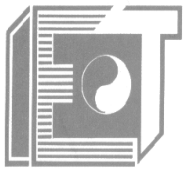


ИНСТИТУТ ЭКОНОМИКИ
ПЕРЕХОДНОГО ПЕРИОДА

В.П. Носко

Эконометрика для начинающих

Основные понятия, элементарные методы,
границы применимости,
интерпретация результатов

Москва
2000

**ИНСТИТУТ ЭКОНОМИКИ
ПЕРЕХОДНОГО ПЕРИОДА**

В.П. Носко

Эконометрика для начинающих

**Основные понятия, элементарные методы,
границы применимости,
интерпретация результатов**

Москва
2000

Институт экономики переходного периода
Основан в 1992 г.
Учредители: Академия народного хозяйства при Правительстве РФ

Директор: Е.Т.Гайдар

Носко Владимир Петрович - кандидат физико-математических наук, старший научный сотрудник механико-математического факультета Московского государственного университета им. М.В.Ломоносова. Автор более 40 научных работ, соавтор учебного пособия "Основные понятия и задачи математической статистики".
Преподает эконометрику с 1994 года. В настоящее время читает курсы лекций по эконометрике на механико-математическом факультете МГУ, на факультете менеджмента Международного университета (г. Москва) и в Институте экономики переходного периода.

Настоящая работа издана на средства гранта, предоставленного Институту экономики переходного периода Агентством США по международному развитию

Компьютерный дизайн: А. Астахов

ISBN 5-93255-027-9

Лицензия на издательскую деятельность ***Серия ИД № 02079 от 19 июня 2000 г.***
103918, Москва, Газетный пер., 5
Тел. (095) 229–6413, FAX (095) 203–8816
E-MAIL – root@iet.ru, **WEB** Site – http://www.iet.ru

ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ

Предлагаемое учебное пособие имеет своей целью обеспечить базу для изучения вводного полугодового курса эконометрики, когда в распоряжении преподавателя имеется всего порядка 12 лекций и некоторое количество часов практических занятий. При этом от читателя не требуется никаких предварительных знаний из теории вероятностей и математической статистики. Что касается математического анализа и линейной алгебры, то желательно, чтобы читатель имел хотя бы некоторое представление о производной и интеграле, а также о матрицах и операциях над ними. Соответственно, акценты в изложении смещаются в сторону разъяснения базовых понятий и основных процедур статистического анализа данных с привлечением большого количества иллюстративных примеров. В этом отношении данное учебное пособие близко по духу к имеющейся в русском переводе книге К. Доугерти «Введение в эконометрику» (1997), которая предназначена для изучения годового курса эконометрики и которую можно рекомендовать для последующего изучения вопросов, не охваченных в рамках настоящего пособия.

С целью постепенного введения студентов в круг понятий и методов эконометрики, в первой части пособия вообще не используются понятия теории вероятностей и математической статистики. И только когда дальнейшее игнорирование этих понятий в процессе анализа данных становится попросту невозможным, дается необходимый минимум сведений из этих дисциплин. Вторая часть пособия посвящена построению и статистическому анализу линейных регрессионных моделей при классических предположениях о модели наблюдений. В третьей части рассматриваются графические и формальные статистические методы выявления ряда нарушений классических предположений и методы коррекции статистических выводов при обнаружении таких нарушений.

Пособие написано на основании курса лекций, который читался автором на протяжении ряда лет в Международном университете (г.

Москва), и лекций для аспирантов Института экономических проблем переходного периода.

ЧАСТЬ 1. ОЦЕНИВАНИЕ И ПОДБОР МОДЕЛЕЙ СВЯЗИ МЕЖДУ ПЕРЕМЕННЫМИ БЕЗ ПРИВЛЕЧЕНИЯ ВЕРОЯТНОСТНО-СТАТИСТИЧЕСКИХ МЕТОДОВ

1.1. ЭКОНОМЕТРИКА И ЕЕ СВЯЗЬ С ЭКОНОМИЧЕСКОЙ ТЕОРИЕЙ

Эконометрика (Econometrics) - 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.

Пусть, например, мы имеем данные о размерах ***располагаемого дохода (disposable personal income) DPI*** и ***расходов на личное потребление* (*personal consumption*) *C*** для n семейных хозяйств, так что DPI_i и C_i, соответственно, представляют располагаемый доход и расходы на личное потребление i -го семейного хозяйства.

Простейшей моделью связи между DPI и C является ***линейная модель связи***

$$\boxed{C = \alpha + \beta \cdot DPI,}$$

где β - некоторая постоянная величина , $0 < \beta < 1$, характеризующая в данном круге семейных хозяйств их ***склонность к потреблению***, связанную с традициями и привычками, а α - *"**автономное потребление**"*.

Однако, если разместить на плоскости в прямоугольной системе координат точки $\left(DPI_i,\ C_i\right)$ с абсциссами DPI_i и ординатами C_i (такое расположение точек называется ***диаграммой рассеяния - scatterplot***), то, как правило, эти точки вовсе не будут лежать на одной прямой вида $C = \alpha + \beta \cdot DPI$, соответствующей линейной модели связи. Вместо этого, они будут образовывать ***облако рассеяния***, вытянутое в некотором направлении (см. Рис.1.1). В таком случае соотношение между DPI_i и C_i принимает форму

$$\boxed{C_i = \left(\alpha + \beta \cdot DPI_i\right) + \varepsilon_i,\ i = 1,\text{K}\ ,n}$$

(модель наблюдений), где слагаемое

$$\boxed{\varepsilon_i = C_i - \left(\alpha + \beta \cdot DPI_i\right)}$$

представляет *отклонение* реально наблюдаемых расходов на потребление C_i от значения $\alpha + \beta \cdot DPI_i$, предсказываемого гипотетической линейной моделью связи для i -го семейного хозяйства. Эти отклонения отражают совокупное влияние на конкретные значения C_i множества дополнительных факторов, не учитываемых принятой моделью связи.

РИС. 1.1

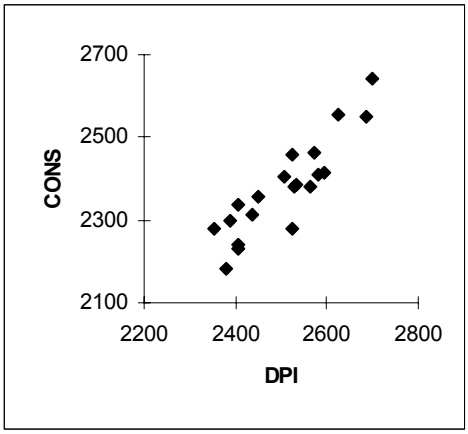


Диаграмма рассеяния на рис.1.1 соответствует данным о годовом располагаемом доходе и годовых расходах на личное потребление (в 1999 г., в условных единицах) 20 семей . Эти данные представлены в таблице 1.1.

ТАБЛ. 1.1

i	DPI	C
1	2508	2406
2	2572	2464
3	2408	2336
4	2522	2281
5	2700	2641
6	2531	2385
7	2390	2297
8	2595	2416
9	2524	2460
10	2685	2549

I	DPI	C
11	2435	2311
12	2354	2278
13	2404	2240
14	2381	2183
15	2581	2408
16	2529	2379
17	2562	2378
18	2624	2554
19	2407	2232
20	2448	2356

Предложив для описания имеющихся статистических данных модель, учитывающую указанные отклонения от теоретической модели линейной связи между DPI_i и C_i (***модель наблюдений)***, мы неизбежно сталкиваемся с вопросом о том,

каковы значения α и β в этой модели. И с этого момента попадаем в поле деятельности ***эконометрики***, предлагающей различные ***методы оценивания параметров*** экономических моделей по имеющимся статистическим данным, а также методы использования оцененной модели для целей экономического ***прогнозирования*** и проведения рациональной экономической политики. Кроме того, методы эконометрики дают возможность ***подбора подходящей модели***, адекватной имеющимся данным, в ситуации, когда в распоряжении исследователя нет ясной экономической теории, описывающей поведение интересующих его отдельных экономических показателей и связи между различными показателями.

1.2. ДВЕ ПЕРЕМЕННЫЕ: МЕРЫ ИЗМЕНЧИВОСТИ И СВЯЗИ

В приводимой ниже таблице 1.2 указаны уровни безработицы (в %) среди белого и цветного населения США в период с марта 1968 г. по июль 1969 г. (месячные данные). В первом столбце расположены номера последовательных наблюдений ($i = 1$ для марта 1968 г., i =17 для июля 1969 г.), во втором столбце - значения BEL_i уровня безработицы среди белого населения в i -ом месяце, а в третьем - значения $ZVET_i$ уровня безработицы среди цветного населения в i -ом месяце.

ТАБЛ. 1.2

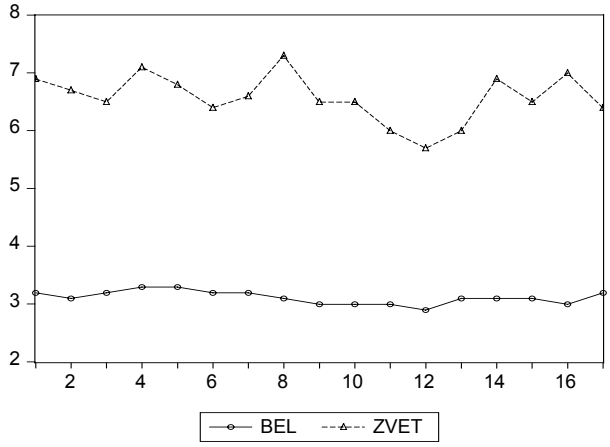
i	BEL	ZVET	i	BEL	ZVET
1	3.2	6.9	**10**	3.0	6.5
2	3.1	6.7	**11**	3.0	6.0
3	3.2	6.5	**12**	2.9	5.7
4	3.3	7.1	**13**	3.1	6.0
5	3.3	6.8	**14**	3.1	6.9
6	3.2	6.4	**15**	3.1	6.5
7	3.2	6.6	**16**	3.0	7.0

8	3.1	7.3	**17**	3.2	6.4
9	3.0	6.5			

Рассмотрим, прежде всего, графики изменения уровней безработицы в обеих группах в течение указанного периода времени (Рис. 1.2).

Первое впечатление от просмотра этих графиков - уровень безработицы среди цветного населения существенно выше и изменяется со временем со значительными колебаниями; уровень безработицы среди белого населения изменяется плавно и в довольно узком диапазоне.

РИС. 1.2



Для того, чтобы использовать обозначения, соответствующие общепринятой практике, мы обозначим через $x_1, x_2, \text{K}, x_{17}$ последовательно наблюдаемые уровни безработицы среди цветного населения, а через $y_1, y_2, \text{K}, y_{17}$ - соответствующие им уровни безработицы среди белого населения США, так что мы можем говорить о наблюдаемых значениях двух переменных: переменной x - уровня безработицы среди цветного на-

селения, и переменной y - уровня безработицы среди белого населения.

Наиболее простыми показателями, характеризующими последовательности $x_1, x_2, \text{K}, x_{17}$ и $y_1, y_2, \text{K}, y_{17}$, являются их ***средние значения (means)***

$$\bar{x} = \frac{1}{17}\sum_{i=1}^{17} x_i = \frac{x_1 + x_2 + \text{K} + x_{17}}{17}, \qquad \bar{y} = \frac{1}{17}\sum_{i=1}^{17} y_i = \frac{y_1 + y_2 + \text{K} + y_{17}}{17},$$

а также ***дисперсии*** (точнее, ***выборочные дисперсии - sample variances***)

$$Var(x) = \frac{1}{n-1}\sum_{i=1}^{17}\left(x_i - \bar{x}\right)^2, \qquad Var(y) = \frac{1}{n-1}\sum_{i=1}^{17}\left(y_i - \bar{y}\right)^2,$$

характеризующие *степень разброса* значений $x_1, x_2, \text{K}, x_{17}$ ($y_1, y_2, \text{K}, y_{17}$) вокруг своего среднего $\bar{x}$ ($\bar{y}$, соответственно), или ***вариабельность*** **(*изменчивость*)** этих переменных на множестве наблюдений. Отсюда обозначение ***Var*** **(*variance*)**. Впрочем, более естественным было бы измерение степени разброса значений переменных в тех же единицах, в которых измеряется и сама переменная. Эту задачу решает показатель, называемый ***стандартным отклонением*** **(*standard deviance - Std.Dev.*)** переменной x (переменной y), определяемый соотношением

$$\text{Std.Dev.(x)} = \sqrt{\text{Var(x)}},$$

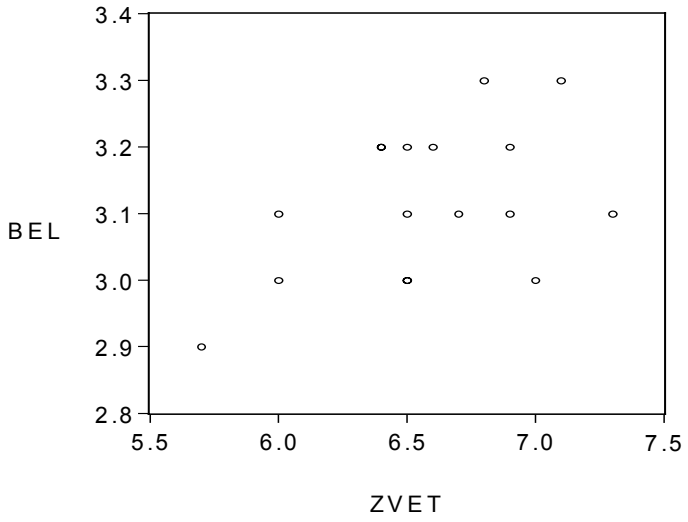
($\text{Std.Dev.(y)} = \sqrt{\text{Var(y)}}$, соответственно).

Вычисления по указанным формулам приводят к значениям $\bar{x} = 6.576$, $\text{Std.Dev.(x)} = 0.416$; $\bar{\text{y}} = 3.118$, $\text{Std.Dev.(y)} = 0.113$. Иными словами, уровень безработицы среди цветного населения, в среднем, более, чем в два раза превышает уровень безработицы среди белого населения. Стандартные отклонения, соответственно, относятся приблизительно как 4:1, что указы-

вает на гораздо более сильную изменчивость (“вариабельность”) уровня безработицы среди цветного населения. Размахи колебаний уровней равны, соответственно, 7.3 - 5.7 = 1.6 и 3.3 - 3.1 = 0.2.

