплаты.

- Предложите спецификацию модели, чтобы влияние членства в профсоюзе могло быть различным для мужчинами и женщинами.

- Предложите спецификацию чтобы эффект от членства в профсоюзе мог зависеть от наличия высшего образования (фиктивная переменная) и от опыта работы.

Упражнение 14. Рассматривается регрессионная модель зависимости логарифма зарплаты ($\ln w$) от уровня образования (edu, в годах) и членства в профсоюзе (фиктивная переменная $union$, равная 1 если член профсоюза). Предложите такую спецификацию модели, чтобы эффект от образования мог зависеть от пола.

Упражнение 15. Рассматривается регрессионная модель зависимости логарифма зарплаты ($\ln w$) от уровня образования (edu, в годах) и пола (фиктивная переменная $male$, равная 1 для мужчин). Предложите такую спецификацию модели, чтобы эффект от образования мог зависеть от того, является ли индивидуум иммигрантом или нет.

Упражнение 16. Рассматривается регрессионная модель зависимости логарифма зарплаты ($\ln w$) от уровня образования (edu, в годах), уровня IQ ($iqscore$) и членства в профсоюзе (фиктивная переменная D). Предложите такую спецификацию модели, чтобы эффект от образования и уровня IQ мог зависеть от пола.

Упражнение 17. Рассматривается регрессионная модель зависимости логарифма уровня зарплаты ($\ln w$) от уровня образования (edu, в годах) и от места работы (фиктивная переменная $smsa$, равная 1 если работает в городе с населением больше 1млн.). Предложите такую спецификацию модели, чтобы эффект от образования и места работы мог зависеть от того, является ли индивидуум иммигрантом или нет.

Упражнение 18. Рассмотрим регрессионную модель 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 $male$

$$\widehat{\ln(Wage)} = \beta_0 + \beta_1 edu + \beta_2 age + \beta_3 age^2 + \beta_4 male + \beta_5 city + \beta_6 (city \cdot male)$$

Модель регрессии была отдельно оценена по выборкам из 35 иммигрантов и из 23 граждан страны и были получены остаточные суммы квадратов $\text{RSS}_{imm} = 40.2$ и $\text{RSS}_{native} = 60.2$. 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%.

Упражнение 19. Используя данные 250 случайно отобранных мужчин и 280 случайно выбранных женщин было оценено зарплатное уравнение

$$\widehat{Wage} = \underset{(0.23)}{12.52} + \underset{(0.36)}{2.12} * Male \quad R^2 = 0.06 \quad \text{SER} = 4.2$$

где $Wage$ – уровень почасовой оплаты труда, $Male$ – фиктивная (бинарная) переменная, равная 1 для мужчин.

- Какая средняя разность в уровне оплаты труда между мужчинами и женщинами?
- Значима ли разница в уровне оплаты труда между мужчинами и женщинами? Уровень значимости 1%.
- Постройте 95% доверительный интервал для разницы в уровне оплаты труда.
- Какой средний выборочный уровень оплаты женщин? Мужчин?

Упражнение 20. В условиях предыдущей задачи другой исследователь на основе тех же данных оценил регрессионное уравнение используя фиктивную переменную $Female$, равную 1 для женщин

$$\widehat{Wage} = \gamma_0 + \gamma_1 Female.$$

- Чему равны оценки коэффициентов $\hat{\gamma}_0$ и $\hat{\gamma}_1$?
- Чему равны R^2 и SER в этой модели?

Упражнение 21. Был проведен эксперимент: школьники младших классов случайным образом были назначены в «обычные» или «маленькие» классы («обычные» состояли из 24 учеников, «маленькие» – из 15) и в конце учебного года им дали одинаковый тест. Пусть $SmallClass$ – бинарная переменная, равная 1 если школьник был назначен в «маленький» класс. Была оценена регрессия

$$\widehat{TestScore} = \underset{(1.6)}{918.0} + \underset{(2.5)}{13.9} \cdot SmallClass \quad R^2 = 0.01 \quad n = 200$$

- Улучшают ли «малые» классы оценку за тест и на сколько?
- Значимо ли назначение в «малый» класс влияет на оценку за тест? Сформулируйте проверяемую гипотезу. Уровень значимости 5%.
- Постройте 99% доверительный интервал для эффекта влияния «малого» класса на оценку за тест.

Упражнение 22. Рассмотрим модель зависимости почасовой оплаты труда $Wage$ от уровня школьного образования ($educ$, в годах), пола (бинарная переменная $female$) и наличия высшего образования (бинарная переменная $heduc$)

$$\widehat{Wage} = 12.3 + \underset{(0.0008)}{0.0023} educ - \underset{(0.09)}{0.23} female + \underset{(0.78)}{2.3}\, heduc - \underset{(0.78)}{0.93}(heduc \cdot female) \quad n = 27.$$

Какая средняя разность в почасовой оплате между мужчинами и женщинами без высшего образования? С высшим образованием? Какой средний эффект от наличия высшего образования для мужчин? Для женщин? Значима ли разность в оплате труда между мужчинами и женщинами без высшего образования? Уровень значимости 5%.

Упражнение 23. 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)

$$\widehat{Wage} = \underset{(1.2)}{11.4} + \underset{(0.0006)}{0.002}\,educ + \underset{(0.17)}{0.4}\,male + \underset{(0.78)}{0.6}\,(male \cdot smsa) - \underset{(1.04)}{2.3}\,imm + \underset{(0.078)}{1.34}\,smsa \quad n = 33.$$

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%.

Упражнение 24. Оценивается зависимость уровня школьного образования $YearSch$ в Швеции. Влияющие переменные: $HighAb$ – фиктивная переменная, показывающая способности выше среднего уровня, $Funsk$ – фиктивная переменная, показывающая, что отец не имел образования, $Reform$ – фиктивная переменная, показывающая обучался ли в новой реформированной школе, $Female$ – фиктивная переменная, показывающая пол человека, $City$ – фиктивная переменная, показывающая вырос ли в городе. На основе случайной выборки 11174 людей была оценена регрессия

$$\widehat{Yearsch} = \underset{(0.1332)}{12.3994} + \underset{(0.0608)}{1.9361}HighAb + \underset{(0.0686)}{0.1519}Female - \underset{(0.1112)}{0.5721}(Female \cdot City) - \underset{(0.0735)}{2.2345}Funsk + \underset{(0.0747)}{0.5121}Reform + \underset{(0.1290)}{0.6921}City, \quad R^2 = 0.1998$$

- Какая средняя разница в уровне образования между мужчинами и женщинами живущими в городе? Вне города?
- Какая средняя разница в уровне образования между новой и старой школьной системами?
- Проверьте гипотезу, что уровень образования мужчин в городе не отличается от уровня образования мужчин вне города. Уро-

вень значимости 1%. Следует ли из этого, что городские жители более или менее образованны?

Упражнение 25. В условиях предыдущей задачи была оценена без учета различия город/не город (без учета факторов $City$ и $Female \cdot City$)

$$\widehat{Yearsch} = \underset{(0.0941)}{12.7108} + \underset{(0.0613)}{1.9408} HighAb - \underset{(0.0592)}{0.0780} Female - \underset{(0.0812)}{2.2898} Unsk + \underset{(0.0747)}{0.5121} Reform, \quad R^2 = 0.1938$$

Тестируйте гипотезу об отсутствии различия в уровне образования между жителями города и сельской местности.

Упражнение 26. Рассмотрим модель регрессии с фиктивной переменной x_i

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

Пусть в выборке влияющей переменной n_1 значений 0 и n_2 значений 1 ($n_1 + n_2 = n$). Обозначим $\bar{y}^{(p)}$ – выборочное среднее и $\widehat{\sigma}_p^2$ – выборочная дисперсия тех значений зависимой переменной, для которых $x_i = p$ ($p = 0, 1$). Пусть $\bar{y}$ – общее среднее, $\widehat{\sigma}^2$ – общая выборочная дисперсия y (по всей выборке).

- Выразите оценки коэффициентов регрессии $\hat{\beta}_0$ и $\hat{\beta}_1$ через n_1, n_2, $\bar{y}^{(0)}$, $\bar{y}^{(1)}$, $\bar{y}$, $\widehat{\sigma}_0^2$, $\widehat{\sigma}_1^2$, $\widehat{\sigma}^2$.

- Можно ли s_1^2 выразить через n_1, n_2, $\bar{y}^{(0)}$, $\bar{y}^{(1)}$, $\bar{y}$, $\hat{\sigma}_0^2$, $\hat{\sigma}_1^2$, $\hat{\sigma}^2$?

Упражнение 27. Модель $\widehat{INF} = \beta_0 + \beta_1 UNEM$ описывает зависимость инфляции (INF) от безработицы ($UNEM$). Оценивается регрессия

$$\widehat{INF} = \gamma_0 + \gamma_1 UNEM + \gamma_2 N,$$

где N – число солнечных пятен.

- Будет ли $\hat{\gamma}_1$ несмещенной оценкой коэффициента β_1?

- Верно ли неравенство $s_{\gamma_1} < s_{\beta_1}$?

Ответ кратко поясните.

Упражнение 28. 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)

$$\widehat{\ln y_t} = 3.4 + 0.3\ln(u_t) + 0.03\ln(z_t) \quad R^2 = 0.69 \quad \text{DW} = 3.22.$$

Будут ли ошибки в модели регрессии автокоррелированны (первого порядка)? Если да, то автокорреляция положительная или отрицательная?

Упражнение 29. 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?

Упражнение 30. 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?

Упражнение 31. 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?

Упражнение 32. 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 ?

Упражнение 33. Рассмотрим кривую Филлипса (Phillips curve) с запаздыванием, связывающую темпы инфляции inf с уровень безработицы $unem$ и уровнем безработицы с лагом один $unem_{t-1}$:

$$\widehat{inf}_t = 0.78 - 0.24unem_t - 0.02unem_{t-1} \quad R^2 = 0.097 \quad n = 57.$$

a) Опишите двухшаговую процедуру тестирования ошибок модели регрессии на автокорреляцию первого порядка и напишете спецификацию вспомогательной регрессии.

b) Пусть во вспомогательной регрессии $R_0^2 = 0.16$. Можно ли сделать вывод об автокоррелированности (первого порядка) ошибок регрессии в кривой Филлипса? Уровень значимости 5%.

c) Можно ли в данной модели регрессии для тестирования ошибок регрессии на автокорреляцию первого порядка использовать тест Durbin – Watson ?

Упражнение 34. 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}

$$\widehat{inf}_t = 1.23 + 0.12inf_{t-1} - 0.12unem_t - 0.002unem_{t-1}$$
$$R^2 = 0.12 \quad n = 60.$$

a) Опишите двухшаговую процедуру тестирования ошибок регрессии на автокорреляцию третьего порядка и напишете 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 Durbin – Watson ?

Упражнение 35 ([29]). Рассмотрим регрессионную модель зависимости доходности краткосрочных казначейских облигаций США $i3$ (3-х месячный T-bill rate) от годового уровня инфляции inf (основанного на consumer price index CPI) и дефицита бюджета def (в процентах от GDP), основанную на данных с 1948 по 2003 года

$$\widehat{\ln(i3_t)} = 1.733 + 0.606 \ln(inf_t) + 0.513 \ln(def_t)$$
$$R^2 = 0.602 \quad \text{DW} = 0.716$$

a) Опишите двухшаговую процедуру тестирования ошибок модели регрессии на автокорреляцию второго порядка и напишете спецификацию вспомогательной регрессии.

b) Пусть во вспомогательной регрессии $R_0^2 = 0.371$. Можно ли сделать вывод об автокоррелированности (второго порядка) ошибок регрессии? Уровень значимости 5%.

c) Можно ли в данной модели регрессии для тестирования ошибок регрессии на автокорреляцию первого порядка использовать тест Durbin–Watson ? Если да, то тестируйте ошибки на автокорреляцию первого порядка.

Глава 4

Модели временны́х рядов

При построении эконометрических регрессионных моделей для временных рядов необходимо учитывать следующие особенности:

- фактор времени естественным образом упорядочивает данные, т.е. важен порядок, в котором записаны данные временного ряда (в отличие от пространственной выборки);
- в отличие от пространственной выборки во временном ряду естественно допускать зависимость элементов ряда в различные моменты времени («эффект памяти»);
- часто приходится оценивать регрессионные модели по небольшим выборкам и нет возможности получить выборочные данные большого объема. Например, такая ситуация может возникать при работе с годовыми макроэкономическими данными.

4.1. Условия Гаусса – Маркова для регрессионных моделей временны́х рядов

Будем рассматривать многофакторную модель регрессии

$$y_t = \beta_0 + \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + u_t, \quad t = 1, \ldots, n \tag{4.1}$$

для временны́х рядов $\{y_t, x_{t1}, \ldots, x_{tk}\}$. Основное отличие этой модели от модели для случайных пространственных выборок (cross-section data) состоит в том, что случайные величины, формирующие временной ряд, не обязаны быть независимыми. Соответственно, условия Гаусса – Маркова должны быть скорректированы.

Введем обозначение

$$\mathbf{x}_t = \begin{pmatrix} x_{t1} \\ \vdots \\ x_{tk} \end{pmatrix} \ (t = 1, \ldots, n), \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

Тогда уравнение (4.1) может быть записано в матричном виде

$$y_t = \beta_0 + \mathbf{x}_t'\beta + u_t.$$

Относительно ошибок регрессии будем предполагать выполнение следующих условий:

1. $\mathsf{M}(u_t|\mathbf{x}_1, \ldots, \mathbf{x}_n) = 0$;
2. $\mathrm{Var}(u_t|\mathbf{x}_1, \ldots, \mathbf{x}_n) = \mathrm{Var}(u_t) = \sigma^2$;
3. $\mathrm{cov}(u_t, u_s|\mathbf{x}_1, \ldots, \mathbf{x}_n) = 0$;
4. $u_t|\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \mathcal{N}(0, \sigma^2)$.

Замечание. отличие от модели регрессии для пространственных выборок состоит в том, что условное математическое ожидание и условная дисперсия берутся при условии $\mathbf{x}_1, \ldots, \mathbf{x}_n$, т.к. допускается коррелирование элементов временного ряда в разные моменты времени.

Теорема. *Пусть ошибки модели регрессии удовлетворяют условиям 1. – 3. и ни один из регрессоров не выражается линейно через остальные. Тогда OLS-оценки коэффициентов в модели* (4.1) *будут BLUE-оценками.*

Замечание. Как и в случае пространственных выборок для несмещенности OLS-оценок достаточно условия 1. на ошибки регрессии.

Теорема. *Пусть для модели регрессии выполнены условия предыдущей теоремы и выполнено условие 4. нормальной распределенности ошибок. Тогда для OLS-оценок коэффициентов в модели регрессии* (4.1) *верны статистические выводы регрессии для пространственных выборок.*

Для модели регрессии (4.1) применимы все описанные выше методы исследования на функциональную форму и гетероскедастичность.

Следует обратить внимание на два вида моделей регрессии для временных рядов:

- *статические модели* (static model) включает объясняющие переменные, взятые за тот же период времени, что и зависимая переменная. Пример - статическая кривая Филлипса, описывающая зависимость уровня инфляции inf от уровня безработицы $unem$

$$inf_t = \beta_0 + \beta_1 unem_t + u_t$$

- *модель (конечных) распределенных лагов* (finite distributed lag, FDL model) содержит лаговые значения регрессоров.

Замечание. Также в модель можно включить лаговые значения зависимой переменной, однако этот случай не описывается вышеизложенной моделью, так нарушаются условия 1. – 3. на ошибки регрессии. Эта модель будет рассмотрена отдельно.

4.2. Модель тренда и сезонность

В экономическом анализе встречаются временные ряды имеющие (в среднем) устойчивую тенденцию к возрастанию с течением временем. Поведение таких временных рядов можно описывать регрессионной моделью тренда, где в качестве объясняющей переменной выступает фактор времени.