Удобным графическим средством анализа данных является ***диаграмма рассеяния (scatterplot*)**, на которой в прямоугольной системе координат располагаются точки x_i, y_i, i = 1, 2, ..., n, где n - количество наблюдаемых пар значений переменных x и y. В нашем примере n = 17, и диаграмма рассеяния имеет вид

РИС. 1.3



Вытянутость облака точек на диаграмме рассеяния вдоль наклонной прямой позволяет сделать предположение о том, что существует некоторая объективная тенденция линейной связи между значениями переменных x и y, выражаемой соотношением

$y = \alpha + \beta \cdot x,$

где x — уровень безработицы среди цветного, а y — среди белого населения. В то же время, указанное соотношение выражает всего лишь тенденцию: реально наблюдаемые значения y_i отличаются от значений $y = \alpha + \beta \cdot x_i$, на величину

$$\boxed{\varepsilon_i = y_i - (\alpha + \beta \cdot x_i)}$$

так что

$$\boxed{y_i = (\alpha + \beta \cdot x_i) + \varepsilon_i,\ i = 1,\text{K}\ ,n.}$$

Последнее соотношение определяет ***линейную модель наблюдений,*** тогда как соотношение

$$\boxed{y = \alpha + \beta \cdot x}$$

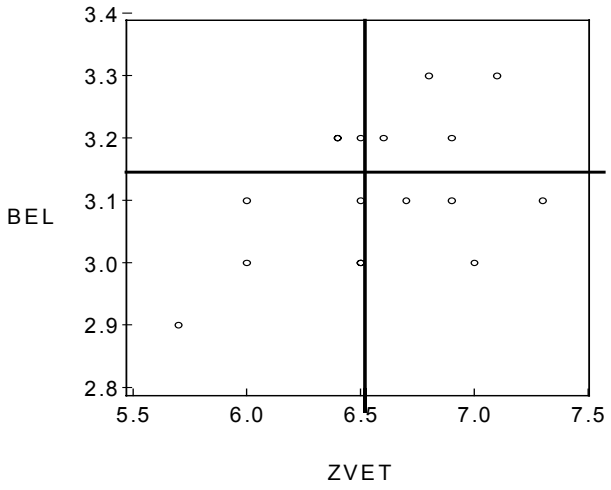
определяет ***линейную модель связи*** между рассматриваемыми переменными.

Заметим, однако, что видимая степень проявления вытянутости облака точек на диаграмме рассеяния (при наличии линейной связи между переменными) существенно зависит от выбора единиц измерения переменных x и y. Поэтому, во-первых, желательно при построении диаграммы выбирать масштабы и интервалы изменения переменных таким образом, чтобы диаграмма имела вид квадрата и чтобы на диаграмме имелись точки, достаточно близко расположенные к каждой из четырех границ квадрата. Во-вторых, желательно иметь какие-то числовые характеристики, которые отражали бы действительное наличие вытянутости облака точек вдоль наклонной прямой и не зависели от шкал, в которых представлены значения переменных.

Одна из характеристик такого рода связана с разбиением диаграммы рассеяния горизонтальной и вертикальной прямыми на 4 прямоугольника.

Разбивающие диаграмму прямые (секущие) проводятся через точку $(\bar{x},\bar{y})$, так что если точка (x_i,y_i) лежит правее вертикальной секущей, то отклонение $x_i-\bar{x}$ имеет знак плюс, а если левее, то знак минус. Аналогично, если точка (x_i,y_i) лежит выше горизонтальной секущей, то отклонение $y_i-\bar{y}$ имеет знак плюс, а если она расположена ниже этой секущей, то знак минус (см. Рис. 1.4).

РИС. 1.4



Пусть m_{++} — количество таких точек среди $(x_1,y_1),\text{K},(x_n,y_n)$, для которых $x_i-x>0$ и $y_i-y>0$ (верхний правый прямоугольник); m_{+-} — количество точек, для которых $x_i-x>0$ и $y_i-y<0$ (нижний правый прямоугольник); m_{-+} — количество точек, для которых $x_i-x<0$ и $y_i-y>0$ (верхний левый прямоугольник); m_{--}- количество точек, для которых $x_i-x<0$ и $y_i-y<0$ (нижний левый прямоуголь-

ник). В нашем примере, $m_{++} = 4$, $m_{+-} = 4$, $m_{-+} = 3$ (точки, соответствующие наблюдениям с номерами 6 и 17, имеют совпадающие координаты), $m_{--} = 6$ (точки, соответствующие наблюдениям с номерами 9 и 10, имеют совпадающие координаты), так что количество точек с совпадающими знаками отклонений $x_i - x$ и $y_i - y$ равно $m_{++} + m_{--} = 10$, а количество точек, у которых знаки отклонений различны, равно $m_{+-} + m_{-+} = 7$.

Количество точек с совпадающими знаками отклонений от средних значений составляет 10/17=0.59, т. е. около 59% общего числа точек, и это служит некоторым указанием на наличие вытянутости облака точек в направлении прямой, имеющей положительный угловой коэффициент. Если бы большинство составляли точки с противоположными знаками отклонений от средних значений, то это служило бы объективным указанием на наличие вытянутости облака точек в направлении прямой, имеющей отрицательный угловой коэффициент. Последняя ситуация часто наблюдается при рассмотрении зависимости спроса на товар от его цены.

Более распространенным является определение степени выраженности линейной связи между произвольными переменными x и y, принимающими значения x_i и y_i, $i = 1,\text{K},n$, посредством **(*выборочного*) *коэффициента корреляции* (*sample correlation coefficient*)**

$$r_{xy} = \frac{Cov(x,y)}{\sqrt{Var(x)}\sqrt{Var(y)}}.$$

Величина $Cov(x,y)$, стоящая в числителе, определяется соотношением

$$Cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

и называется ***(выборочной) ковариацией*** переменных x и y, так что, формально,

$$\boxed{Cov(x,x) = Var(x),} \quad \boxed{Cov(y,y) = Var(y).}$$

Если указанная тенденция выражена на диаграмме рассеяния довольно ясно, то значения r_{xy} по абсолютной величине близки к единице (т. е. значения r_{xy} близки к +1 или к –1). Если же наличие линейной тенденции связи обнаруживается на диаграмме рассеяния с трудом, то тогда значения r_{xy} близки к нулю. Как мы увидим позднее, значения r_{xy} уже не зависят от выбора шкал измерения переменных x и y (если, конечно, эти шкалы линейны).

В нашем примере $Var(x) = 0.1732$, $Var(y) = 0.0128$, $Cov(x,y) = 0.0204$, откуда находим

$$r_{xy} = \frac{0.0204}{\sqrt{0.1732}\sqrt{0.0128}} = 0.4608,$$

т. е. получаем значение r_{xy}, расположенное приблизительно посередине между 0 и 1.

Замечание

Мы определили *Var* и *Cov,* деля соответствующие суммы квадратов на *n*-1. Это имеет свое объяснение, которое пока выходит за рамки нашего обсуждения. Вместе с тем, в разных руководствах по эконометрике *Var* и *Cov* определяются поразному. Деление на *n* - 1 используется, например, в книгах Доугерти (1997), Айвазяна и Мхитаряна (1998), тогда как в книге Магнуса, Катышева и Пересецкого (1997) соответствующие суммы квадратов делятся не на *n* - 1, а на *n.* К счастью, и *Cov* и *Var* будут играть у нас лишь вспомогательную роль, а величина более существенного для нас коэффициента корреляции r_{xy} не

зависит от того, каким из двух способов мы будем определять *Var* и *Cov,* лишь бы только при определении обеих этих характеристик использовался один и тот же способ.

1.3. МЕТОД НАИМЕНЬШИХ КВАДРАТОВ. ПРЯМОЛИНЕЙНЫЙ ХАРАКТЕР СВЯЗИ МЕЖДУ ДВУМЯ ЭКОНОМИЧЕСКИМИ ФАКТОРАМИ

Теперь мы обсудим вопрос о том, каким образом можно (хотя бы приблизительно) восстановить гипотетическую линейную связь между переменными, если таковая действительно существует.

Мы уже заметили, что при наличии объективной тенденции поддержания линейной связи между переменными x и y естественно рассмотреть ***линейную модель наблюдений***

$$\boxed{y_i = (\alpha + \beta \cdot x_i) + \varepsilon_i,\ i = 1,\text{K}\ ,n.}$$

Если α и β — «истинные» значения параметров линейной модели связи, то

$$\boxed{\varepsilon_i = y_i - (\alpha + \beta \cdot x_i)}$$

представляет собой ***ошибку*** в i - м наблюдении (***error,*** или ***disturbance***). Однако, даже при действительном существовании линейной связи, параметры α и β такой связи остаются неизвестными, и мы можем судить об их истинных значениях лишь приближенно, оценивая значения α и β на основании ограниченного количества имеющихся данных наблюдений (статистических таблиц).

Поиск подходящих оценок для α и β можно осуществлять, например, путем поиска на диаграмме рассеяния прямой, проходящей через точку $(\bar{x}, \bar{y})$ — «центр» системы точек $(x_1, y_1),\text{K}\ ,(x_n, y_n)$ и «наилучшим образом» выражающей на-

правление вытянутости этой системы (облака) точек. Пусть прямая

$$y = \alpha^* + \beta^* x$$

рассматривается в числе прочих в процессе такого поиска. Для i - го наблюдения мы будем наблюдать тогда расхождение («невязку»)

$$\varepsilon_i^* = y_i - \left(\alpha^* + \beta^* x_i\right),$$

причем значения ε_i^* могут быть как положительными, так и отрицательными. При изменении значений α^* и β^* будет изменяться и алгебраическая сумма невязок $\sum_{i=1}^{n} \varepsilon_i^*$. С этой точки зрения, мы можем остановить свой выбор на прямой, для которой соблюдается баланс положительных и отрицательных невязок, так что

$$\sum_{i=1}^{n} \varepsilon_i^* = 0.$$

Соответствующие этой прямой значения α^* и β^* будем обозначать как $\bar{\alpha}$ и $\bar{\beta}$. Итак, прямая

$$y = \bar{\alpha} + \bar{\beta} x$$

проходит через точку $(\bar{x}, \bar{y})$, и если обозначить еще

$$e_i = y_i - \left(\bar{\alpha} + \bar{\beta} x_i\right),$$

то тогда

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

Значение e_i называется ***остатком*** в i - м наблюдении. Для реальных данных, как правило, <u>все</u> остатки отличны от нуля,

так что часть из них имеет положительный знак, а остальные — отрицательный.

Оказывается, что ту же самую прямую $y = \hat{\alpha} + \hat{\beta} x$ можно получить, исходя из другого принципа — ***принципа наименьших квадратов***. Согласно этому принципу, среди всех возможных значений α^*, β^*, претендующих на роль оценок параметров α и β, следует выбирать такую пару α^{**}, β^{**}, для которой

$$\sum_{i=1}^{n} (y_i - \alpha^{**} - \beta^{**} x_i)^2 = \min_{\alpha^*, \beta^*} \sum_{i=1}^{n} (y_i - \alpha^* - \beta^* x_i)^2.$$

Иначе говоря, выбирается такая пара α^{**}, β^{**}, для которой сумма квадратов невязок оказывается наименьшей. Получаемые при этом оценки называются ***оценками наименьших квадратов***, и можно показать, что они совпадают с ранее определенными оценками $\hat{\alpha}$ и $\hat{\beta}$, так что

$$\alpha^{**} = \hat{\alpha}, \quad \beta^{**} = \hat{\beta}.$$

Заметим, что при построении оценок наименьших квадратов заранее не требуется, чтобы соответствующая прямая проходила через точку $(\bar{x}, \bar{y})$; этот факт является свойством оценок наименьших квадратов. Наличие такого свойства мы докажем чуть позднее, а сейчас обратимся к вопросу о том, как практически найти указанные оценки $\hat{\alpha}$ и $\hat{\beta}$.

Если исходить из первого определения, то прежде всего следует заметить, что если прямая $y = \alpha^* + \beta^* x$ проходит через точку $(\bar{x}, \bar{y})$, то тогда $\bar{y} = \alpha^* + \beta^* \bar{x}$, так что

$$\alpha^* = \bar{y} - \beta^* \bar{x},$$

и для поиска «наилучшей» прямой достаточно определить ее угловой коэффициент β^*. Изменяя значения β^* и следя за изменением значений $\sum_{i=1}^{n} \varepsilon_i^*$, мы можем, в принципе, найти искомое $\hat{\beta}$ с любой наперед заданной точностью.

Использование непосредственного перебора значений α^*, β^* с целью минимизации суммы квадратов

$$Q(\alpha^*, \beta^*) = \sum_{i=1}^{n} (y_i - \alpha^* - \beta^* x_i)^2$$

при реализации метода наименьших квадратов также возможно, хотя и требует, конечно, существенно больших вычислительных усилий.

Было бы идеальным, если бы существовала возможность прямого вычисления значений $\hat{\alpha}$ и $\hat{\beta}$ по какой-нибудь формуле на основании известных значений x_i, y_i, $i = 1,\text{K},n$. Такую возможность нам предоставляет еще один подход к поиску параметров $\hat{\alpha}$, $\hat{\beta}$ «наилучшей» прямой.

Заметим, что через каждую пару точек $(x_i, y_i), (x_k, y_k)$ на диаграмме рассеяния можно провести прямую. Всего таких прямых (с учетом совпадающих точек) будет ровно столько, сколько различных пар индексов (i,k) можно образовать на основе n индексов $1,\text{K},n$. А количество таких пар индексов равно числу сочетаний из n элементов по два. Из комбинаторной математики известно, что последняя величина равна $N = n(n-1)/2$. Пусть прямая, проходящая через j - ю пару точек, имеет вид

$$y = \alpha_j + \beta_j x,$$

а точки, через которые она проводится, имеют абсциссы $x_1(j)$ и $x_2(j)$, соответственно.

Обратимся опять к диаграмме рассеяния. Из этой диаграммы видно, что параметры $\bar{\alpha}$ и $\bar{\beta}$ будут очень сильно отличаться для различных пар, и для многих пар не будут иметь ничего общего с параметрами $\bar{\alpha}$, $\bar{\beta}$ «наилучшей» прямой. Оказывается, однако, что эти значения $\bar{\alpha}$ и $\bar{\beta}$ можно получить как взвешенные суммы значений параметров отдельных прямых:

$$\bar{\alpha} = \sum_{j=1}^{N} w_j \alpha_j \ , \quad \bar{\beta} = \sum_{j=1}^{N} w_j \beta_j \ ,$$

где $\sum_{j=1}^{n} w_j = 1$ и веса $w_1, \text{K}\ , w_n$ имеют вид

$$w_j = \frac{\left(x_2(j) - x_1(j)\right)^2}{\sum_{k=1}^{N}\left(x_2(k) - x_1(k)\right)^2} \ ,$$

Нетрудно заметить, что большие веса придаются тем прямым, которые строятся по точкам с далеко разнесенными абсциссами.

Итак, мы имеем возможность получать оценки наименьших квадратов чисто аналитически, сначала вычисляя параметры α_j, β_j отдельных прямых, а затем взвешивая полученные значения. Однако, существует еще один способ получения точных формул для $\bar{\alpha}$ и $\bar{\beta}$, исходящий из принципа наименьших квадратов.

Согласно этому принципу, оценки $\bar{\alpha}$ и $\bar{\beta}$ находятся путем минимизации суммы квадратов

$$Q(\alpha,\beta)=\sum_{i=1}^{n}(y_i-\alpha-\beta\ x_i)^2$$

по всем возможным значениям α и β при заданных (наблюдаемых) значениях $x_1,\text{K},x_n,\ y_1,\text{K},y_n$. Функция $Q(\alpha,\beta)$ как функция двух переменных описывает поверхность $z=Q(\alpha,\beta)$ в трехмерном пространстве с прямоугольной системой координат $\alpha,\beta,z,$ и дело сводится к известной математической задаче поиска точки минимума функции двух переменных.

Такая точка находится путем приравнивания нулю частных производных функции $z=Q(\alpha,\beta)$ по переменным α и β, т. е. приравниванием нулю производной функции $Q(\alpha,\beta)$ как функции только от α при фиксированном β,

$\partial\,\mathrm{Q}(\alpha,\beta)/\partial\beta=0$,

и производной функции $Q(\alpha,\beta)$ как функции только от β при фиксированном α,

$\partial\,\mathrm{Q}(\alpha,\beta)/\partial\beta=0$,

Это приводит к так называемой ***системе нормальных уравнений***

$\partial Q(\alpha,\beta)/\partial\alpha=0$, $\partial\,\mathrm{Q}(\alpha,\beta)/\partial\beta=0$,

решением которой и является пара $\hat{\alpha},\ \hat{\beta}$. Остается заметить, что согласно правилам вычисления производных,

$$\partial Q(\alpha,\beta)/\partial\alpha=2\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)(-1),$$

$$\partial Q(\alpha,\beta)/\partial\beta=2\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)(-x_i),$$

так что искомые значения $\hat{\alpha}$, $\hat{\beta}$ удовлетворяют соотношениям

$$\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i) = 0\,,\ \sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)x_i = 0\,.$$

Эту систему двух уравнений можно записать также в виде

$$\begin{cases} n\hat{\alpha} + \left(\sum_{i=1}^{n} x_i\right)\hat{\beta} = \sum_{i=1}^{n} y_i \\ \left(\sum_{i=1}^{n} x_i\right)\hat{\alpha} + \left(\sum_{i=1}^{n} x_i^2\right)\hat{\beta} = \sum_{i=1}^{n} y_i x_i \end{cases}.$$

Последняя система является системой двух линейных уравнений с двумя неизвестными и может быть легко решена, например, методом подстановки.

Из первого уравнения системы находим:

$$\hat{\alpha} = \tfrac{1}{n}\sum_{i=1}^{n} y_i - \tfrac{1}{n}\hat{\beta}\sum_{i=1}^{n} x_i = \bar{y} - \hat{\beta}\,\bar{x}\,,$$

так что точка $(\bar{x}, \bar{y})$ действительно лежит на прямой $y = \hat{\alpha} + \hat{\beta}x$. Подстановка полученного выражения для $\hat{\alpha}$ во второе уравнение системы дает

$$\tfrac{1}{n}\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right) - \tfrac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\hat{\beta} + \left(\sum_{i=1}^{n} x_i^2\right)\hat{\beta} = \sum_{i=1}^{n} y_i x_i\,,$$

откуда

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

Заметим еще, что

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

$$\sum_{i=1}^{n}\left(y_i-\bar{y}\right)\left(x_i-\bar{x}\right)=\sum_{i=1}^{n}y_ix_i-\bar{y}\sum_{i=1}^{n}x_i-\bar{x}\sum_{i=1}^{n}y_i+n\bar{y}\bar{x}=\sum_{i=1}^{n}y_ix_i-n\bar{y}\bar{x}\ .$$

Последние соотношения позволяют получить более употребительную форму записи выражения для $\hat{\beta}$ (в отклонениях от средних значений)

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

которая в паре с выражением

$$\boxed{\hat{\alpha}=\bar{y}-\hat{\beta}\bar{x}}$$

дает явное и простое решение задачи отыскания оценок $\hat{\alpha}$, $\hat{\beta}$ на основе принципа наименьших квадратов.

Разумеется, такое решение может существовать только при выполнении условия

$$\boxed{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\neq 0\ ,}$$

что равносильно <u>отличию от нуля определителя системы</u>. Действительно, этот определитель равен

$$n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2=n\left[\sum_{i=1}^{n}x_i^2-n\bar{x}^2\right]=n\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\ .$$

Последнее условие называется ***условием идентифицируемости*** модели наблюдений $y_i=\left(\alpha+\beta\cdot x_i\right)+\varepsilon_i$, $i=1,\text{K},n$, и означает попросту, что <u>не все значения</u> x_1,K,x_n <u>совпадают между собой</u>. При нарушении этого условия <u>все</u> точки

(x_i, y_i), $i = 1,\text{K}\,,n$, лежат на одной вертикальной прямой $x = \bar{x}$.

Оценки $\hat{\alpha}$ и $\hat{\beta}$ обычно называют ***оценками наименьших квадратов*** **(*least squares estimates*)**, или LS — оценками. Обратим еще раз внимание на полученное выражение для $\hat{\beta}$. Нетрудно видеть, что в это выражение входят уже знакомые нам суммы квадратов, участвовавшие ранее в определении выборочной дисперсии $Var(x) = \sum_{i=1}^{n} (x_i - \bar{x})^2 \Big/ (n-1)$ и выборочной ковариации $Cov(x,y) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \Big/ (n-1)$, так что, в этих терминах,

$$\boxed{\hat{\beta} = \frac{Cov(x,y)}{Var(x)}}.$$

Отсюда, в частности, видно, что значения $\hat{\beta}$ близки к нулю, если ковариация между наблюдаемыми значениями переменных x и y близка к нулю. (Однако, близость $\hat{\beta}$ к нулю здесь следует понимать как относительную, с учетом реальных значений выборочной дисперсии $Var(x)$.) Кроме того, знак $\hat{\beta}$ совпадает со знаком ковариации $Cov(x,y)$, поскольку $Var(x) > 0$.

Вычисление значений $\hat{\alpha}$ и $\hat{\beta}$ для нашего примера дает значения

$$\hat{\beta} = 0.020415 / 0.162976 = 0.125\,,$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 3.118 - 0.125 \cdot 6.576 = 2.294\,.$$

Таким образом, «наилучшая» прямая имеет вид

$y = 2.294 + 0.125x$,

и мы принимаем ее в качестве аппроксимации для «истинной» модели линейной связи между переменными x и y. Эта аппроксимация указывает на то, что при изменении переменной x на 1 единицу (измерения x) переменная y изменяется «в среднем» на 0.125 единиц (измерения y).

Факт горизонтальности прямой $y = \hat{\alpha} + \hat{\beta}x$ при $\hat{\beta} = 0$ $\left(Cov(x, y) = 0\right)$ и наличие у этой прямой наклона при $\hat{\beta} \neq 0$ $\left(Cov(x, y) \neq 0\right)$, позволяют произвести некоторую детализацию структуры остатков $e_i = y_i - \hat{\alpha} - \hat{\beta}x_i$. С этой целью, опять рассмотрим диаграмму рассеяния, сосредоточившись на какой-нибудь одной точке. Пусть в нашем примере это точка A = (7.1, 3.3). Опустим из этой точки перпендикуляр на ось абсцисс. Он пересечет прямую $y = \bar{x}$ в точке B = (7.1, 3.118) и прямую $y = \hat{\alpha} + \hat{\beta}x$ в точке C = (7.1, 3.183), так что расстояние по вертикали от точки A до прямой $y = \bar{x}$, равное AB = 3.3 — 3.118= 0.182, раскладывается в сумму

$AB = AC + BC$.

Отсюда находим, что расстояние по вертикали от точки A до прямой $y = \hat{\alpha} + \hat{\beta}x$ равно $AC = AB — CB$ = 0.182 — (3.183 — 3.118) = 0.117.

Вообще, для любой точки $\left(x_i, y_i\right)$ на диаграмме рассеяния можно записать:

$y_i - \bar{y} = \left(y_i - \hat{y}_i\right) + \left(\hat{y}_i - \bar{y}\right)$,

где $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$ - ордината точки «наилучшей» прямой, имеющей абсциссу x_i. Возведем обе части последнего пред-

ставления в квадрат и просуммируем левые и правые части полученных для каждого *i* равенств:

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

Входящая в правую часть сумма

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

называется чаще всего ***остаточной суммой квадратов* (*residual sum of squares*)** и имеет аббревиатуру ***RSS*** (Доугерти, Айвазян-Мхитарян, Себер), хотя в литературе по эконометрике можно встретить и такие варианты аббревиатур как *SSR* (Green), а также *ESS* (*error sum of squares* — Harvey, Chatterjie) и *SSE* (Магнус-Катышев-Пересецкий). Поэтому, при чтении различных руководств по эконометрике следует обратить особое внимание на то, какие именно термины и обозначения используются авторами.

Заметим, что если $\hat{\beta}=0$, то $\hat{\alpha}=\bar{x}$ и $\hat{y}_i\equiv\bar{x}$. Следовательно, при $\hat{\beta}=0$

$$\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2=\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2 .$$

При $\hat{\beta}\neq 0$, по самому определению прямой $y=\hat{\alpha}+\hat{\beta}x$, имеем

$$\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2<\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2 .$$

Тенденция линейной связи между x и y выражена в максимальной степени, если $RSS=0$. При этом, все точки $\left(x_i,y_i\right)$, i = 1, 2,..., n, располагаются на одной прямой $y=\hat{\alpha}+\hat{\beta}x$. Тенденция линейной связи между переменными x и y не обна-

руживается вовсе, если RSS совпадает с $\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$.

Таким образом, есть определенные основания предложить в качестве «меры выраженности» в данных наблюдений линейной связи между переменными величину

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

называемую ***коэффициентом детерминации***. Этот коэффициент изменяется в пределах от 0 (при $\hat{\beta} = 0$, т. е. $RSS = TSS$) до 1 (при $RSS = 0$),

$$0 \le R^2 \le 1.$$

Вернемся, однако, к полученному ранее представлению $\sum_{i=1}^{n} (y_i - \bar{y})^2$ в виде

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

и рассмотрим третью сумму в правой части этого представления. Имеем:

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

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

Но

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} \left(y_i - (\hat{\alpha} + \hat{\beta} x_i)\right) = 0$$

(см. первое уравнение из системы нормальных уравнений). К тому же,

$$\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)x_i=\sum_{i=1}^{n}\left(y_i-\left(\hat{\alpha}+\hat{\beta}x_i\right)\right)x_i=0$$

(см. второе уравнение из системы нормальных уравнений). Таким образом,

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

и, следовательно, справедливо представление

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

так что

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

т. е. получено второе представление для R^2 в виде

$$\boxed{R^2=\frac{\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}\ ,}$$

Стоящую здесь в числителе сумму квадратов мы будем называть ***суммой квадратов, объясненной моделью*** **(*explained sum of squares*)**, и будем использовать для ее обозначения аббревиатуру ***ESS***, так что

$$\boxed{ESS=\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2\ .}$$

Сумму квадратов, стоящую в знаменателе, будем называть ***полной суммой квадратов* (*total sum of squares*)** и будем использовать для ее обозначения аббревиатуру ***TSS***, так что

$$TSS = \sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2 .$$

Напомним также, что нами уже была определена ***остаточная сумма квадратов***

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

Все эти три суммы квадратов связаны соотношением

$$TSS = ESS + RSS ,$$

которое представляет собой разложение полной суммы квадратов на сумму квадратов, объясненную моделью, и остаточную сумму квадратов. Используя эти три суммы, мы находим также, что

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

Таким образом, значение R^2 тем выше, чем больше доля объясненной моделью суммы квадратов *ESS* по отношению к полной сумме квадратов *TSS*.

Термины «полная» и «объясненная моделью» суммы квадратов имеют следующее происхождение. Полная сумма квадратов соответствует значению *RSS* в ситуации, когда $\hat{\beta} = 0$ и «наилучшая» прямая имеет вид $y = \bar{y}$, отрицающий наличие линейной зависимости y от x. Вследствие этого, привлечение информации о значениях переменной x не дает ничего нового для объяснения изменений значений y от наблюдения к наблюдению. Степень этой изменчивости мы уже характеризовали значением выборочной дисперсии

$$Var(y) = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \frac{TSS}{n-1} \ ;$$

при этом, $TSS = RSS$ и $ESS = 0$.

В ситуации, когда $\hat{\beta} \neq 0$, мы имеем нетривиальное представление $TSS = ESS + RSS$, с $ESS \neq 0$, и поэтому можно записать:

$$Var(y) = \frac{TSS}{n-1} = \frac{ESS}{n-1} + \frac{RSS}{n-1} \ .$$

Но

$$\frac{ESS}{n-1} = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{n-1} = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{\hat{y}}\right)^2}{n-1} = Var(\hat{y}) \ ,$$

где $\hat{y}$ — переменная, принимающая в i - м наблюдении значение $\hat{y}_i$. (Здесь мы использовали тот факт, что $\sum_{i=1}^{n} e_i = 0$, так что $\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right) = 0$, $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n}\hat{y}_i$ и $\bar{y} = \bar{\hat{y}}$.) К тому же,

$$\frac{RSS}{n-1} = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n-1} = \frac{\sum_{i=1}^{n} e_i^2}{n-1} = \frac{\sum_{i=1}^{n}\left(e_i - \bar{e}\right)^2}{n-1} = Var(e) \ ,$$

где e — переменная, принимающая в i - м наблюдении значение e_i. (Здесь мы использовали тот факт, что $\bar{e} = \sum_{i=1}^{n} e_i \ / \ n = 0$.)

В итоге, мы получаем разложение

$$\boxed{Var(y) = Var(\hat{y}) + Var(e) \ ,}$$

показывающее, что изменчивость переменной y (степень которой характеризуется значением $Var(y)$) частично объясняется изменчивостью переменной $\hat{y}$ (степень которой характеризуется значением $Var(\hat{y})$). Не объясненная переменной $\hat{y}$ часть изменчивости переменной y соответствует изменчивости переменной e (степень которой характеризуется значением $Var(e)$).

Таким образом, вспомогательная переменная $\hat{y}$ берет на себя объяснение некоторой части изменчивости значений переменной y, и эта объясненная часть будет тем больше, чем выше значение коэффициента детерминации R^2, который мы теперь можем записать также в виде

$$\boxed{R^2 = \frac{Var(\hat{y})}{Var(y)} = 1 - \frac{Var(e)}{Var(y)}}.$$

Поскольку переменная $\hat{y}$ получается линейным преобразованием переменной x, то изменчивость $\hat{y}$ однозначно связана с изменчивостью x, так что, в конечном счете, построенная модель объясняет часть изменчивости переменной y изменчивостью переменной x. Поэтому, принять говорить в таком контексте о переменной y как об ***объясняемой*** переменной, а о переменной x — как об ***объясняющей*** переменной.

Вернемся опять к нашему примеру. В этом примере

$$\begin{aligned} ESS &= 0.043474 \\ \underline{RSS &= 0.161231} \\ TSS &= 0.204705, \end{aligned}$$

так что

$Var(\hat{y}) = 0.043474/16 = 0.002717,$

$Var(e) = 0.161231/16 = 0.010077,$

$Var(y) = 0.012784,$

$R^2 = 0.043474/0.204705 = 0.212374.$

Значение коэффициента детерминации оказалось достаточно малым, и один из последующих вопросов будет состоять в том, сколь близким к нулю должно быть значение R^2, чтобы мы могли говорить о практическом отсутствии линейной связи между переменными.

1.4. СВОЙСТВА ВЫБОРОЧНОЙ КОВАРИАЦИИ, ВЫБОРОЧНОЙ ДИСПЕРСИИ И ВЫБОРОЧНОГО КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ

Вернемся теперь к определению выборочной ковариации и отметим некоторые ее свойства.