Модель линейного тренда задается уравнением

$$y_t = \beta_0 + \beta_1 t + u_t, \quad t = 1, \ldots, n.$$

Будем предполагать, что ошибки $\{u_t\}$ удовлетворяют условиям теоремы Гаусса – Маркова. Тогда к модели линейного тренда применимы выводы стандартной линейной модели регрессии. В частности, среднее значение $\mathsf{M}y_t$ линейно зависит от времени t:

$$\mathsf{M}y_t = \beta_0 + \beta_1 t.$$

Коэффициент β_1 имеет следующую интерпретацию: это есть среднее приращение временного ряда за один период времени

$$\triangle \mathsf{M}y_t = \mathsf{M}y_t - \mathsf{M}y_{t-1} = \beta_1.$$

Следовательно, с увеличением времени,

- при $\beta_1 > 0$ во временном ряду есть «тенденция к возрастанию»,

- при $\beta_1 < 0$ во временном ряду есть «тенденция к убыванию»,

причем средняя скорость изменения временного ряда за один период времени постоянна.

Модель экспоненциального тренда задается уравнением

$$\ln(y_t) = \beta_0 + \beta_1 t + u_t.$$

Будем предполагать, что ошибки $\{u_t\}$ удовлетворяют условиям теоремы Гаусса – Маркова. Тогда к модели линейного тренда применимы выводы стандартной линейной модели регрессии. В частности, среднее значение зависит от t экспоненциально

$$\mathsf{M}\ln(y_t) = \beta_0 + \beta_1 t.$$

Для коэффициента β_1 получаем следующую интерпретацию:

$$\triangle\mathsf{M}\ln(y_t) = \mathsf{M}\ln(y_t) - \mathsf{M}\ln(y_{t-1}) = \mathsf{M}\ln\left(\frac{y_t}{y_{t-1}}\right) = \beta_1$$

Следовательно, за один период времени (в среднем) значение y_t изменяется в $\exp(\beta_1)$ раз.

Если β_1 мало, то $\exp(\beta_1) \approx 1 + \beta_1$ и за один период времени в среднем значение y_t изменяется (в первом приближении) на $\beta_1 \cdot 100\%$.

Другие модели тренда. Наряду с линейным и экспоненциальным трендом в прикладных задачах могут встречаться и другие функциональные формы трендов. Например, **квадратичный тренд**, задаваемый уравнением

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + u_t.$$

Для выбора функциональной модели тренда применимы методы и тесты модели регрессии на функциональную форму.

Использование временны́х рядов с трендом в регрессионных моделях

Применение трендовых временных рядов в качестве зависимой и объясняющих переменных имеет важную особенность. Поясним ее на примере модели с одной объясняющей переменной. Итак, пусть

$$\begin{aligned} y_t &= \alpha_0 + \alpha_1 t + u_t, & \alpha_1 &\neq 0, \\ x_t &= \gamma_0 + \gamma_1 t + v_t, & \gamma_1 &\neq 0, \end{aligned}$$

и оценивается линейная модель регрессии

$$y_t = \beta_0 + \beta_1 x_t + \text{error}\,.$$

Но тогда мы имеем *проблему невключения значимого фактора* (который «коррелирует» с x_t), а именно фактора времени t. Это приводит к смещению OLS-оценок параметров регрессии, в частности коэффициент β_1 может оказаться значимым, хотя из экономических соображений факторы должны быть независимыми. Описанная проблема называется **ложной регрессией** (spurious regression problem). Необходимо учесть тренд (включить в модель значимый фактор времени) и оценивать регрессию

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 t + \text{error}\,.$$

Пример (Housing investment & Prices [29]). На основе годовых данных с 1947 по 1988 года ($n = 42$) была оценена лог-линейная модель зависимости инвестиций в строительство ($invpc$) от индекса цен на дома ($price$, равен 1 для 1982 г.):

$$\widehat{\ln(invpc)} = -0.550 + 1.241 \ln(price)$$
$$s_0 = 0.043, \quad s_1 = 0.382, \quad R^2 = 0.208.$$

Согласно этой модели, эластичность $invpc$ по $price$ значима и положительна. Оба временных ряда имеют возрастающие значимые тренды:

$$\widehat{\ln(invpc)}_t = \hat{\alpha}_0 + 0.0081t, \qquad s_1 = 0.0018$$
$$\widehat{\ln(price)}_t = \hat{\gamma}_0 + 0.0044t, \qquad s_1 = 0.0004.$$

Чтобы учесть трендовое поведение факторов в модель необходимо включить временной тренд

$$\widehat{\ln(invpc)} = -0.913 - 0.381 \ln(price) + 0.0098t$$
$$s_0 = 0.136, \quad s_1 = 0.697, \quad s_2 = 0.0035, \quad R^2 = 0.307$$

В этой модели эластичность отрицательна и незначима. Временной тренд значим и показывает увеличение $invpc$ за год в среднем на 0.98%.

Замечание. Включение в модель трендовой переменной может и «повышать значимость» существенных объясняющих переменных.

Используя формулы для двухфакторной регрессии несложно показать, что включение в модель трендовой переменной равносильно следующей двухшаговой процедуре:

1. «детрендируем» зависимую и объясняющую переменные: вычисляем $\dot{y}_t$ и $\dot{x}_t$ – остатки в моделях тренда

$$y_t = \alpha_0 + \alpha_1 t + u_t$$
$$x_t = \gamma_0 + \gamma_1 t + v_t,$$

   соответственно;

2. оцениваем парную модель регрессию

$$\dot{y}_t = \beta_0 + \beta_1 \dot{x}_t + \text{error}$$

Тогда оценки коэффициентов β_0 и β_1 в регрессии y_t на x_t, t и в регрессии $\dot{y}_t$ на $\dot{x}_t$ совпадают.

Сезонность

В некоторых временных рядах, особенно полученных на основе месячных или квартальных (иногда недельных или дневных) данных может наблюдаться сезонность или периодичность.

Пример. Объем продаж мороженного имеет выраженную сезонность, что связано с погодными условиями. Число постояльцев курортного отеля также может иметь выраженную сезонность, что также связано с погодными условиями. Однако в финансовых данных (доходности и т.д.) как правило сезонность не наблюдается.

Для учета сезонности и периодичности в модель регрессии вводят фиктивные переменные.

Замечание. Следует отметить, что часто статистические данные публикуются с поправкой на сезонность (seasonally adjusted), так что учитывать ее не нужно. Например, квартальные данные U.S. GDP публикуются с исключением сезонности.

4.3. Модель распределенных лагов

Как уже отмечалось, *модель распределенных лагов* (finite distributed lag, FDL model) содержит в качестве регрессоров лаговые значения объясняющих переменных. При выполнении условий на ошибки регрессии к данной модели применимы все статистические выводы многофакторной регрессии.

Интерпретацию коэффициентов в модели распределенных лагов рассмотрим на примере модели FDL(2) с одним регрессором

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 x_{t-1} + \beta_3 x_{t-2} + u_t$$

Пусть в момент времени t значение объясняющей переменной увеличилось на единицу. Тогда ожидаемое изменение зависимой переменной в тот же момент времени t равен

$$\triangle \mathsf{M} y_t = \beta_1,$$

а ожидаемое изменение в будущие периоды времени равны

$$\triangle \mathsf{M} y_{t+1} = \beta_2, \quad \triangle \mathsf{M} y_{t+2} = \beta_3,$$
$$\triangle \mathsf{M} y_{t+l} = 0, \quad l \geqslant 3$$

Коэффициент β_1 есть таким образом отклик за один период или краткосрочный мультипликатор (impact multiplier).

Пусть начиная с момента времени t значения регрессора увеличилось на единицу. Тогда ожидаемое изменение зависимой переменной равно

$$\begin{aligned} \triangle \mathsf{M} y_t &= \beta_1 \\ \triangle \mathsf{M} y_{t+1} &= \beta_1 + \beta_2 \\ \triangle \mathsf{M} y_{t+l} &= \beta_1 + \beta_2 + \beta_3, \quad l \geqslant 2. \end{aligned}$$

Долгосрочный отклик или долгосрочный мультипликатор (long-run multiplier) для модели FDL(2) таким образом равен $\beta_1 + \beta_2 + \beta_3$. Это можно рассматривать как наличие «долгосрочной зависимости» между факторами вида

$$y^* = \beta_0 + (\beta_1 + \beta_2 + \beta_3) x^*.$$

Аналогично определяется долгосрочные мультипликаторы и «долгосрочная зависимость» для произвольной модели FDL.