Пусть a — некоторая постоянная, а x_i, y_i, z_i — переменные, принимающие в i - м наблюдении значения x_i, y_i, z_i, $i = 1,K,n$ (n — количество наблюдений). Тогда a можно рассматривать как переменную, значения которой в i - м наблюдении a_i равно a, и

$$Cov(x,a) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(a_i - \bar{a}) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(a - a),$$

так что

$$\boxed{Cov(x,a) = 0\ .}$$

Далее, очевидно, что

$$\boxed{Cov(x,y) = Cov(y,x)}$$

и что

$$\boxed{Cov(x,x) = Var(x)\ .}$$

Кроме того,

$$Cov(ax,y) = \frac{1}{n-1}\sum_{i=1}^{n}(ax_i - a\bar{x})(y_i - \bar{y}) = a\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}),$$

так что

$$\boxed{Cov(ax, y) = a\ Cov(x, y)\ .}$$

Наконец,

$$Cov(x, y+z) = \tfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i + z_i - (\bar{y} + \bar{z}))$$

$$= \tfrac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left((y_i - \bar{y}) + (z_i - \bar{z})\right)$$

$$= \tfrac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) + \tfrac{1}{n\text{-}1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(z_i - \bar{z}\right),$$

так что

$$\boxed{Cov(x, y+z) = Cov(x, y) + Cov(x, z)\ .}$$

На основе этих свойств, в частности, находим, что

$$\boxed{Var(a) = 0}$$

(постоянная не обладает изменчивостью),

$$\boxed{Var(ax) = a^2 Var(x), \quad Std.Dev.(ax) = a \cdot Std.Dev(x)}$$

(при изменений единицы измерения переменной в a раз, во столько же раз изменяется и величина стандартного отклонения этой переменной),

$$\boxed{Var(x + a) = Var(x)}$$

(сдвиг начала отсчета не влияет на изменчивость переменной).

Наконец,

$$Var(x + y) = Cov(x + y, x + y) =$$

$$= Cov(x, x) + Cov(x, y) + Cov(y, x) + Cov(x, y),$$

т. е.

$$\boxed{Var(x + y) = Var(x) + Var(y) + 2Cov(x, y)}$$

(дисперсия суммы двух переменных отличается от суммы дисперсий этих переменных на величину, равную удвоенному значению ковариации между этими переменными).

Что касается выборочного коэффициента корреляции r_{xy}, то если *изменяются начало отсчета и единица измерения*, скажем, переменной x, так что вместо значений x_1, K, x_n мы получаем значения

$\tilde{x}_i = a + bx_i, \quad i = 1, \text{K}, n, \quad (b > 0)$

переменной $\tilde{x} = a + bx$, то тогда

$$r_{\tilde{x}y} = \frac{Cov(\tilde{x}, y)}{\sqrt{Var(\tilde{x})}\sqrt{Var(y)}} = \frac{Cov(a + bx, y)}{\sqrt{Var(a + bx)}\sqrt{Var(y)}} =$$

$$= \frac{bCov(x, y)}{\sqrt{b^2 Var(x)}\sqrt{Var(y)}} = r_{xy}.$$

Иными словами, выборочный коэффициент корреляции r_{xy}, *инвариантен* относительно выбора единиц измерения и начала отсчета переменных x и y.

В то же время, этого нельзя сказать об оценке $\hat{\beta}_x$ коэффициента β в модели наблюдений $y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \text{K}, n.$. Действительно, если, скажем, мы переходим к новой единице измерения переменной x, так что вместо значений x наблюдаются значения переменной $\tilde{x} = bx$, то тогда оценка $\hat{\beta}_{\tilde{x}}$ коэффициента β в модели наблюдений $y_i = \alpha + \beta \tilde{x}_i + \varepsilon_i$, $i = 1, \text{K}, n$, равна

$$\hat{\beta}_{\tilde{x}} = \frac{Cov(\tilde{x}, y)}{Var(\tilde{x})} = \frac{Cov(bx, y)}{Var(bx)} = \frac{bCov(x, y)}{b^2 Var(x)} = \frac{1}{b}\beta_x.$$

Таким образом, изменяя единицу измерения переменной x (или переменной y), мы можем получать существенно раз-

личные значения $\hat{\beta}$, от сколь угодно малых до сколь угодно больших. (Желательно выбирать единицы измерения таким образом, чтобы сравниваемые переменные имели одинаковый порядок.) Близость значений $\hat{\beta}$ к нулю всегда должна интерпретироваться с оглядкой на используемые единицы измерения переменных x и y.

Отметим, в этой связи, полезное представление $\hat{\beta}$ в виде

$$\boxed{\hat{\beta} = r_{xy}\sqrt{\frac{Var(y)}{Var(x)}}}.$$

Действительно,

$$\hat{\beta} = \frac{Cov(x,y)}{Var(x)} = \frac{r_{xy}\sqrt{Var(x)}\sqrt{Var(y)}}{Var(x)},$$

откуда и вытекает указанное представление. Из этого представления получаем, в частности, что при $Var\ (x) = Var\ (y)$ имеет место равенство $\hat{\beta} = r_{xy}$, и тогда выраженность линейной связи между x и y непосредственно отражается в близости значения $\hat{\beta}$ к 1 или -1.

Рассмотрим теперь коэффициент корреляции $r_{y\hat{y}}$ между переменными y и $\hat{y}$, где $\hat{y} = \hat{\alpha} + \hat{\beta}x$, а $\hat{\alpha}$ и $\hat{\beta}$ — оценки наименьших квадратов параметров α и β гипотетической линейной связи между переменными x и y. Замечая, что $y = \hat{y} + e$ (т.к. $e_i = y_i - \hat{y}_i$ по определению), находим:

$$r_{y\hat{y}} = \frac{Cov(y,\hat{y})}{\sqrt{Var(y)}\sqrt{Var(\hat{y})}} = \frac{Cov(\hat{y}+e,\hat{y})}{\sqrt{Var(y)}\sqrt{Var(\hat{y})}}$$

$$= \frac{Cov(\hat{y},\hat{y}) + Cov(e,\hat{y})}{\sqrt{Var(y)}\sqrt{Var(\hat{y})}}.$$

Но ранее мы уже получили (при выводе разложения для TSS) соотношение

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

которое, с учетом соотношения $\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right) = 0$, приводит к равенству

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

левая часть которого есть не что иное как
$Cov(e, \hat{y}) = Cov(y - \hat{y}, \hat{y})$.

Следовательно,

$$r_{y\hat{y}} = \frac{Var(\hat{y})}{\sqrt{Var(y)}\sqrt{Var(\hat{y})}} = \sqrt{\frac{Var(\hat{y})}{Var(y)}} ,$$

так что

$$\boxed{r_{y\hat{y}}^2 = \frac{Var(\hat{y})}{Var(y)} = R^2 .}$$

Последнее соотношение показывает, что коэффициент детерминации равен квадрату коэффициента корреляции между переменными y и $\hat{y}$, так что при достаточно сильно выраженной линейной связи между переменными x и y, что соответствует значению R^2, близкому к 1, оказывается близким к 1 и коэффициент корреляции между переменными y и $\hat{y}$.

По причинам, которые будут ясны из дальнейшего рассмотрения, $r_{y\hat{y}}$ называют ***множественным коэффициентом корреляции*** (***multiple-R***, ***множественный-R***).

Отметим также, что переменная $\hat{y}$ измеряется в тех же единицах, что и переменная y, и при изменении масштаба из-

мерения переменной y значение $r_{y\hat{y}}$ не изменяется. Отсюда вытекает, что коэффициент детерминации R^2 *инвариантен* относительно изменения масштаба и начала отсчета переменных x и y.

Заметим, наконец, что

$$r_{y\hat{y}} = \frac{Cov(y,\hat{y})}{\sqrt{Var(y)}\sqrt{Var(\hat{y})}} = \frac{Cov(y,\hat{\alpha}+\hat{\beta}\,x)}{\sqrt{Var(y)}\sqrt{Var(\hat{\alpha}+\hat{\beta}\,x)}}$$

$$= \frac{\hat{\beta}\,Cov(y,x)}{\sqrt{Var(y)}\sqrt{\hat{\beta}^2 Var(x)}} = \frac{\text{sign}(\hat{\beta})\cdot Cov(y,x)}{\sqrt{Var(y)}\sqrt{Var(x)}} .$$

(здесь sign(z)=-1 для z<0, sign(z)=0 для z=0, sign(z)=1 для z>0)
Поскольку же

$$\hat{\beta} = \frac{Cov(x,y)}{Var(x)} ,$$

то $\text{sign}(\hat{\beta}) = \text{sign}(Cov(x,y))$, и

$$\boxed{r_{y\hat{y}} = \text{sign}(Cov(x,y))\cdot r_{xy} ,}$$

так что

$$\boxed{r^2_{y\hat{y}} = r^2_{xy} = R^2 ,}$$

и мы можем установить значение R^2 еще до построения модели линейной связи.

Замечание

Если $r_{xy} < 0$, то $\text{sign}(Cov(y,x)) = -1$ и $r_{y\hat{y}} > 0$; если $r_{xy} > 0$, то $\text{sign}(Cov(y,x)) = 1$ и $r_{y\hat{y}} > 0$, так что всегда $\boxed{r_{y\hat{y}} > 0 .}$

1.5. «ОБРАТНАЯ» МОДЕЛЬ ПРЯМОЛИНЕЙНОЙ СВЯЗИ

Пусть наша задача состоит в оценивании модели прямолинейной связи между некоторыми переменными x и y на основе наблюдений n пар (x_i, y_i), $i = 1,\text{K}\,,n$, значений этих переменных. Мы уже рассмотрели вопрос об оценивании параметров такой связи, исходя из модели наблюдений $y_i = (\alpha + \beta\, x_i) + \varepsilon_i$, $i = 1,\text{K}\,,n$. Что изменится, если мы будем исходить из «обратной» модели $x_i = (\alpha + \beta\, y_i) + \varepsilon_i$, $i = 1,\text{K}\,,n$?

Пусть $\hat{\alpha}_{xy}, \hat{\beta}_{xy}$ — оценки параметров α и β в модели наблюдений $x_i = (\alpha + \beta\, y_i) + \varepsilon_i$, $i = 1,\text{K}\,,n$, а $\hat{\alpha}_{yx}, \hat{\beta}_{yx}$ — оценки параметров в модели наблюдений $y_i = (\alpha + \beta\, x_i) + \varepsilon_i$, $i = 1,\text{K}\,,n$. Тогда

$$\hat{\beta}_{xy} \cdot \hat{\beta}_{yx} = \frac{Cov(x,y)}{Var(y)} \cdot \frac{Cov(x,y)}{Var(x)} = \left(\frac{Cov(x,y)}{\sqrt{Var(y)}\sqrt{Var(x)}} \right)^2 ,$$

т. е.

$$\boxed{\hat{\beta}_{xy} \cdot \hat{\beta}_{yx} = r_{xy}^2 ,}$$

или

$$\boxed{\hat{\beta}_{xy} \cdot \hat{\beta}_{yx} = R^2 .}$$

В то же время, по первой модели наблюдений мы получаем наилучшую прямую

$$x = \hat{\alpha}_{xy} + \hat{\beta}_{xy} y ,$$

а по второй — прямую

$$y = \hat{\alpha}_{yx} + \hat{\beta}_{yx} x .$$

Первую прямую мы можем записать в виде

$$y = -\frac{\hat{\alpha}_{xy}}{\hat{\beta}_{xy}} + \frac{1}{\hat{\beta}_{xy}} x \ .$$

Сравнивая коэффициенты при x в двух последних уравнениях, находим, что эти коэффициенты равны в том и только в том случае, когда выполнено соотношение

$$\hat{\beta}_{yx} = \frac{1}{\hat{\beta}_{xy}} \ ,$$

т. е.

$$\hat{\beta}_{yx} \cdot \hat{\beta}_{xy} = 1 \ ,$$

или, с учетом предыдущего, когда $R^2 = 1$.

Что касается отрезков на осях, то они будут совпадать тогда и только тогда, когда

$$\hat{\alpha}_{yx} = -\frac{\hat{\alpha}_{xy}}{\hat{\beta}_{xy}} ,$$

или

$$\hat{\alpha}_{yx} \cdot \hat{\beta}_{xy} = -\hat{\alpha}_{xy} \ .$$

Но

$$\hat{\alpha}_{yx} = \bar{y} - \hat{\beta}_{yx} \bar{x} \ ,$$

так что

$$\hat{\alpha}_{yx} \cdot \hat{\beta}_{xy} = (\bar{y} - \hat{\beta}_{yx} \bar{x}) \hat{\beta}_{xy} = \bar{y} \hat{\beta}_{xy} - \hat{\beta}_{yx} \hat{\beta}_{xy} \bar{x} \ .$$

При $R^2 = 1$ получаем

$$\hat{\alpha}_{yx} \cdot \hat{\beta}_{xy} = \bar{y} \hat{\beta}_{xy} - \bar{x} \ .$$

В то же время,

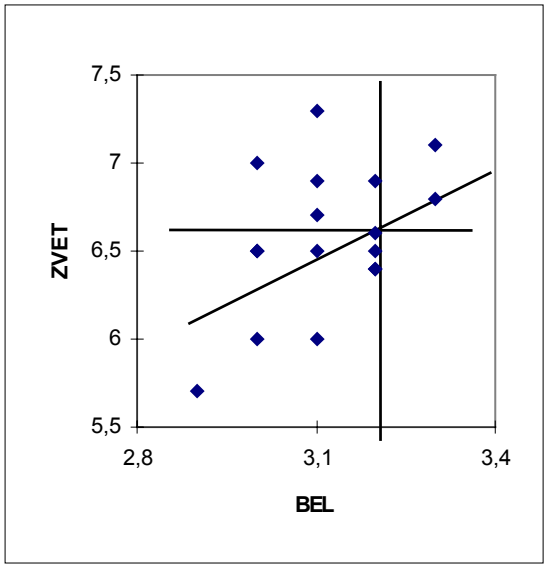
$$\hat{\alpha}_{xy} = -\bar{x} + \hat{\beta}_{xy} \bar{y} \ ,$$

так что при $R^2 = 1$ совпадают и отрезки на осях, т. е. наилучшая прямая одна и та же при обеих моделях наблюдений, и это есть прямая, на которой расположены все наблюдаемые точки (x_i, y_i), $i = 1,\text{K}\ ,n$.

Иными словами, наилучшие прямые, построенные по двум альтернативным моделям, совпадают в том и только в том случае, когда все точки (x_i, y_i), $i = 1,\text{K}\ ,n$, расположены на одной прямой (так что $e_1,\text{K}\ ,e_n = 0$); при этом, $R^2 = 1$. В противном случае, $R^2 \neq 1$ и подобранные «наилучшие» прямые имеют разные угловые коэффициенты.

Кстати, в рассмотренном нами примере с уровнями безработицы, диаграмма рассеяния с переставленными осями (соответствующими модели наблюдений $x_i = (\alpha + \beta\ y_i) + \varepsilon_i$, $i = 1,\text{K}\ ,n$) имеет вид

РИС. 5



Количество точек с совпадающими знаками отклонений координат от средних значений равно 10 (4+ 6, с учетом совпадений), а число точек с противоположными знаками отклонений координат от средних значений равно 7 (4+3, с учетом совпадений). Соответственно, «облако точек» имеет некоторую вытянутость вдоль наклонной прямой, проведенной через «центр» облака. «Наилучшая» прямая имеет вид

$x = 1.291 + 1.695y$;

коэффициент детерминации равен

$R^2 = 0.212374$.