4.4. Модель авторегрессии временных рядов

При анализе экономических и финансовых показателей могут возникать ситуации, когда изучаемый процесс не имеет устойчивой тенденции к росту или убыванию (тренда), а представляет собой «случайные колебания» вокруг некоторого среднего уровня. Такие явления могут иметь место, например, в случае, когда экономические показатели характеризуются случайными отклонениями под воздействием внешних факторов от положения равновесия, т.е. описывают случайные колебания системы, находящейся в равновесии. Для описания таких классов временных рядов используются вероятностные модели стационарных временных рядов.

4.4.1. Стационарные временные ряды

Определение. Временной ряд y_t называется *стационарным* (*в широком смысле*), если

$$\mathsf{M}y_t \equiv \text{const}, \quad \text{cov}(y_t, y_{t+h}) = \gamma(h) \quad (h = 0, \pm 1, \pm 2 \ldots)$$

Понятие стационарного временного ряда означает, что его среднее значение не изменяется во времени, т.е. временной ряд не имеет тренда. Кроме того, ковариация между разными элементами временного ряда (как между случайными величинами) зависит только от того, насколько сильно они отдалены друг от друга во времени. Величина h, характеризующая разницу во времени между элементами временного ряда, называется *лаговой переменной* или *запаздыванием*. Так как

$$\gamma(0) = \text{cov}(y_t, y_t) = \text{Var}(y_t),$$

то дисперсия стационарного временного ряда также не меняется со временем.

Определение. Функция $\gamma(h)$ как функция от лаговой переменной, называется *автоковариационной функцией* временного ряда.

Она определена как для положительных, так и для отрицательных лагов h. Так как

$$\gamma(-h) = \text{cov}(y_t, y_{t-h}) = \text{cov}(y_{t-h}, y_t) = \text{cov}(y_\tau, y_{\tau+h}) = \gamma(h),$$

то $\gamma(h)$ – четная функция. Для произвольных моментов времени t и s очевидно равенство

$$\operatorname{cov}(y_t, y_s) = \gamma(t-s)$$

Вычислим теперь коэффициенты корреляции между разными элементами стационарного временного ряда с временным лагом h:

$$\operatorname{corr}(y_t, y_{t+h}) = \frac{\operatorname{cov}(y_t, y_{t+h})}{\sqrt{\operatorname{Var}(y_t) \cdot \operatorname{Var}(y_{t+h})}} = \frac{\gamma(h)}{\sqrt{\gamma(0) \cdot \gamma(0)}} = \frac{\gamma(h)}{\gamma(0)}.$$

Также как и в случае коэффициента ковариации, коэффициент корреляции между разными элементами стационарного временного ряды зависит только от лага между ними.

Определение. Функция $\rho(h) = \operatorname{corr}(y_t, y_{t+h})$ называется *автоковариационной функцией* (autocorrelation function, ACF) стационарного временного ряда.

Очевидно, что она также является четной функцией лаговой переменной и $\rho(0) = 1$. Для коэффициента автокорреляции очевидно:

$$\operatorname{corr}(y_s, y_t) = \rho(s-t).$$

Предложение. *Для произвольного стационарного ряда существует предел автокорреляционной функции*

$$\lim_{h \to \pm\infty} \rho(h) = 0.$$

Это означает, что с ростом временного лага элементы временного ряда становятся «менее коррелированными». Это можно интерпретировать следующим образом: с ростом времени t временной ряд «забывает свои прошлые состояния», так как $\operatorname{corr}(y_s, y_t) = \rho(t-s) \to 0$ при $t \to +\infty$ если s фиксированно.

Определение. *Коррелограммой* стационарного временного ряда называется график функции $\rho(h)$.

Пример. Рассмотрим основной пример стационарного временного ряда. Пусть $\{y_t\}$ – последовательность некоррелируемых, нормально распределенных случайных величин с нулевым средним и одинаковой

дисперсией. Временной ряд с такими свойствами называется (гауссовским) *белым шумом* (white noise). Из определения следует, что ACF для белого шума имеет простой вид

$$\rho(h) = \begin{cases} 1, & h = 0 \\ 0, & h \neq 0 \end{cases}$$

т.е. белый шум «мгновенно забывает» свои прошлые значения.

Замечание. Из определения видно, что временной ряд (гауссовского) белого шума удовлетворяет условиям Гаусса – Маркова на ошибки модели регрессии.

4.4.2. Модель авторегрессии

Рассмотрим одну из основных моделей стационарных временных рядов – модель авторегрессии.

Определение. Модель *авторегрессии*[1] AR(p) *порядка* p определяется разностным регрессионным уравнением между значениями временного ряда и имеет следующий вид:

$$y_t = \mu + \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + \varepsilon_t, \quad \beta_p \neq 0, \tag{4.2}$$

ε_t – белый шум.

Замечание. В модели авторегрессии ряд белого шума иногда называется также *обновляющей* или *инновационной последовательностью.*

Согласно модели (4.2), значения временного ряда y_t складывается из суммы прошлых p значений временного ряда и величины ε_t, отвечающей за влияние внешних факторов в момент времени t.

В общем случае временной ряд, определяемый разностным уравнением (4.2), не будет стационарным. Необходимо накладывать дополнительные условия. Для формулировки условий стационарности модели авторегрессии AR(p) введем многочлен степени p следующим равенством

$$\beta(z) = 1 - \beta_1 z - \cdots - \beta_p z^p.$$

[1] AR = AutoRegression

Теорема (Условия стационарности)**.** *Временной ряд, определяемый разностным уравнением* (4.2)*, является стационарным тогда и только тогда, когда* ***все*** *корни многочлена* $\beta(z)$ *(в том числе и комплексные) по модулю больше единицы*

Пример. Рассмотрим модель авторегрессии первого порядка AR(1), задаваемую уравнением

$$y_t = \mu + \beta_1 y_{t-1} + \varepsilon_t, \quad \beta_1 \neq 0.$$

Для этой модели многочлен $\beta(z) = 1 - \beta_1 z$ имеет единственный корень $z_0 = 1/\beta_1$. Следовательно, регрессионная модель AR(1) определяет стационарный временной ряд в том и только в том случае, когда

$$|z_0| > 1 \Leftrightarrow |\beta_1| < 1.$$

При $\beta_1 = 1$ и $\mu = 0$ модель авторегрессии первого порядка определяет нестационарный временной ряд

$$y_t = \varepsilon_1 + \varepsilon_{t-1} + \cdots + \varepsilon_t = \sum_{s=1}^{t} \varepsilon_s,$$

называемый *случайным блужданием* (random walk), для которого, очевидно, $\mathsf{M}y_t = 0$, но $\operatorname{Var}(y_t) = t \operatorname{Var}(\varepsilon)$, т.е. со временем дисперсия временного ряда возрастает.

Пример. Рассмотрим модель авторегрессии второго порядка AR(2), задаваемую уравнением

$$y_t = 3 + 2y_{t-1} - y_{t-2} + \varepsilon_t.$$

Для нее многочлен $\beta(z) = 1 - 2z + z^2$ имеет единственный корень $z_0 = 1$ кратности два. Следовательно, уравнение авторегрессии задает нестационарный временной ряд.