Произведение угловых коэффициентов 0.125265 и 1.695402 наилучших прямых в «прямой» и «обратной» моделях наблюдений равно 0.212374 и совпадает со значением R^2.

Отметим, что несовпадение наилучших прямых, конечно, связано с тем, что в этих двух альтернативных моделях наблюдений мы минимизировали различные суммы квадратов: в «прямой» модели мы минимизировали сумму квадратов отклонений точек от подбираемой прямой в направлении, параллельном оси y, а во втором — в направлении, параллельном оси x.

1.6. ПРОПОРЦИОНАЛЬНАЯ СВЯЗЬ МЕЖДУ ПЕРЕМЕННЫМИ

Хотя на практике не рекомендуется отказываться от включения свободного члена в уравнение подбираемой прямолинейной связи, если только его отсутствие не обосновывается надежной теорией (как в физике — закон Ома), мы все же иногда сталкиваемся с необходимостью подбора прямой, проходящей через начало координат. Позднее мы приведем соответствующие примеры.

Итак, пусть мы имеем наблюдения (x_i, y_i), $i = 1,\text{K},n$, и предполагаем, что гипотетическая линейная связь между переменными x и y имеет вид

$$\boxed{y = \beta x}$$

(***пропорциональная связь*** между переменными), так что ей соответствует модель наблюдений

$$\boxed{y_i = \beta x_i + \varepsilon_i,\ i = 1,\text{K},n.}$$

Применение метода наименьших квадратов в этой ситуации сводится к минимизации суммы квадратов невязок

$$Q(\beta) = \sum_{i=1}^{n} (y_i - \beta x_i)^2$$

по всем возможным значениям β. Последняя сумма квадратов является функцией единственной переменной β (при известных значениях x_i, y_i, $i = 1,\text{K},n$), и точка минимума этой функции легко находится. Для этого мы приравниваем нулю производную $Q(\beta)$ по β:

$$2\sum_{i=1}^{n}(y_i - \hat{\beta} x_i)(-x_i) = 0,$$ (нормальное уравнение)

откуда получаем:

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

или

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

Отсюда видно, что при таком подборе

$$\tilde{\beta} \neq \frac{Cov(x,y)}{Var(x)} ,$$

и точка $(\bar{x},\bar{y})$ уже не лежит, как правило, на подобранной прямой

$$\boxed{y = \tilde{\beta} x .}$$

Более того, в такой ситуации

$$\boxed{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 \neq \sum_{i=1}^{n}\left(y_i - \tilde{y}_i\right)^2 + \sum_{i=1}^{n}\left(\tilde{y}_i - \bar{y}\right)^2 ,}$$

где

$$\boxed{\tilde{y}_i = \tilde{\beta} x_i ,}$$

и поэтому использовать для вычисления коэффициента детерминации выражение

$$R^2 = \frac{\sum_{i=1}^{n}\left(\tilde{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

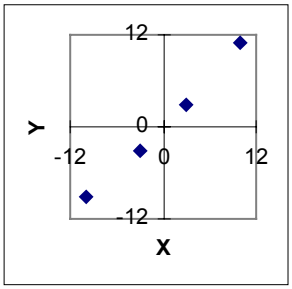
не имеет смысла. В этой связи полезно рассмотреть следующий искусственный пример.

Пример

Пусть переменные x и y принимают в четырех наблюдениях значения, приведенные в следующей таблице

i	1	2	3	4
x_i	10	3	–10	-3
y_i	11	3	-9	-3

соответствующей диаграмме рассеяния



и мы предполагаем пропорциональную связь между этими переменными, что соответствует модели наблюдений $y_i = \beta x_i + \varepsilon_i$, $i = 1,2,3,4.$ Для этих данных

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

так что $\hat{y}_i \equiv x_i$, $i = 1,\text{K}\ ,n$. При этом,

RSS = (11—10)^2 + (3-3)^2+ (-9+10)^2+ (-3+3)^2 = 2,

TSS = (11-0.5)^2+ (3-0.5)^2+ (-9-0.5)^2+ (-3-0.5)^2 = 219,

ESS = (10-0.5)^2+ (3-0.5)^2+ (-10-0.5)^2+ (-3-0.5)^2 = 219,

так что здесь $RSS + ESS \neq TSS$, и вычисление R^2 по формуле

$$R^2 = ESS/TSS$$

приводит к значению $R^2 = 1$. Но последнее возможно только если все точки $(x_i, y_i), i = 1,2,3,4,$ лежат на одной прямой, а у нас это не так. Заметим также, что в этом примере сумма остатков $e_1 + e_2 + e_3 + e_4 = 2 \neq 0$, что невозможно в модели с включением в правую часть постоянной составляющей.

Можно, конечно, попытаться справиться с возникающим при оценивании модели без постоянной составляющей затруднением, попросту игнорируя нарушение соотношения

$RSS + ESS = TSS$ и определяя коэффициент детерминации соотношением

$$R^2 = 1 - (RSS/TSS),$$

и именно такое значение R^2 приводится в протоколах некоторых пакетов программ анализа статистических данных, например пакета ECONOMETRIC VIEWS (TSP). Для нашего иллюстративного примера с четырьмя наблюдениями использование последнего приводит к значению $R^2 = 1 - (2/219) = 0.990860$, которое не противоречит интуиции и представляется разумным. Однако, к сожалению, и такой подход к определению коэффициента детерминации не решает проблемы, поскольку, в принципе, при оценивании модели без постоянной составляющей возможны ситуации, когда $RSS > TSS$, что приводит к отрицательным значениям R^2.

Пример

Пусть переменные x и y принимают в четырех наблюдениях значения, приведенные в следующей таблице

i	1	2	3	4
x_i	0	0.2	0.4	3
y_i	0.5	0.8	1.2	2

что соответствует диаграмме рассеяния

и мы предполагаем пропорциональную связь между этими переменными, что соответствует модели наблюдений

$y_i = \beta x_i + \varepsilon_i$, $i = 1,2,3,4$. Для этих данных $\hat{\beta} = 0.721739$. При этом, $RSS = 1.537652$, $TSS = 1.2675$, и вычисление R^2 по формуле $R^2 = 1 - (RSS/TSS)$ приводит к отрицательному значению $R^2 = -0.213138$.

Преодолеть возникающие затруднения можно, если определить R^2 в модели наблюдений без постоянной составляющей формулой

$$R^2 = 1 - \frac{RSS}{\sum_{i=1}^{n} y_i^2},$$

в которой используется сумма квадратов нецентрированных значений переменной y (отклонений значений переменной y от «нулевого уровня»). При таком определении, неотрицательность коэффициента R^2 гарантируется наличием соотношения

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

которое отражает геометрическую сущность метода наименьших квадратов (аналог знаменитой теоремы Пифагора для многомерного простанства) и выполняется как для модели без постоянной составляющей, так и для модели с наличием постоянной составляющей в правой части модели наблюдений. Деля обе части последнего равенства на $\sum_{i=1}^{n} y_i^2$, приходим к соотношению

$$1 = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n} y_i^2} + \frac{\sum_{i=1}^{n}\hat{y}_i}{\sum_{i=1}^{n} y_i^2},$$

из которого непосредственно следует, что

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n} y_i^2} = \frac{\sum_{i=1}^{n}\hat{y}_i}{\sum_{i=1}^{n} y_i^2} \geq 0 .$$

(Доказать заявленное равенство не сложно. Действительно,

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n}\left(y_i - \hat{y}_i + \hat{y}_i\right)^2 = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 + \sum_{i=1}^{n}\hat{y}_i^2 + 2\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)\hat{y}_i .$$

Но

$$\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)\hat{y}_i = \sum_{i=1}^{n}\left(y_i - \hat{\beta} x_i\right)\hat{\beta} x_i = \hat{\beta}\sum_{i=1}^{n}\left(y_i - \hat{\beta} x_i\right)x_i = 0 ,$$

(см. нормальное уравнение), что и приводит к искомому результату.)

В последнем примере использование определения R^2 с <u>не центрированными</u> y_i дает $R^2 = 1 - \left(1.537652/6.33\right) = 0.242$.

1.7. ПРИМЕРЫ ПОДБОРА ЛИНЕЙНЫХ МОДЕЛЕЙ СВЯЗИ МЕЖДУ ДВУМЯ ФАКТОРАМИ. ФИКТИВНАЯ ЛИНЕЙНАЯ СВЯЗЬ

В этом разделе мы рассмотрим примеры подбора линейных моделей связи для конкретных данных.

Пример 1

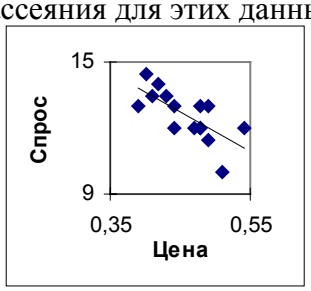
В следующей таблице приведены данные об изменении потребительского спроса на куриные яйца семи семейных хо-

зяйств в зависимости от цены на этот продукт в течение 15 недель:

i	1	2	3	4	5	6	7	8	9	10
Спрос	12	10	13	11.5	12	13	12	12	12	13
Цена	0.54	0.51	0.49	0.49	0.48	0.48	0.48	0.47	0.44	0.44

i	11	12	13	14	15
Спрос	13.5	14	13.5	14.5	13
Цена	0.43	0.42	0.41	0.40	0.39

(спрос измерялся в дюжинах, цена — в долларах). Диаграмма рассеяния для этих данных имеет следующий вид:



Предполагая, что модель наблюдений имеет вид $y_i = \alpha + \beta x_i + \varepsilon_i$, $i = 1,\text{K}\ ,n,$ где y_i — спрос в *i*-ю неделю, а x_i — цена в *i*-ю неделю, мы получаем следующие оценки для неизвестных параметров α и β модели линейной связи между ценой и спросом: $\hat{\alpha} = 21.100,\ \hat{\beta} = -18.559$. Таким образом, подобранная модель линейной связи имеет вид $y = 21.100 - 18.559\ x$. При этом,

$TSS = 17.6,\ RSS = 8.562,\ ESS = 9.038,$

так что коэффициент детерминации оказывается равным $R^2 = 0.514,$ т. е. изменчивость цен объясняет 51.4% изменчивости спроса на куриные яйца. На диаграмме рассеяния изо-

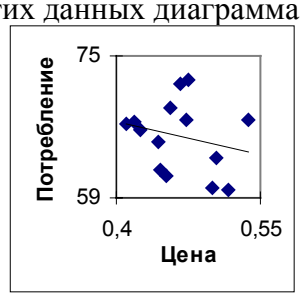
бражена прямая линия, соответствующая подобранной модели линейной связи.

Пример 2

В следующей таблице приведены данные о годовом потреблении свинины y на душу населения в США (в фунтах) и оптовых ценах на свинину x (в долларах за фунт) за период с 1948 по 1961 год:

Год	Потр.	Цена		Год	Потр.	Цена
1948	67.8	0.5370		1955	66.6	0.4256
1949	67.7	0.4726		1956	67.4	0.4111
1950	69.2	0.4556		1957	61.5	0.4523
1951	71.9	0.4655		1958	60.2	0.4996
1952	72.4	0.4735		1959	67.6	0.4183
1953	63.5	0.5047		1960	65.2	0.4433
1954	60.0	0.5165		1961	62.2	0.4448

Для этих данных диаграмма рассеяния имеет вид



Предполагая, что модель наблюдений имеет вид $y_i = \alpha + \beta x_i + \varepsilon_i$, $i = 1,\text{K},n$, где y_i — потребление свинины в i-й год рассматриваемого периода, а x_i — оптовая цена на свинину в этом году, мы получаем следующие оценки для неизвестных параметров α и β модели линейной связи между оптовой ценой и потреблением: $\hat{\alpha} = 77.552$, $\hat{\beta} = -24.925$. Таким образом, подобранная модель линейной связи имеет вид $y = 77.552 - 24.925\,x$. При этом,

$TSS = 208.194,\ RSS = 196.701,\ ESS = 11.493,$

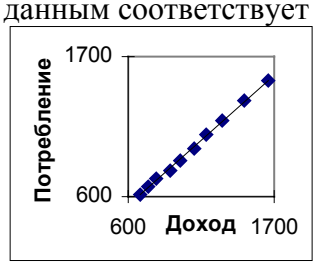
так что коэффициент детерминации здесь оказывается равным $R^2 = 0.055$. Изменчивость оптовой цены объясняет здесь лишь 5.5% изменчивости потребления свинины.

Пример 3

Рассмотрим данные о размерах совокупного располагаемого дохода и совокупных расходах на личное потребление в США в период с 1970 по 1979 год. Обе величины выражены в текущих долларах США.

Год	Расп. доход	Потребление
1970	695.2	621.7
1971	751.9	672.4
1972	810.3	737.1
1973	914.0	811.7
1974	998.1	887.9
1975	1096.2	976.6
1976	1194.3	1084.0
1977	1313.5	1204.0
1978	1474.3	1346.7
1979	1650.5	1506.4

Этим данным соответствует диаграмма рассеяния



Предполагая, что модель наблюдений имеет вид $y_i = \alpha + \beta x_i + \varepsilon_i$, $i = 1,\text{K},n,$ где y_i — совокупные расходы на личное потребление в *i*-й год рассматриваемого периода, а x_i — совокупный располагаемый доход в этом году, мы получаем следующие оценки для неизвестных параметров α и β мо-

дели линейной связи между совокупным располагаемым доходом и совокупными расходами на личное потребление: $\hat{\alpha} = -30.534,\ \hat{\beta} = 0.932.$ Таким образом, подобранная модель линейной связи имеет вид $y = -30.534 - 0.932\,x$. При этом,

$TSS = 791138.545,\ RSS = 740.320,\ ESS = 790398.225,$

так что коэффициент детерминации здесь оказывается равным $R^2 = 0.9995$. Изменчивость совокупного располагаемого дохода объясняет здесь более 99.95% изменчивости совокупных расходов на личное потребление.

Впрочем, не следует слишком оптимистически интерпретировать близкие к единице значения коэффициента детерминации R^2 как указание на то, что изменения значений объясняемой переменной практически полностью определяются именно изменениями значений объясняющей переменной. В этой связи, рассмотрим следующий поучительный пример.