Найдем теперь среднее значение стационарного временного ряд, описываемого моделью авторегрессии (4.2). Для этого возьмем математическое ожидание от обеих частей регрессионного уравнения:

$$\mathsf{M}y_t = \mu + \beta_1 \mathsf{M}y_{t-1} + \cdots + \beta_p \mathsf{M}y_{t-p} + \mathsf{M}\varepsilon_t.$$

Так как $\mathsf{M}\varepsilon_t = 0$ и для стационарного временного ряда $\mathsf{M}y_t \equiv \text{const}$, то получаем

$$\mathsf{M}y_t = \frac{\mu}{1 - \beta_1 - \cdots - \beta_p}.$$

Можно доказать следующий результат

Лемма. *Для стационарной модели авторегрессии*

$$\operatorname{cov}(\varepsilon_t, y_{t-h}) = 0, \quad h > 0,$$

т.е. обновляющая последовательность ***не коррелирует*** *с прошлыми значениями временного ряда*

Найдем как связаны коэффициенты авторегрессионного уравнения и ACF. Для этого проведем следующие выкладки:

$$\begin{aligned}\gamma(h) = \gamma(-h) = \operatorname{cov}(y_t, y_{t-h}) = \\ \operatorname{cov}(\mu + \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + \varepsilon_t, y_{t-h}) = \\ \beta_1 \operatorname{cov}(y_{t-1}, y_{t-h}) + \cdots + \beta_p \operatorname{cov}(y_{t-p}, y_{t-h}) + \operatorname{cov}(\varepsilon_t, y_{t-h}) = \\ \beta_1\gamma_1(h-1) + \cdots + \beta_p\gamma_p(h-p).\end{aligned}$$

Разделив на $\gamma(0)$ получим линейное уравнение

$$\rho(h) = \beta_1\rho(h-1) + \beta_2\rho(h-2) + \cdots + \beta_p\rho(h-p).$$

Объединив эти уравнения при $h = 1, \ldots, p$ и воспользовавшись четностью функции $\rho(h)$, получим *систему линейных уравнений Юла-Уолкера*, которая показывает зависимость между коэффициентами регрессии β_k и коэффициентами автокорреляции $\rho(h)$:

$$\begin{cases} \rho(1) &= \beta_1\rho(0) + \beta_2\rho(1) + \cdots + \beta_p\rho(p-1) \\ \rho(2) &= \beta_1\rho(1) + \beta_2\rho(0) + \cdots + \beta_p\rho(p-2) \\ &\vdots \\ \rho(p) &= \beta_1\rho(p-1) + \beta_2\rho(p-2) + \cdots + \beta_p\rho(0) \end{cases}$$

Модель AR(1)

Рассмотрим подробнее модель авторегрессии первого порядка

$$y_t = \mu + \beta_1 y_{t-1} + \varepsilon_t.$$

Как мы знаем, эта модель определяет стационарный временной ряда только в случае $|\beta_1| < 1$.

Уравнение Юла–Уолкера для модели AR(1) имеет вид

$$\rho(1) = \beta_1\rho(0),$$

откуда с учетом равенства $\rho(0) = 1$, получаем $\rho(1) = \beta_1$.

Далее, так как $\rho(h) = \beta_1 \rho(h-1)$, то имеем

$$\rho(h) = \beta_1 \rho(h-1) = \beta_1^2 \rho(h-2) = \cdots = \beta_1^h \rho(0) = \beta_1^h, \quad h > 0$$

Следовательно, ACF для модели AR(1) представляет собой бесконечно убывающую геометрическую прогрессию при положительных лагах h. Для произвольных лагов

$$\rho(h) = \beta_1^{|h|}.$$

Для модели авторегрессии первого порядка существую два разных случая поведения выборочного временного ряда: случай положительного и отрицательного коэффициента β_1.

Случай $\beta_1 > 0$ По определению, $\mathrm{corr}(y_t, y_{t+1}) = \rho(1) = \beta_1 > 0$, т.е. соседние члены временного ряда **положительно** коррелированы. Это означает, что в выборочных значениях временного ряда имеет место следующая закономерность:

Во временном ряду есть ***«тенденция к сохранению знака»*** *относительно среднего уровня (математического ожидания): если значение временного ряда больше среднего уровня, то, «типичная ситуация», что и последующее значение также больше среднего уровня и наоборот*

Особенно наглядно эта закономерность выглядит в случае нулевого среднего значения (случай $\mu = 0$): «типичная ситуация», что значение временного ряда имеет тот же знак (положительный или отрицательный), что и предыдущее значение.

Случай $\beta_1 < 0$ По определению, $\mathrm{corr}(y_t, y_{t+1}) = \rho(1) = \beta_1 < 0$, т.е. соседние члены временного ряда **отрицательно** коррелированы. Это означает, что в выборочных значениях временного ряда имеет место следующая закономерность:

Во временном ряду есть ***«тенденция к смене знака»*** *относительно среднего уровня (математического ожидания): если значение временного ряда больше среднего уровня, то, «типичная ситуация», что последующее значение меньше среднего уровня и наоборот.*

В случае нулевого среднего значения (случай $\mu = 0$) эта закономерность особенно наглядна: «типичная ситуация», что значение временного ряда имеет противоположный знак по сравнению с предыдущим значением.

4.4.3. Прогнозирование авторегрессионных случайных процессов

Рассмотрим теперь задачу прогнозирования стационарного временного ряда в рамках модели авторегрессии. Под прогнозом будем понимать наилучший линейный прогноз, т.е. прогноз с наименьшей дисперсией, выражаемый линейным образом через известные прошлые значения временного ряда. Пусть нам полностью известно поведение временного ряда до периода времени n. Будущее значение временного ряда определяется уравнением авторегрессии

$$y_{n+1} = \mu + \beta_1 y_n + \beta_2 y_{n-1} + \cdots + \beta_p y_{n-p+1} + \varepsilon_{n+1}$$

Так как будущее значение инновационной последовательности ε_{n+1} не коррелирует с прошлыми значениями временного ряда

$$\operatorname{corr}(\varepsilon_{n+1}, y_{n-h}) = 0 \quad (h \geqslant 0),$$

то в качестве прогноза естественно взять

$$\widehat{y}_{n+1} = \mu + \beta_1 y_n + \beta_2 y_{n-1} + \cdots + \beta_p y_{n-p+1},$$

Предложение. *Определенный таким образом линейный прогноз будет наилучшим на период времени n+1 (на один шаг) в смысле среднеквадратичного отклонения.*

Среднеквадратичная ошибка прогноза, очевидно, равна

$$\mathsf{M}\,|\widehat{y}_{n+1} - y_{n+1}|^2 = \mathsf{M}\varepsilon_{n+1}^2 = \operatorname{Var}(\varepsilon)$$

Рассмотри теперь задачу прогнозирования на несколько шагов. Наилучший (в смысле средне-квадратичного отклонения) линейный прогноз на период времени $n + 2$ (на два шага) дается формулой

$$\widehat{y}_{n+2} = \mu + \beta_1 \widehat{y}_{n+1} + \beta_2 y_n + \cdots + \beta_p y_{n-p+2};$$

на период времени $n+3$ (на три шага) - формулой

$$\widehat{y}_{n+3} = \mu + \beta_1 \widehat{y}_{n+2} + \beta_2 \widehat{y}_{n+1} + \cdots + \beta_p y_{n-p+3}$$

и т.д. В общем случае получаем:

Записываем для будущего значения авторегрессионное уравнение, отбрасываем ε_t и подставляем вместо y_t известные нам прошлые значения временного ряда, либо прогнозные значения в предыдущие периоды времени.

Пример. Рассмотрим модель авторегрессии первого порядка

$$y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t.$$

Тогда прогноз на один шаг имеет вид

$$\widehat{y}_{n+1} = \beta_0 + \beta_1 y_n.$$

Прогноз на $l > 1$ шагов вычисляется рекурсивно

$$\widehat{y}_{n+l} = \beta_0 + \beta_1 \widehat{y}_{n+l-1}.$$

Пример. Рассмотрим модель второго порядка

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \varepsilon_t.$$

Прогноз на один шаг определяется как

$$\widehat{y}_{n+1} = \beta_0 + \beta_1 y_n + \beta_2 y_{n-1}.$$

Прогноз на два шага вычисляется по формуле

$$\widehat{y}_{n+2} = \beta_0 + \beta_1 \widehat{y}_{n+1} + \beta_2 y_n.$$

Прогноз на три и более шагов определяется рекурсивно

$$\widehat{y}_{n+l} = \beta_0 + \beta_1 \widehat{y}_{n+l-} + \beta_2 \widehat{y}_{n+l-2}, \quad l > 2.$$

Пример. Для модели третьего порядка

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_3 y_{t-3} + \varepsilon_t$$

прогноз на один шаг вычисляется по формуле

$$\widehat{y}_{n+1} = \beta_0 + \beta_1 y_n + \beta_2 y_{n-1} + \beta_3 y_{n-2}$$

Прогноз на два шага выглядит следующим образом

$$\widehat{y}_{n+2} = \beta_0 + \beta_1 \widehat{y}_{n+1} + \beta_2 y_n + \beta_3 y_{n-1},$$

на три шага:

$$\widehat{y}_{n+3} = \beta_0 + \beta_1 \widehat{y}_{n+2} + \beta_2 \widehat{y}_{n+1} + \beta_3 y_n$$

Прогноз на большее число шагов вычисляется рекурсивно

$$\widehat{y}_{n+l} = \beta_0 + \beta_1 \widehat{y}_{n+l-1} + \beta_2 \widehat{y}_{n+l-2} + \beta_3 \widehat{y}_{n+l-3}, \quad (l > 3).$$

Линейный прогноз стационарных авторегрессионных временных рядов обладает следующим важным свойством

Теорема. *В модели авторегрессии существует предел*

$$\widehat{y}_{n+l} \xrightarrow[\mathsf{P}]{} \mathsf{M} y_t \ (l \to +\infty)$$

Таким образом, при прогнозировании на «много» шагов вперед модель дает прогноз, мало отличающийся от среднего значения временного ряда (тривиальный прогноз). Другими словами, модель авторегрессии дает нетривиальный содержательный результат только при краткосрочных прогнозах.

4.4.4. Эконометрические методы исследования стационарных временных рядов

Перейдем теперь к рассмотрению эконометрических методов исследования авторегрессионных моделей стационарных временных рядов. Выделим три основных задачи

- Идентификация модели: оценивание порядка модели p и «грубое» оценивание коэффициентов модели.
- Оценивание модели: уточнение параметров модели.
- Диагностика модели: проверка адекватности и соответствия основным предположениям регрессионной модели (условиям Гаусса – Маркова).

Автоковариационная функция временного ряда оценивается по обычным формулам выборочного коэффициента ковариации (как обычно, n – объем выборки)

$$\hat{\gamma}(h) = \frac{1}{n-h} \sum_{t=1}^{n-h} (y_t - \bar{y})\,(y_{t+h} - \bar{y})\,.$$

Выборочная автокорреляционная функция ACF вычисляется по формуле

$$r(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}$$

В большинстве эконометрических программных пакетах (Eviews, SPSS и др.) функция ACF до заданного порядка вычисляются автоматически.

Рассмотрим вначале задачу оценивания параметров модели авторегрессии, считая известным ее порядок p.

Теорема. *Если инновационная последовательность ε_t удовлетворяет условиям теоремы Гаусса-Маркова, то* ***асимптотически*** *наилучшими линейными (BLUE) оценки коэффициентов $\mu, \beta_1, \ldots, \beta_p$ в уравнении авторегрессии будут OLS-оценки.*

Оценки коэффициентов $\beta_1, \ldots, \beta_p$ можно получить как решения системы уравнений Юла–Уолкера

$$\begin{cases} r(1) &= \beta_1 r(0) + \beta_2 r(1) + \cdots + \beta_p r(p-1) \\ r(2) &= \beta_1 r(1) + \beta_2 r(0) + \cdots + \beta_p r(p-2) \\ & \vdots \\ r(p) &= \beta_1 r(p-1) + \beta_2 r(p-2) + \cdots + \beta_p r(0) \end{cases}$$

а оценка коэффициента μ равна

$$m = \bar{y}\left(1 - \hat{\beta}_1 - \cdots - \hat{\beta}_p\right)$$

Пример. Для модели авторегрессии первого порядка получаем

$$\hat{\beta}_1 = r(1), \quad m = \bar{y}\big(1 - \hat{\beta}_1\big) = \bar{y}\big(1 - r(1)\big)$$

Следовательно, модель первого порядка будет задаваться авторегрессионным уравнением

$$\hat{y}_t = \bar{y}\big(1 - r(1)\big) + r(1) y_{t-1}.$$

Пример. Для временного ряда были вычислены $\bar{y} = 1.3$ и $r(1) = -0.7$. Тогда уравнение авторегрессии для модели первого порядка будет следующим

$$\hat{y}_t = 1.3\big(1 - (-0.7)\big) + (-0.7)y_{t-1} = 2.21 - 0.7y_{t-1}$$

Пример. Для модели второго порядка (с учетом $r(0) = 1$) система уравнений Юла-Уолкера имеет вид

$$\begin{cases} \beta_1 + r(1)\beta_2 &= r(1) \\ r(1)\beta_1 + \beta_2 &= r(2) \end{cases}$$

Решая ее получаем

$$\hat{\beta}_1 = \frac{r(1) - r(1)r(2)}{1 - r^2(1)}, \quad \hat{\beta}_2 = \frac{r(2) - r^2(1)}{1 - r^2(1)}.$$

Модель второго порядка будет задаваться авторегрессионным уравнением

$$\hat{y}_t = \bar{y}\left(1 - \hat{\beta}_1 - \hat{\beta}_2\right) + \hat{\beta}_1 y_{t-1} + \hat{\beta}_2 y_{t-2}.$$

Пример. Для временного ряда были вычислены $\bar{y} = 2.1$, $r(1) = 0.6$ и $r(2) = -0.2$. Коэффициенты для модели второго порядка равны

$$\hat{\beta}_1 = \frac{0.6 - 0.6(-0.2)}{1 - 0.6^2} = 1.125, \quad \hat{\beta}_2 = \frac{-0.2 - 0.6^2}{1 - 0.6^2} = -0.875$$

Модель второго порядка будет задаваться уравнением

$$\hat{y}_t = 2.1\big(1 - 1.125 - (-0.875)\big) + 1.125y_{t-1} + (-0.875)y_{t-2}$$

Формулы для вычисления прогнозируемых значений получаются из формул прогнозирования для стационарного временного ряда с заменой коэффициентов β_k и μ на их OLS-оценки:

$$\widehat{y}_{n+1} = m + \hat{\beta}_1 y_n + \hat{\beta}_2 y_{n-1} + \cdots + \hat{\beta}_p y_{n-p+1}$$
$$\widehat{y}_{n+2} = m + \hat{\beta}_1 \widehat{y}_{n+1} + \hat{\beta}_2 y_n + \cdots + \hat{\beta}_p y_{n-p+2}$$

К авторегрессионной модели временных рядов асимптотически применимы все методы регрессионного анализа (проверка значимости коэффициентов модели и модели в целом и т.д.), кроме теста Durbin – Watson (т.к. модель содержит лаговые значения зависимой переменной).

Идентификация модели

Перейдем теперь к задаче идентификации модели, т.е. к задаче оценивания порядка авторегрессии. Наиболее часто используется подход, основанный на идее информационных критериев. Он также применим к более общим моделям стационарных временных рядов (моделям скользящего среднего MA и моделям ARMA).

Процедура выбора модели на основе информационных критериев состоит в следующем: оцениваются несколько моделей авторегрессии разных порядков и для каждой модели считается числовой показатель информационного критерия. Выбирается та модель, для которой этот числовой показатель минимален.