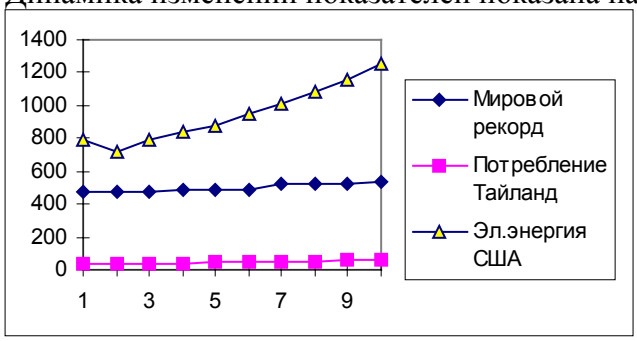
Пример 4

Рассмотрим динамику изменений в период с 1957 по 1966 годы трех совершенно различных по природе показателей: E — суммарного производства электроэнергии в США (в млрд. квт-час), C — совокупных потребительских расходов в Тайланде (в млрд. бат) и H — мирового рекорда на конец года в прыжках в высоту с шестом среди мужчин (в см). Значения этих показателей приведены в таблице:

Год	Потребление Тайланд млрд бат	Эл. энергия США млрд квт-час	Мир. рекорд (прыжки с шестом) см
1957	34.9	716	478
1958	35.9	724	478
1959	37.9	797	478
1960	41.1	844	481
1961	43.5	881	483
1962	46.7	946	493

Год	Потребление Тайланд млрд бат	Эл. энергия США млрд квт-час	Мир. рекорд (прыжки с шестом) см
1963	48.9	1011	520
1964	52.0	1083	528
1965	56.1	1157	528
1966	62.6	1249	534

Динамика изменений показателей показана на графике:



По этим данным мы можем формально, используя метод наименьших квадратов, подобрать модели линейной зависимости каждого из трех показателей от каждого из остальных показателей. Это приводит, например, к моделям

$$E = -2625.5 + 7.131H\,, \qquad R^2 = 0.900;$$

$$C = -129.30 + 0.350H, \qquad R^2 = 0.871;$$

$$E = \ 23.90 + 19.950C, \qquad R^2 = 0.993;$$

$$C = -0.860 + 0.0498E, \qquad R^2 = 0.993.$$

(Заметим, кстати, что произведение угловых коэффициентов двух последних прямых, соответствующих моделям линейной связи, в которых объясняемая и объясняющая переменная меняются местами, равно $19.950 \cdot 0.0498 = 0.993$ и совпадает со значением коэффициента детерминации R^2 в этих двух подобранных моделях.)

Мы видим, что во всех подобранных моделях значения коэффициента детерминации весьма высоки, и это формально означает, что изменчивость «объясняющих» переменных в этих моделях составляет значительный процент от изменчивости «объясняемой» переменной, стоящей в левой части уравнения. Однако, вряд ли мы всерьез можем полагать, что динамика роста суммарного производства электроэнергии в США действительно объясняется динамикой роста мирового рекорда по прыжкам в высоту с шестом, несмотря на высокое значение 0.9 коэффициента детерминации в первом из четырех уравнений.

В ситуациях, подобных последнему примеру, принято говорить о ***фиктивной* (*ложной, паразитной — spurious*)** линейной связи между соответствующими показателями. И такие ситуации часто встречаются при рассмотрении показателей, динамика изменений которых обнаруживает заметный тренд (убывание или возрастание) — именно такой характер имеют исследуемые показатели в последнем примере.

Чтобы понять, почему это происходит, вспомним полученное в свое время равенство

$R^2 = r_{yx}^2$.

Из этого равенства вытекает, что близкие к единице значения коэффициента детерминации соответствуют близким по абсолютной величине к единице значениям коэффициента корреляции между переменными y и x. Но этот коэффициент корреляции равен

$$r_{yx} = \frac{Cov(y,x)}{\sqrt{Var(y)}\sqrt{Var(x)}} ,$$

где

$$Cov(y,x) = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right).$$

При фиксированных значениях $Var(x)$ и $Var(y)$, значение r_{xy} будет тем ближе к 1, чем большим будет значение $Cov(y,x) > 0$. Последнее же обеспечивается совпадением знаков разностей $y_i - \bar{y}$ и $x_i - \bar{x}$ для максимально возможной доли наблюдений переменных y и x, что как раз и имеет место, когда в процессе наблюдения обе переменные возрастают или обе переменные убывают по величине. (В этом случае превышение одной из переменных своего среднего значения сопровождается, как правило, и превышением второй переменной своего среднего значения. Напротив, если одна из переменных принимает значение, меньшее среднего значения этой переменной, то и вторая переменная,как правило, принимает значение, меньшее своего среднего.)

Аналогичным образом, значение r_{xy} будет тем ближе к -1, чем меньшим будет значение $Cov(y,x) < 0$. Последнее же обеспечивается несовпадением знаков разностей $y_i - \bar{y}$ и $x_i - \bar{x}$ для максимально возможной доли наблюдений переменных y и x, что имеет место, когда в процессе наблюдения одна из переменных возрастает, а вторая убывает. (В этом случае, если одна из переменных принимает значение, меньшее среднего значения этой переменной, то вторая переменная,как правило, принимает значение, большее своего среднего.)

Из сказанного следует, что близость к единице наблюдаемого значения коэффициента детерминации не обязательно означает наличие причинной связи между двумя рассматриваемыми переменными, а может являться лишь следствием тренда значений обеих переменных.

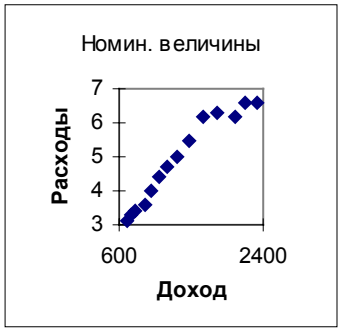
Последнее обстоятельство часто наблюдается при анализе различных экономических показателей, вычисленных без поправки на инфляцию (***недефлированные*** данные). Проиллюстрируем это следующим примером.

Пример 5

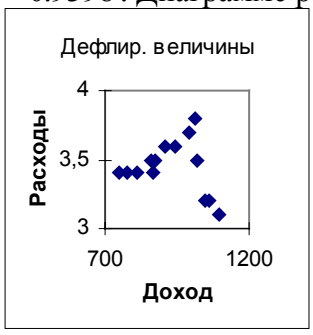
Обратимся к данным о совокупном располагаемом доходе и совокупных личных расходах на местный транспорт в США за период с 1970 по 1983 год. Данные представлены как в текущих долларах США, так и в долларах 1972 года — пересчет к последним выполнен с учетом динамики индекса потребительских цен в указанном периоде. (Уровень цен в 1972 г. принят за 100%.)

Год	Распол. доход номинал.	Расходы номинал.	Распол. доход дефлир.	Расходы дефлир.
1970	695.2	3.1	751.6	3.4
1971	751.9	3.3	779.2	3.4
1972	810.3	3.4	810.3	3.4
1973	914.0	3.6	864.7	3.4
1974	998.1	4.0	857.5	3.5
1975	1096.2	4.4	874.5	3.5
1976	1194.3	4.7	906.4	3.6
1977	1313.5	5.0	942.9	3.6
1978	1474.3	5.5	988.8	3.7
1979	1650.5	6.2	1015.7	3.8
1980	1828.7	6.3	1021.6	3.5
1981	2040.9	6.2	1049.3	3.2
1982	2180.1	6.6	1058.3	3.2
1983	2333.2	6.6	1095.4	3.1

Диаграмма рассеяния для недефлированных величин имеет вид



Соответствующая модель линейной связи: $y = 1.743 + 0.0023\,x$. Коэффициент детерминации равен $R^2 = 0.9398$. Диаграмме рассеяния дефлированных величин



соответствует модель линейной связи $y = 3.758 - 0.0003\,x$. Коэффициент детерминации равен на этот раз всего лишь $R^2 = 0.0353$.

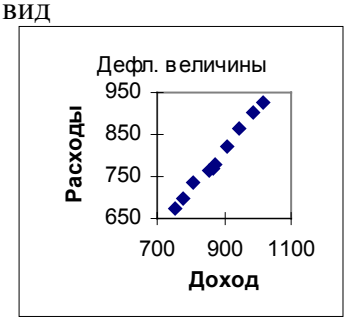
В связи с последним примером, вернемся к примеру 3 и выясним, не является ли обнаруженная там сильная линейная связь между совокупным располагаемым доходом и совокупными расходами на личное потребление лишь следствием использования недефлированных величин.

Для этого рассмотрим дефлированные значения, представленные следующей таблицей, в последнем столбце которой

приведены значения индекса потребительских цен (уровень цен 1972 г. принят за 100%).

Год	Дефлир. доход	Дефлир. потребл.
1970	695.2	621.7
1971	751.9	672.4
1972	810.3	737.1
1973	914.0	811.7
1974	998.1	887.9
1975	1096.2	976.6
1976	1194.3	1084.0
1977	1313.5	1204.0
1978	1474.3	1346.7
1979	1650.5	1506.4

Соответствующая этой таблице диаграмма рассеяния имеет вид



Подобранная модель линейной связи $y = -67.655 + 0.979\,x$. Коэффициент детерминации при переходе от номинальных величин к дефлированным остается очень высоким: $R^2 = 0.9918$. Следовательно, наличие сильной линейной связи между совокупным располагаемым доходом и совокупными расходами на личное потребление не является только лишь следствием инфляционных процессов.

1.8. ОЧИСТКА ПЕРЕМЕННЫХ. ЧАСТНЫЙ КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

Возникновение паразитной линейной связи между двумя переменными часто можно объяснить тем, что хотя эти переменные и не связаны друг с другом причинным образом, изменение каждой из них достаточно хорошо объясняется изменением значений некоей третьей переменной, «координирующей» динамику изменения первых двух переменных. Проиллюстрируем это на примере данных, использованных в примере 4 из предыдущего раздела.

При рассмотрении указанного примера мы подобрали модель линейной связи между значениями суммарного производства электроэнергии в США (***E***) и мирового рекорда на конец года в прыжках в высоту с шестом среди мужчин (***H***). Коэффициент детерминации для этой модели оказался весьма высоким, равным 0.900.

Поскольку динамика изменения этих двух показателей на периоде наблюдений обнаруживает видимый положительный тренд, попытаемся приблизить каждый из них линейной функцией от времени. Подбор методом наименьших квадратов приводит к моделям:

$$E = 613.333 + 59.539\ t\ , \quad H = 459.067 + 7.461\ t\ ,$$

где t обозначает t-й год на периоде наблюдений. При этом, в первом случае коэффициент детерминации равен 0.9812, а во втором коэффициент детерминации равен 0.8705. Иначе говоря, наблюдаемая изменчивость переменных E и H достаточно хорошо «объясняется» изменением переменной t, фактически являющейся здесь выразителем «технического и спортивного прогресса».

Чтобы найти «объективную» связь между показателями E и H, «очищенную» от влияния на эти показатели фактора времени, естественно поступить следующим образом.

Возьмем ряд остатков

$e_E(t) = E_t - (613.333 + 59.539\ t)$,

получаемых при подборе первой модели, и ряд остатков

$e_H(t) = H_t - (459.067 + 7.461\ t)$,

получаемых при подборе второй модели. Тогда переменные e_E и e_H, принимающие значения $e_E(t)$ и $e_H(t)$,соответственно, $t = 1,\text{K},10$, можно интерпретировать, как результат «очистки» переменных *E* и *H* от линейного тренда во времени. Соответственно, «истинная» линейная связь между переменными *E* и *H*, если таковая имеется, должна, скорее всего, измеряться коэффициентом корреляции r_{e_x,e_y} между «очищенными» переменными e_E и e_H.

Подобранная линейная связь между e_E и e_H имеет вид

$e_E = 0.0000 + 1.420\ e_H$;

при этом получаем значение

$R^2 = 0.2454$

против значения 0.900 в модели с «неочищенными» переменными. Коэффициент корреляции между «очищенными» переменными e_E и e_H

$r_{e_E,e_H} = \sqrt{0.2454} = 0.4954$

почти вдвое меньше коэффициента корреляции $r_{E,H} = \sqrt{0.900} = 0.9487$ между «неочищенными» переменными *E* и *H*.

Коэффициент корреляции r_{e_x,e_y} между «очищенными» переменными e_E и e_H называется ***частным коэффициентом корреляции*** между переменными *E* и *H* при исключении влияния на них переменной t .

В дальнейшем мы покажем, что значение $r_{e_E, e_H} = 0.4954$ при $n = 10$ «слишком мало» для того, чтобы можно было отвергнуть гипотезу о том, что коэффициент при e_H в линейной модели связи

$e_E = \gamma + \delta \cdot e_H$

в действительности равен нулю.

1.9. ПРОЦЕНТНОЕ ИЗМЕНЕНИЕ ФАКТОРОВ В ЛИНЕЙНОЙ МОДЕЛИ СВЯЗИ

Вернемся к примеру с совокупным располагаемым доходом (***DPI***) и совокупными расходами на личное потребление (***C***) и будем использовать для анализа дефлированные данные, принимая за базовый 1972 год.

Мы подобрали по таким данным за 1970—1979 годы модель линейной связи

$$C = -67.66 + 0.98\ DPI$$

(мы здесь округлили полученные ранее значения до сотых долей). В соответствии с такой моделью, увеличение реального совокупного располагаемого дохода на 1 млрд. долларов (в единицах 1972 г.) приводит к увеличению совокупного личного потребления на 980 млн. долларов (остальные 20 млн. долларов сохраняются в виде сбережений). Разумеется, имеется в виду только тенденция; ежегодные реальные цифры будут отличаться от предсказываемых моделью. Величина $\beta = 0.98$ оценивает ***склонность к потреблению*** по отношению к располагаемому доходу (***propensity to consumption***).

Зададимся теперь таким вопросом: на сколько процентов изменится совокупный объем потребления C при увеличении совокупного располагаемого дохода на 1% (опять имеем в виду дефлированные величины)?

Итак, предположим, что совокупный располагаемый доход, имевший значение DPI, увеличился на один процент и стал равным $DPI + \Delta DPI$, где ΔDPI — абсолютное приращение совокупного располагаемого дохода, так что

$$\left(\Delta DPI / DPI\right) \cdot 100 = 1 ,$$

откуда $\Delta DPI = 0.01 \cdot DPI$. Такому абсолютному приращению совокупного располагаемого дохода соответствует «в среднем» абсолютное приращение совокупных расходов на потребление

$$\Delta C = 0.98 \cdot \Delta DPI = 0.98 \cdot 0.01 DPI = 0.0098\ DPI ,$$

что соответствует процентному изменению совокупных расходов на потребление, равному

$$\frac{\Delta C}{C} \cdot 100 = \left(\frac{0.0098 DPI}{-67.66 + 0.98 DPI} \right) \cdot 100$$

$$= \frac{0.98 DPI}{0.98 DPI\left(1 - \left(67.66 / 0.98 DPI\right)\right)} = \frac{1}{1 - \left(69.04 / DPI\right)} .$$

Мы видим, что при увеличении DPI на 1%, процентное изменение C оказывается различным и зависит от того, каким было исходное значение DPI. При $DPI < 69.04$ оно даже становится отрицательным, а при $DPI > 69.04$ изменяется, уменьшаясь от $+\infty$ до 1. Если бы у нас значение параметра α было положительным, то тогда

$$\frac{\Delta C}{C} \cdot 100 = \frac{\beta DPI}{\alpha + \beta DPI} = \frac{1}{1 + \left(\alpha / \beta DPI\right)} \le 1 ,$$

и процентное изменение совокупных расходов на потребление возрастало бы от 0 до 1 при увеличении DPI от 0 до $+\infty$.

Впрочем, в интервале наблюдавшихся значений DPI в период с 1970 по 1979 год величина $(\Delta C/C)\cdot 100$ изменяется незначительно: от значения

$$\frac{0.98\cdot 751.6}{-67.66+0.98\cdot 751.6} = 1.10$$

до значения

$$\frac{0.98\cdot 1015.7}{-67.66+0.98\cdot 1015.7} = 1.07 .$$

Обратимся еще раз к примеру с безработицей. В этом примере мы подобрали модель

$$BEL = 2.294 + 0.125\ ZVET ,$$

где BEL — процент безработных среди белого населения США, а $ZVET$ — процент безработных среди цветного населения США.

В соответствии с этой моделью, если количество безработных среди цветного населения вырастет с $ZVET$ % до $(ZVET + 1)$ %, то количество безработных среди белого населения вырастет («в среднем») с BEL % до $(BEL + 0.125)$ %.

В то же время, если речь идет об относительном росте безработицы, то при увеличении доли безработных среди цветного населения на 1%, доля безработных среди белого населения возрастает на

$$\frac{\beta\ ZVET}{\alpha + \beta\ ZVET} = \frac{0.125\ ZVET}{2.294 + 0.125\ ZVET}$$

процентов. Значения $ZVET$ изменяются на периоде наблюдений от 5.7 до 7.3, так что последнее отношение изменяется от

$$\frac{0.125\cdot 5.7}{2.294+0.125\cdot 5.7} = 0.31$$

до

$$\frac{0.125\cdot 7.3}{2.294+0.125\cdot 7.3} = 0.40 .$$

В примере с куриными яйцами (*SPROS* — спрос, *CENA* — цена)

$SPROS = 21.1 - 18.6\ CENA$.

Увеличение цены на 1% приводит к возрастанию цены (в долларах) на

$\Delta CENA = 0.01\ CENA$.

Это, в свою очередь, приводит изменению спроса (в среднем) на

$\Delta SPROS = -18.6 \cdot 0.01\ CENA$,

т. е. к уменьшению спроса (в среднем) на $0.186\ CENA$ дюжин, что составляет

$$\left(\frac{0.186 \cdot CENA}{SPROS}\right) \cdot 100 = \frac{18.6 \cdot CENA}{21.1 - 18.6 \cdot CENA}$$

процентов.

В диапазоне цен от \$0.39 до \$0.54последняя величина изменяется от 0.524 до 0.908, что говорит о ***неэластичном*** (по цене) спросе. Последнее означает, что убытки от продажи яиц по более низкой цене не перекрываются дополнительным доходом от возрастания объема реализации: объем реализации возрастает, но в недостаточной степени.

В то же время, в примере с совокупным располагаемым доходом и совокупными расходами на личное потребление расходы на потребление формально оказываются ***эластичными*** по располагаемому доходу (при изменении совокупного располагаемого дохода на 1% совокупные расходы на личное потребление изменяются в среднем более, чем на 1%).

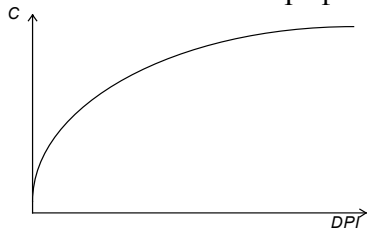
1.10. НЕЛИНЕЙНАЯ СВЯЗЬ МЕЖДУ ПЕРЕМЕННЫМИ

Разумеется, связь между конкретными экономическими факторами вовсе не обязана быть линейной.

Например, если мы рассматриваем зависимость от располагаемого дохода DPI не всех затрат на личное потребление, а лишь затрат C на некоторый продукт питания (или группу продуктов питания), например, на куриные яйца, то уже по чисто физиологическим причинам функция связи

$$\boxed{C = f(DPI)}$$

скорее всего, должна замедлять свой рост при возрастании DPI, так что возможный график этой функции имеет вид



В такой ситуации нельзя говорить о склонности к потреблению данного продукта как о постоянной величине. Вместо этого, в рассмотрение вводят понятие ***предельной* (*marginal*) *склонности к потреблению*** (***MPC***), которая для заданной величины DPI располагаемого дохода определяется формулой

$$MPC(DPI) = \lim_{\Delta DPI \to 0} \frac{f(DPI + \Delta DPI) - f(DPI)}{\Delta DPI}.$$

Иначе говоря,

$$\boxed{MPC(DPI) = \frac{dC}{dDPI} = f'(DPI).}$$

Замедление скорости роста функции $f(DPI)$ соответствует убыванию $MPC(DPI)$ с возрастанием DPI. Уточняя пред-

положения о поведении MPC, можно получить ту или иную форму связи между переменными DPI и C.

Среди прочих возможных форм связи между DPI и C отметим ***степенную*** связь

$$\boxed{C = f(DPI) = \alpha \cdot (DPI)^{\beta}},$$

в которой $\alpha > 0,\ 0 < \beta < 1$. Для такой связи

$$\boxed{MPC(DPI) = \alpha\beta\, DPI^{\beta-1}},$$

так что предельная склонность к потреблению монотонно убывает с ростом DPI.

Степенную форму связи можно привести к линейной форме, если вместо уровней дохода и расходов на потребление рассмотреть логарифмы уровней по какому-нибудь (но одному и тому же!) основанию (например, натуральные или десятичные логарифмы).

Действительно, переходя к логарифмам уровней, получаем соотношение

$$\boxed{\log C = \log\alpha + \beta \cdot \log DPI},$$

или, обозначая $\log C = C^*$, $\log\alpha = \alpha^*$, $\log DPI = DPI^*$,

$C^* = \alpha^* + \beta \cdot DPI^*$.

Линейной модели связи в логарифмах соответствует линейная модель наблюдений

$C^* = \alpha^* + \beta \cdot DPI^* + \varepsilon_i$, $i = 1,\text{K}\ ,n,$

которую мы уже умеем оценивать.

Заметим, что коэффициент β в последних выражениях есть не что иное как

$$\boxed{\beta = \frac{d\log C}{d\log DPI}};$$

эта величина не зависит от выбора основания логарифмов, так что

$$\beta = \frac{d \ln C}{d \ln DPI} ,$$

где используются натуральные логарифмы.

Вообще, если мы имеем связь между какими-то переменными экономическими факторами X и Y в виде

$$\boxed{Y = f(X) ,}$$

то мы определяем функцию

$$\boxed{MPY(X) = \frac{dY}{dX} = f'(X)}$$

как ***предельную склонность Y по отношению к X***.

В экономической теории существенную роль играет ***функция эластичности***, определяемая как предел

$$\boxed{\eta(X) = \lim_{\Delta X \to 0} \frac{\dfrac{f(X + \Delta X) - f(X)}{f(X)} \cdot 100}{\dfrac{\Delta X}{X} \cdot 100}}$$

отношения процентного изменения Y к процентному изменению X, когда последнее стремится к нулю. Правую часть последнего соотношения можно записать в виде

$$\boxed{\eta(X) = \frac{X}{Y} \cdot \frac{dY}{dX} = \frac{X}{Y} \cdot MPY(X) .}$$

Заметим также, что

$$\frac{d \ln f(X)}{d \ln X} = \left(\frac{d \ln f(X)}{dX} \right) \Big/ \left(\frac{d \ln X}{dX} \right) = \frac{X}{Y} \cdot \frac{dY}{dX} ,$$

так что

$$\boxed{\eta(X) = \frac{d \ln Y}{d \ln X} = \frac{X}{Y} \cdot MPY(X) = \frac{dY/Y}{dX/X} .}$$

Значение $MPC(X_0)$ равно угловому коэффициенту касательной к графику функции $Y = f(X)$ при $X = X_0$, тогда как значение $\eta(X_0)$ равно угловому коэффициенту касательной к графику зависимости $\ln Y$ от $\ln X$ при $X = X_0$. Как следствие, условие постоянства $MPC(X)$, т. е. $MPC(X) \equiv \beta$, означает линейную связь между уровнями факторов

$Y = \alpha + \beta X$,

а условие постоянства эластичности $\eta(X) \equiv \beta$ означает линейную связь между логарифмами уровней

$\ln Y = \alpha + \beta \ln X$,

соответствующую степенной связи между уровнями

$$\boxed{Y = exp(\alpha + \beta\, lnX) = Const \cdot X^{\beta}},$$

выражающей степенное возрастание (при $\beta > 0$) или убывание (при $\beta < 0$) уровней фактора Y при возрастании уровней фактора X.

Заметим, что если $\eta(X) \equiv \beta$, то эту постоянную можно трактовать как процентное изменение уровня фактора Y при изменении фактора X на 1%.

Отметим также, что в модели $Y = \alpha + \beta X$ функция эластичности имеет вид

$$\eta(X) = \frac{X}{Y} \cdot \beta = \frac{\beta X}{\alpha + \beta X} = \frac{1}{\frac{\alpha}{\beta X} + 1}$$

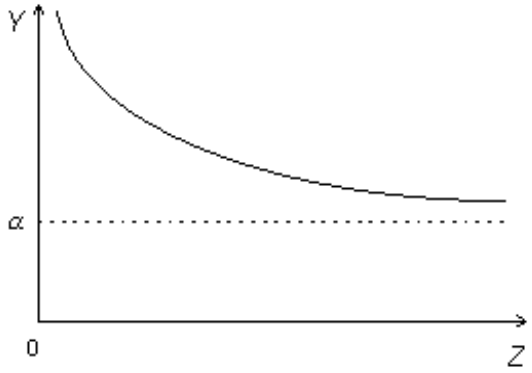
и при $\alpha\beta > 0$ возрастает от 0 до 1 с возрастанием значений X от 0 до ∞. Если $\alpha = 0$, то $\eta(X) \equiv \beta$. При $\alpha\beta < 0$ функция эластичности $\eta(X)$ убывает от $+\infty$ до 1, когда X изменяется от $-\alpha/\beta$ до $+\infty$.

К линейной форме связи можно привести и некоторые другие виды зависимости, характерные для экономических моделей.

Так, если Y — объем плановых инвестиций, а Z — норма процента, то между ними существует связь, которая иногда может быть выражена в форме

$$\boxed{Y = \alpha + \frac{\beta}{Z}\ , \quad \alpha > 0,\ \beta > 0,}$$

и имет графическое представление



Заменой переменной $X = 1/Z$ приводим указанную связь к линейной форме $Y = \alpha + \beta X$. В этой модели эластичность Y по Z отрицательна и меньше единицы по абсолютной величине:

$$\eta(Z) = \frac{dY}{dZ} \cdot \frac{Z}{Y} = \left(-\frac{\beta}{Z^2} \right) \cdot \frac{Z}{\alpha + \frac{\beta}{Z}} = -\frac{\beta}{\beta + \alpha Z}$$

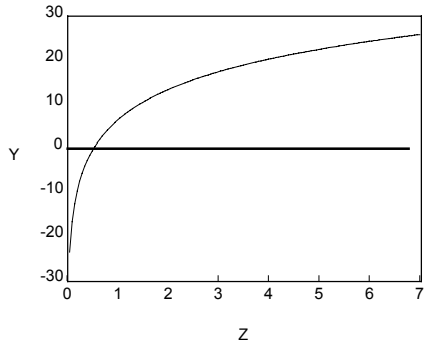
(«объем плановых инвестиций неэластичен по отношению к норме процента»).

В моделях «доход — потребление», относящихся к потреблению продуктов питания, линейная модель в логарифмах

уровней, выражающая уменьшение $MPC(DPI)$ с возрастанием DPI, все же не всегда удовлетворительна, поскольку эластичность в такой модели постоянна. Опять же по чисто физиологическим причинам, скорее более подходящей будет модель связи с убывающей (в конечном счете) эластичностью. Такого рода связь между факторами Y и Z может иметь вид

$$Y = \alpha + \beta \ln Z, \quad \alpha > 0, \ \beta > 0 .$$

(См. следующий график, построенный при $\alpha = 5, \beta = 10$.)



Действительно,

$$\eta(Z) = \frac{dY}{dZ} \cdot \frac{Z}{Y} = \left(\frac{\beta}{Z}\right) \cdot \frac{Z}{\alpha + \beta \ln Z} \xrightarrow[Z \to \infty]{} 0 ;$$

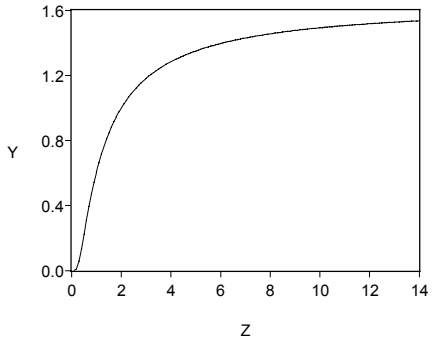
однако, здесь возникают проблемы с отрицательными значениями Y при малых значениях Z.

Последнего недостатка нет в модели

$$\ln Y = \alpha - \frac{\beta}{Z} , \quad \beta > 0 ,$$

т. е.

$$Y = exp\left(\alpha - \frac{\beta}{Z}\right).$$



(График построен при значениях α=0.1, β=1.) Здесь

$$\eta(Z) = \frac{\beta}{Z}$$

(закон Энгеля убывания эластичности потребления продуктов питания по доходу).

Обе последние модели сводятся к линейной форме связи путем перехода от уровней переменных к их логарифмам или обратным величинам.

Замечание

Если исследователь принимает модель наблюдений

$\ln Y_i = \alpha^* + \beta \ln X_i + \varepsilon_i$,

то тем самым, он соглашается тем, что

$Y_i = e^{\alpha^*} \cdot X_i^{\beta} \cdot e^{\varepsilon_i}$,

или

$$Y_i = \alpha \cdot X_i^{\beta} \cdot \nu_i ,$$

т. е. соглашается с мультипликативным вхождением ошибок ν_i в нелинейное уравнение для Y_i.

В то же время, не исключено, что по существу дела модель должна иметь вид

$$Y_i = \alpha \cdot X_i^{\beta} + \nu_i ,$$

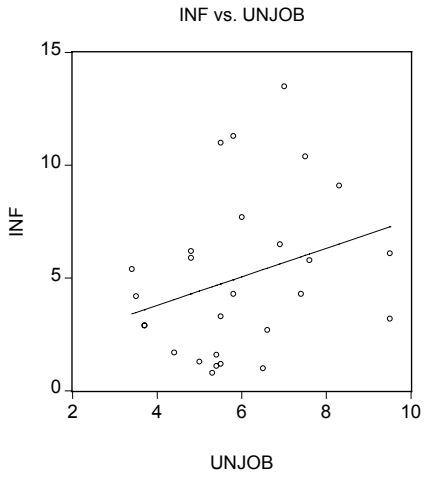
т. е. имеет аддитивные ошибки. В последнем случае взятие логарифмов от обеих частей не приводит к линейной модели наблюдений. В такой ситуации оценки наименьших квадратов параметров α и β приходится получать ***итерационными методами***, в процессе реализации которых производится ***последовательное приближение*** к минимуму суммы квадратов

$$Q(a,b) = \sum_{i=1}^{n} \left(Y_i - a\, X_i^b \right)^2 .$$

1.11. ПРИМЕР ПОДБОРА МОДЕЛЕЙ НЕЛИНЕЙНОЙ СВЯЗИ, СВОДЯЩИХСЯ К ЛИНЕЙНОЙ МОДЕЛИ.

Суть политики Кеннеди-Джонсона (Джон Кеннеди — президент США с 1961 по 1963 г., Линдон Джонсон — президент США с 1963 по 1969 г.) состояла в сокращении налогов, увеличении расходов на оборону и ускорении роста количества денег в обращении. Предполагалось, что это вызовет оживление экономики США и будет способствовать снижению нормы безработицы (т. е. доли безработных в общей численности рабочей силы). Ожидалось также, что возрастание темпов инфляции будет при этом не очень сильным.