Информационный критерий Акаике (Akaike, 1973) порядок авторегрессии p выбирается из условия

$$AIC(m) = \ln s^2(m) + \frac{2m}{n} \longrightarrow \min .$$

Здесь $s^2(m)$ - оценка дисперсии ошибок в модели AR(m). Исторически это первый информационный критерия. Его основной недостаток состоит в том, что при больших объемах выборочных данных критерий может завышать порядок авторегрессии

Информационный критерий Шварца (Shwarz [25] 1978) порядок авторегрессии p выбирается из условия

$$SIC(m) = \ln s^2(m) + \frac{m \ln n}{n} \longrightarrow \min .$$

Информационный критерий Хеннана-Куина (Hannan, Quinn [16], 1979) порядок авторегрессии p выбирается из условия

$$HQ(m) = \ln s^2(m) + 2c \cdot \frac{m \ln(\ln n)}{n} \longrightarrow \min, \quad c > 1.$$

Это критерий может недооценивать порядок p при небольших объемах выборки

Пример. По временному ряду объема $n = 100$ были оценены авторегрессионные модели до четвертого порядка и для них получены следующие оценки дисперсий ошибок:

$$s^2(1) = 0.9,\ s^2(2) = 0.7,\ s^2(3) = 0.5,\ s^2(4) = 0.46.$$

Выберем порядок модели авторегрессии на основе информационного критерия Shwarz. Вычислим показатель SIC для моделей авторегрессии до четвертого порядка:

$$SIC(1) = \ln s^2(1) + \frac{1 \cdot \ln n}{n} = \ln 0.9 + \frac{1 \cdot \ln 100}{100} \approx -0.059$$
$$SIC(2) = \ln s^2(2) + \frac{2 \cdot \ln n}{n} = \ln 0.7 + \frac{2 \cdot \ln 100}{100} \approx -0.265$$
$$SIC(3) = \ln s^2(3) + \frac{3 \cdot \ln n}{n} = \ln 0.5 + \frac{3 \cdot \ln 100}{100} \approx -0.555$$
$$SIC(4) = \ln s^2(4) + \frac{4 \cdot \ln n}{n} = \ln 0.46 + \frac{4 \cdot \ln 100}{100} \approx -0.592$$

Следовательно, необходимо сделать выбор в пользу модели авторегрессии четвертого порядка, так как значение информационного критерия для нее минимально

Пример. В условиях предыдущего примера произведем выбор модели с помощью информационного критерия Акаике. Вычислим показатель AIC для моделей авторегрессии до четвертого порядка:

$$AIC(1) = \ln s^2(1) + \frac{2 \cdot 1}{n} = \ln 0.9 + \frac{2}{100} \approx -0.085$$
$$AIC(2) = \ln s^2(2) + \frac{2 \cdot 2}{n} = \ln 0.7 + \frac{4}{100} \approx -0.317$$
$$AIC(3) = \ln s^2(3) + \frac{2 \cdot 3}{n} = \ln 0.5 + \frac{6}{100} \approx -0.633$$
$$AIC(4) = \ln s^2(4) + \frac{2 \cdot 4}{n} = \ln 0.46 + \frac{8}{100} \approx -0.697$$

Следовательно, необходимо выбрать модель авторегрессии четвертого порядка, так как для нее значение информационного критерия минимально.

Проверка адекватности модели

Проверка адекватности, т.е. проверка согласованности выбранной и оцененной модели с наблюдениями, как и в регрессионном анализе,

основано на исследовании остатков. А именно, остатки должны моделировать процесс нормально распределенного белого шума.

Так как модель содержит лаговые значения зависимой переменной, то критерий Durbin–Watson для исследования ошибок на автокорреляцию неприменим. Статистика DW будет смещенной в сторону уменьшения.

Приведем метод исследования, предложенный Боксом и Льюнгом (Box, Ljung [5]) и основанный на применении Q-статистик. Пусть $r_e(h)$ – выборочные коэффициенты автокорреляции ряда остатков e_t в модели AR(p). Q-статистики определяются равенствами

$$Q = n(n+2)\sum_{h=1}^{M}\frac{r_e^2(h)}{n-h}, \quad M = 1, 2, \ldots$$

При выполнении условия Гаусса – Маркова на ошибки модели авторегрессии Q-статистики **асимптотически** имеют распределение χ^2_{M-p}. Следовательно, при заданном уровне значимости (и при большом объеме выборки) гипотеза о независимости и одинаковой распределенности ошибок регрессии отвергается при $Q > \chi^2_{\text{кр}}$, где критическое значение $\chi^2_{\text{кр}} = \chi^2(\alpha; M-p)$.

Пример. Для временного ряда длины $n = 100$ была оценена модель второго порядка ($p = 2$) и вычислены коэффициенты автокорреляции остатков

$$r_e(1) = 0.001\ r_e(2) = 0.001\ r_e(3) = 0.0006$$
$$r_e(4) = 0.0004\ r_e(5) = 0.0003.$$

Проверим адекватность модели по критерию Бокса – Льюнга. Вычислим Q-статистику с $M = 5$:

$$Q = n(n+2)\left(\frac{r_e^2(1)}{n-1} + \frac{r_e^2(2)}{n-2} + \frac{r_e^2(3)}{n-3} + \frac{r_e^2(4)}{n-4} + \frac{r_e^2(5)}{n-5}\right) =$$
$$100 \cdot 102 \cdot \left(\frac{0.001^2}{100-1} + \frac{0.001^2}{100-2} + \frac{0.0006^2}{100-3} + \frac{0.0004^2}{100-4} + \frac{0.0003^2}{100-5}\right) \approx$$
$$\approx 0.00027$$

Критическое значение распределения хи-квадрат с $M - p = 5 - 2 = 3$ степенями свободы при 5%-м уровне значимости равно

$$\chi^2_{\text{кр}} = \chi^2(5\%; 3) \approx 7.875.$$

Так как $Q < \chi^2_{\text{кр}}$, то данные согласованы с условиями Гаусса – Маркова для модели AR(2).

4.5. Динамические модели стационарных временных рядов

Динамические модели ADL (autoregression distributed lags) содержат лаговые значение зависимой переменной (а также, возможно, и регрессоров):

$$y_t = \mu + \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + \beta_1 z_{t1} + \cdots + \beta_k z_{tk} + u_t. \qquad (4.3)$$

1. Как отмечалось выше, для таких моделей нарушаются условия на ошибки регрессии. Однако, статистические выводы многофакторной модели регрессии будут верны *асимптотически*, т.е. при больших выборках.

Обозначим

$$\mathbf{z}_t = \begin{pmatrix} z_{t1} \\ \vdots \\ z_{tk} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

и запишем модель в виде

$$y_t = \mu + \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + \mathbf{z}_t'\beta + u_t$$

Теорема. *Пусть выполнены следующие условия*

- $\mathbf{z}_t$ – *(многомерный) стационарный временной ряд;*
- u_t – *процесс белого шума,* $\mathsf{M}u_t = 0$, $\mathrm{Var}(u_t) = \sigma^2$*;*
- $\mathrm{cov}(u_t, z_{tj}) = 0$, $j = 1 \ldots, k$*;*
- $\mathrm{cov}(u_t, y_{t-s}) = 0$, $s = 1, \ldots, p$*;*
- ***все*** *корни многочлена* $\alpha(z) = 1 - \alpha_1 z - \cdots - \alpha_p z^p$ *по модулю больше единицы (условие стационарности);*
- *существует предел (условие эргодичности)*

$$\frac{1}{n}\mathbf{z}_t\mathbf{z}_t' \xrightarrow[\mathsf{P}]{} \Sigma, \quad \det\Sigma \neq 0$$

Тогда OLS-оценки коэффициентов модели регрессии (4.3) ***асимптотически*** *нормальны и имеют минимальную дисперсию.*

Таким образом, при выполнении условие теоремы для больших выборок, мы можем пользоваться стандартными статистическими процедурами.

Замечание. Так как модель регрессии содержит лаговые значения зависимой переменной, то статистика DW будет смещена и тест Durbin – Watson к данной модели **неприменим**.

2. Рассмотрим подробнее частный случай ADL с одним регрессором

$$y_t = \mu + \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + +\beta_0 x_t + \cdots + \beta_q x_{t-q} + u_t$$

и пусть выполнены все условия теоремы. Тогда, аналогично модели FDL, между факторами устанавливается «долговременная связь», описываемая соотношением

$$y^* = \delta_0 + \delta_1 x^*,$$
$$\delta_0 = \frac{\mu}{1 - \alpha_1 - \cdots - \alpha_p} = \frac{\mu}{\alpha(1)}, \quad \delta_1 = \frac{\beta_0 + \cdots + \beta_q}{1 - \alpha_1 - \cdots - \alpha_p}$$

Коэффициент δ_1 называется *долгосрочным мультипликатором* (long-time multiplier).

4.6. Задачи

Упражнение 1. Для следующих моделей FDL найдите долгосрочные мультипликаторы и напишите уравнение «долгосрочной связи»

1. $\widehat{y}_t = \beta_0 + \beta_1 x_t + \beta_2 x_{t-1} + \beta_3 x_{t-2} + \beta_4 x_{t-3}$;
2. $\widehat{y}_t = \beta_0 + \beta_1 x_t + \beta_2 x_{t-1} + \beta_3 z_t + \beta_4 z_{t-1}$;
3. $\widehat{y}_t = \beta_0 + \beta_1 x_t + \beta_2 x_{t-1} + \beta_3 x_{t-2} + \beta_4 z_t + \beta_5 z_{t-1} + \beta_6 z_{t-2}$;
4. $\widehat{y}_t = \beta_0 + \beta_1 x_t + \beta_2 z_t + \beta_3 z_{t-1} + \beta_4 z_{t-2} + \beta_5 w_t + \beta_6 w_{t-1}$.

Упражнение 2. Является временной ряд, заданный авторегрессионным разностным уравнением, стационарным?

1. $y_t = 7 + 0.5y_{t-1} + \varepsilon_t$.
2. $y_t = 10 + 0.25y_{t-2} + \varepsilon_t$.
3. $y_t = 10 + y_{t-1} - 0.25y_{t-2} + \varepsilon_t$.
4. $y_t = \frac{3}{2}y_{t-1} - \frac{3}{4}y_{t-2} + \frac{1}{8}y_{t-3} + \varepsilon_t$.
5. $y_t = 3 + 0.4y_{t-1} - 0.04y_{t-2} + \varepsilon_t$.
6. $y_t = 5 - 3y_{t-1} - 3y_{t-2} - y_{t-3} + \varepsilon_t$.
7. $y_t = -2y_{t-1} + 1.25y_{t-2} - 0.25y_{t-3} + \varepsilon_t$.

Упражнение 3. Написать формулы для прогноза

1. на $l = 3$ шага для процесса $y_t = 5 + 0.5y_{t-1} + \varepsilon_t$.
2. на $l = 4$ шага для процесса $y_t = 2 + 0.25y_{t-2} + \varepsilon_t$.
3. на $l = 3$ шага для процесса $y_t = \frac{1}{27}y_{t-3} + \varepsilon_t$.
4. на $l = 5$ шагов для процесса $y_t = 0.5y_{t-2} - 0.05y_{t-3} + 0.001y_{t-4} + \varepsilon_t$.
5. на $l = 6$ шагов для процесса $y_t = 0.5y_{t-3} + 0.001y_{t-4} + \varepsilon_t$.

Упражнение 4. Написать в общем виде уравнения Юла – Уолкера для моделей AR(2), AR(3) и AR(4).

Упражнение 5. По временному ряду длины $n = 60$ были оценены следующие авторегрессионные модели:

1. $\hat{y}_t = 2 + 0.7y_{t-1}$, $s^2 = 2.1$;
2. $\hat{y}_t = 2.3 + 0.6y_{t-1} - 0.3y_{t-2}$, $s^2 = 1.9$;
3. $\hat{y}_t = 1.8 + 0.55y_{t-1} - 0.25y_{t-2} + 0.01y_{t-3}$, $s^2 = 1,85$.

Какую модель вы выберете?

Упражнение 6. Для временного ряда были вычислены коэффициенты автокорреляции

$$\rho(1) = 0.7; \quad \rho(2) = 0.4; \quad \rho(3) = -0.2$$

и выборочное среднее значение $\bar{y} = 1.7$. Найти оценки коэффициентов в модели

1. AR(1);

2. AR(2);

3. AR(3).

Упражнение 7. Для модели временного ряда длины $n = 50$ были оценены несколько моделей и в каждой из моделей вычислены коэффициенты автокорреляции остатков. Исследовать адекватность этих моделей

1. $r_e(1) = 0.001$; $r_e(2) = 0.0006$; $r_e(3) = 0.0002$; $r_e(4) = 0.001$ в модели AR(2).

2. $r_e(1) = 0.04$; $r_e(2) = 0.02$; $r_e(3) = 0.006$ в модели AR(1).

3. $r_e(1) = 0.02$; $r_e(2) = 0.0008$; $r_e(3) = 0.003$; $r_e(4) = 0.001$ в модели AR(3).

Приложение А

Статистические таблицы

Приведем список функции MS Excel и OpenOffice для вычисления критических значений стандартных распределений с уровнем значимости α

Распределение	MS Excel (Рус)	MS Excel (Eng) OpenOffice
Двустороннее $\mathcal{N}(0,1)$	НОРМСТОБР$(1-\alpha/2)$	NORMSINV$(1-\alpha/2)$
Одностороннее $\mathcal{N}(0,1)$	НОРМСТОБР$(1-\alpha)$	NORMSINV$(1-\alpha)$
χ^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 α								
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 χ^2_k

	Уровень значимости α						
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 α					
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
∞	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
∞	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
∞	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
∞	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
∞	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
∞	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
∞	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

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» ([21], 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).

Литература

[1] Берндт Э., *Практика эконометрики. Классика и современность*, М.: Юнити, 2005.

[2] Вербик М., *Путеводитель по современной эконометрике*, М.: Научна книга, 2008.

[3] Магнус Я.Р., Катышев П.К., Пересецкий А.А., *Эконометрика. Начальный курс*, М.: Дело, 2007.

[4] Носко В. П., *Эконометрика для начинающих*, М.: Изд. ИЭПП, 2000

[5] G. E. Box, G. M. Ljung, *On a measure of lack of fit in time series models*, Biometrica, vol. 65, No 2, pp. 297–303, 1978

[6] T. S. Breusch, A. R. Pagan, *A simple test for heteroscedasticity and random coefficient variation* Econometrica, vol. 47, No 5, pp. 1287–1294, 1979.

[7] 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

[8] 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

[9] R. Davidson, J. G. MacKinnon, *Econometric theory and methods*, Oxford University Press, USA, 2003

[10] 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

[11] J. Durbin, G. S. Watson, *Testing for serial correlation in least squares regression. I*, Biometrica, vol. 37, No 3/4, pp. 409–428, 1950

[12] J. Durbin, G. S. Watson, *Testing for serial correlation in least squares regression. II*, Biometrica, vol. 38, No 1/2, pp. 159–177, 1951

[13] W. Enders, *Applied Econometric Time Series*, John Wiley & Sons; 2nd edition, 2004

[14] R. Frisch, *Editor's note*, Econometrica, vol. 1, pp. 1–4, 1933

[15] J. D. Hamilton, *Time series analysis*, Princeton University Press, 1994

[16] E. J. Hannan, B. G. Quinn, *The Determination of the order of an autoregression*, J. of the Royal Statistic Society. Series B, v. 41, No 2, pp.190–195, 1979

[17] S. M. Goldfeld, R. M. Quandt, *Some tests for homoscedasticity*, Journal of the American Statistical Association, Vol. 60, No.310, pp. 539-547, 1965

[18] 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

[19] 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

[20] 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

[21] W. J. Greene, *Econometric Analysis*, Prentice Hall, 6th edition, 2007

[22] G. S. Maddala, *Introduction to Econometrics*, Second Edition, Macmillian Publishing, 1992

[23] 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

[24] 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

[25] G. Shwarz, *Estimating the dimension of a model*, Ann. Statist, 6, pp.461–464, 1978

[26] J. H. Stock, M. W. Watson, *Introduction to Econometrics*, Addison Wesley, 2nd edition, 2006

[27] G. Tintner, *The Definition of Econometrics*, Econometrica, vol. 21, No. 1, pp. 31-40, 1953

[28] H. White, *A heteroskedasticity-consisten covariance matrix estimator and a direct test for heteroskedasticity*, Econometrica, v. 48, No 4, pp.817–838, 1980.

[29] J. M. Wooldridge, *Introductory Econometrics. A modern approach*, Forth Edition. 2009

[30] J. M. Wooldridge, *Econometric analysis of cross section and panel data*, The MIT Press, 2001.