Рассмотрим прежде всего диаграмму рассеяния для переменных ***UNJOB*** (процент безработных в общей численности рабочей силы) и ***INF*** (темп инфляции):

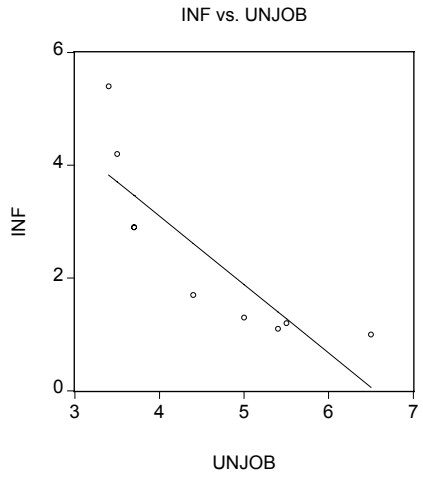


Облако рассеяния довольно округло, и это согласуется с весьма низким значением коэффициента детерминации $R^2 = 0.0864$, получаемым при подборе модели линейной зависимости *INF* от *UNJOB*.

Форма облака рассеяния не указывает и на какой-либо другой тип зависимости между этими двумя переменными на периоде наблюдений с 1958 по 1984 год.

В то же время, в период с 1961 по 1969 год наблюдалась следующая картина.

Год	1961	1962	1963	1964	1965	1966	1967	1968	1969
INF	1.0	1.1	1.2	1.3	1.7	2.9	2.9	4.2	5.4
UNJOB	6.5	5.4	5.5	5.0	4.4	3.7	3.7	3.5	3.4



Характер диаграммы рассеяния явно указывает на наличие нелинейной связи между рассматриваемыми переменными в период с 1961 по 1969 год (*кривая Филлипса*). Изображенная на диаграмме прямая, подобранная методом наименьших квадратов, очевидным образом не соответствует характеру статистических данных, хотя значение коэффициента детерминации $R^2 = 0.7184$ и представляется достаточно высоким. (Позднее мы сможем более квалифицированно говорить о том, действительно ли получаемое при подборе модели значение коэффициента детерминации достаточно велико.) В связи с этим, при подборе моделей к реальным статистическим данным следует обращать внимание не только на коэффициент детерминации, но и (**обязательно**!) на соответствие подобранной модели характеру статистических данных. Далее мы специально обсудим эту проблему, известную как ***проблема адекватности*** полученной модели имеющимся статистическм данным.

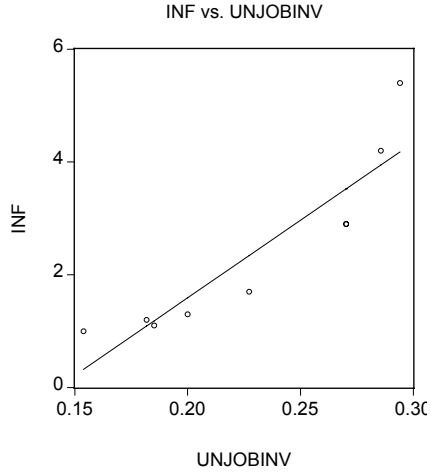
Поскольку, на первый взгляд, расположение точек напоминает график обратной пропорциональной зависимости, можно попробовать рассмотреть модель наблюдений

$INF_i = \alpha + \beta\left(1/UNJOB_i\right) + \varepsilon_i$, $i = 1,\text{K}\ ,n$,

соответствующую линейной связи между переменными INF и $UNJOBINV = 1/UNJOB$. Подбор такой связи приводит к модели

$INF = -3.90 + 27.47\ \left(1/UNJOB\right)$

с достаточно высоким коэффициентом детерминации $R^2 = 0.8307$. Однако, характер диаграммы рассеяния переменных INF и $UNJOBINV$



указывает на неадекватность и этой модели.

Обратившись еще раз к диаграмме рассеяния исходных переменных INF и $UNJOB$ (для данных за 1961—1969 годы), можно заметить, что кривая зависимости INF от $UNJOB$ по-видимому имеет вертикальную асимптоту $INF \cong 3$. Учесть

последнее обстоятельство можно в рамках *модели Michaelis-Menton*

$$INF = \frac{\theta_1 \cdot UNJOB}{\theta_2 + UNJOB} ,$$

которую можно преобразовать к виду

$$INF = \theta_1 - \frac{\theta_1 \cdot \theta_2 \cdot UNJOB}{\theta_2 + UNJOB} ,$$

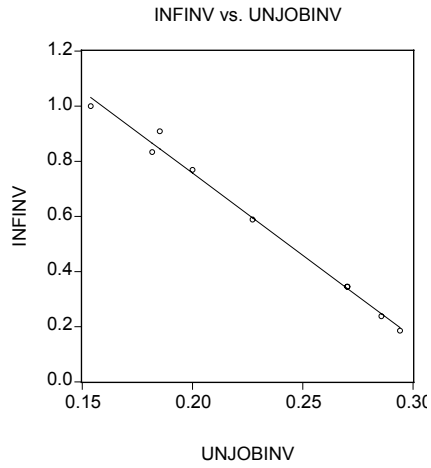
учитывающему наличие и вертикальной и горизонтальной асимптот. Такая модель связи линеаризуется переходом к обратным величинам $Y = 1/INF$, $X = 1/UNJOB$. Действительно, тогда

$$Y = \frac{1}{INF} = \frac{\theta_2 + UNJOB}{\theta_1 \cdot UNJOB} = \frac{1 + \left(\theta_2 / UNJOB\right)}{\theta_1}$$

$$= \frac{1}{\theta_1} + \frac{\left(\theta_2 / \theta_1\right)}{UNJOB} = \alpha + \beta X ,$$

где $\alpha = 1/\theta_1$, $\beta = \theta_2 / \theta_1$.

Диаграмма рассеяния для обратных величин $Y = 1/INF$, $X = 1/UNJOB$ имеет вид



Теперь уже точки на диаграмме рассеяния весьма хорошо следуют прямой линии, подобранной методом наименьших квадратов:

$INFINV = 1.947 - 5.952 \cdot UNJOBINV$,

$R^2 = 0.9914$. Здесь $\alpha = 1.947$, $\beta = -5.952$, так что $\theta_1 = 1/\alpha = 0.515$, $\theta_2 = \beta\,\theta_1 = -3.057$, и оцененная модель *Michaelis-Menton* имеет вид

$$INF = \frac{0.514 \cdot UNJOB}{-3.057 + UNJOB}\ .$$

Модель *Michaelis-Menton* хороша тем, что учитывает наличие асимптот и линеаризуется. С другой, стороны, она является лишь частным случаем более общей модели связи

$$INF = \theta_1 + \frac{\theta_3}{\theta_2 + UNJOB}$$

с тремя свободно изменяющимися параметрами. Действительно, в модели *Michaelis-Menton*

$\theta_3 = \theta_1 \cdot \theta_2$,

и она только двухпараметрическая, так что модель с тремя свободными параметрами является более гибкой. Но, вместе с тем, трехпараметрическая модель уже не линеаризуется, и параметры $\theta_1, \theta_2, \theta_3$ приходится оценивать, используя итерационную процедуру последовательного уменьшения суммы квадратов

$$Q(\theta_1, \theta_2, \theta_3) = \sum_{i=1}^{n} \left(INF_i - \theta_1 - \frac{\theta_3}{\theta_2 + UNJOB_i} \right)^2 .$$

(Конечно, в предположении аддитивности ошибок ε_i.) «Стартовые» значения параметров θ_1, θ_2 в этой процедуре можно взять близкими к оценкам θ_1, θ_2, полученным при оценивании предыдущей модели, например, $\theta_1 = 0.5,\ \theta_2 = -3.0$, а стартовое значение θ_3 можно положить равным 1.

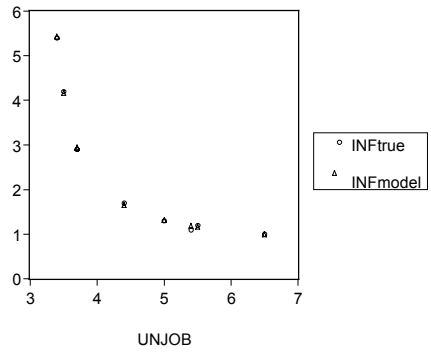
Реализация итерационной процедуры приводит к следующим оценкам параметров:

$\theta_1 = 0.581,\ \theta_2 = -3.117,\ \theta_3 = 1.370$;

при этом, $R^2 = 0.9992$. Оцененная модель имеет вид

$$INF = 0.581 + \frac{1.370}{UNJOB - 3.117} .$$

На следующей диаграмме показаны наблюдаемые значения переменной *INF* (*INFtrue*) и значения (*INFmodel*), получаемые по оцененной модели.



Подобранная модель показывает, что экспансионистские экономические мероприятия первоначально обеспечивают снижение нормы безработицы и реальный экономический рост при умеренной инфляции. Однако, удержать норму безработицы ниже ее естественного значения в течение продолжительного времени можно лишь за счет постоянно ускоряющегося темпа инфляции. К окончанию срока пребывания у власти Линдона Джонсона темп инфляции начал стремительно возрастать, что потребовало смены экономической политики.

Соответственно, наблюдать кривые Филлипса в указанном виде удается только на краткосрочных интервалах.

1.12. ЛИНЕЙНЫЕ МОДЕЛИ С НЕСКОЛЬКИМИ ОБЪЯСНЯЮЩИМИ ПЕРЕМЕННЫМИ

Рассмотрим статистические данные о потреблении текстиля (текстильных изделий) в Голландии в период между двумя мировыми войнами с 1923 по 1939 годы. В приведенной ниже таблице ***T*** — реальное потребление текстиля на душу населения, ***DPI*** — реальный располагаемый доход на душу населения, ***P —*** относительная цена текстиля. Все показатели выражены в индексной форме, в процентах к 1925 году.

Год	T	DPI	p

Год	T	DPI	p

1923	99.2	96.7	101.0
1924	99.0	98.1	100.1
1925	100.0	100.0	100.0
1926	111.6	104.9	90.6
1927	122.2	104.9	86.5
1928	117.6	109.5	89.7
1929	121.1	110.8	90.6
1930	136.0	112.3	82.8
1931	154.2	109.3	70.1

1932	153.6	105.3	65.4
1933	158.5	101.7	61.3
1934	140.6	95.4	62.5
1935	136.2	96.4	63.6
1936	168.0	97.6	52.6
1937	154.3	102.4	59.7
1938	149.0	101.6	59.5
1939	165.5	103.8	61.3

Для объяснения изменчивости потребления текстиля в указанном периоде мы можем привлечь в качестве объясняющей переменной как располагаемый доход *DPI*, так и относительную цену на текстильные изделия *P*. Если исходить из предположения о постоянстве эластичностей потребления текстиля по доходу и цене, то тогда следует подбирать линейные модели для логарифмов индексов, а не для самих индексов. Подбор таких моделей методом наименьших квадратов приводит к следующим результатам (использовались десятичные логарифмы):

$$\lg T = 1.442 + 0.348 \cdot \lg DPI\,, \quad R^2 = 0.0096,$$

$$ESS = 0.000959,\ RSS = 0.099185,\ TSS = 0.100144,\ R^2 = 0.0096;$$

$$\lg T = 3.564 - 0.770 \cdot \lg P,\ R^2 = 0.8760,$$

$$ESS = 0.087729,\ RSS = 0.012415,\ TSS = 0.100144,\ R^2 = 0.8760.$$

Вторая модель, несомненно, лучше описывает наблюдаемую динамику потребления текстиля. Однако, естественно возникает вопрос о том, нельзя ли для объяснения изменчивости переменной *T* использовать одновременно и располагаемый доход и относительную цену текстиля, улучшит ли это объяснение изменчивости потребления текстиля.

Чтобы привлечь для объяснения изменчивости потребления текстиля обе переменные *DPI* и *T*, мы рассматриваем ***модель линейной связи логарифмов*** этих величин

$$\lg T = \alpha + \beta \cdot \lg DPI + \gamma \cdot \lg P$$

и соответствующую ей ***модель наблюдений***

$$\lg T_i = \alpha + \beta \cdot \lg DPI_i + \gamma \cdot \lg P_i + \varepsilon_i , \quad i = 1,\text{K} ,n.$$

Оценки параметров α, β, γ можно опять находить методом наименьших квадратов, путем минимизации по всем возможным значениям α, β, γ суммы квадратов

$$Q(\alpha, \beta, \gamma) = \sum_{i=1}^{n} \left(\lg T_i - \alpha - \beta \ \lg DPI_i - \gamma \lg P_i \right)^2 .$$

Минимум этой суммы достигается на некотором наборе $\alpha = \hat{\alpha}, \ \beta = \hat{\beta}, \ \gamma = \hat{\gamma},$ так что

$$Q(\hat{\alpha}, \hat{\beta}, \hat{\gamma}) = \min_{\alpha, \beta, \gamma} Q(\alpha, \beta, \gamma) .$$

Это минимальное значение мы опять обозначаем

$$RSS = \sum_{i=1}^{n} \left(\lg T_i - \hat{\alpha} - \hat{\beta} \lg DPI_i - \hat{\gamma} \lg P_i \right)^2$$

и называем остаточной суммой квадратов.

Коэффициент детерминации R^2 определяется, как и в модели связи между двумя переменными:

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

Здесь

$$TSS = \sum_{i=1}^{n} \left(\lg T_i - \overline{\lg T} \right)^2 ,$$

$$RSS = \sum_{i=1}^{n} \left(\lg T_i - \hat{\lg T_i} \right)^2 ,$$

где

$$\overline{\lg T} = \frac{1}{n}\sum_{i=1}^{n} \lg T_i \ ,$$

$$\hat{\lg T_i} = \hat{\alpha} + \hat{\beta} \cdot \lg DPI_i \ + \hat{\gamma} \cdot \lg P_i \ , \quad i = 1,\text{K} \ ,n.$$

При этом,

$$TSS = RSS + ESS,$$

где

$$ESS = \sum_{i=1}^{n} \left(\hat{\lg T_i} - \overline{\lg T} \right)^2 \ ,$$

так что

$$R^2 = \frac{ESS}{TSS}$$

(и опять, разложение $TSS = RSS + ESS$ справедливо только при включении постоянной составляющей α в правую часть соотношения, определяющего линейную модель связи). При этом также

$$R^2 = r^2_{\lg T, \hat{\lg T}} \ ,$$

т. е. коффициент детерминации R^2 равен квадрату (обычного) выборочного коэффициента корреляции между переменными $\lg T$ и $\hat{\lg T}$.

Разности

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

называются ***остатками***.

По поводу получения явных выражений для оценок наименьших квадратов мы поговорим несколько позднее, а сейчас просто приведем результаты оценивания для нашего примера:

$$\lg T = 1.374 + 1.143 \cdot \lg DPI - 0.829 \cdot \lg P \ ,$$

$ESS = 0.097577$, $RSS = 0.02567$, $R^2 = 0.9744$.

Мы видим, что в результате привлечения для объяснения изменчивости потребления текстиля сразу двух показателей DPI и P произошло заметное увеличение коэффициента детерминации по сравнению с лучшей из двух моделей, использовавших только один показатель — от значения 0.8760 до значения 0.9744.

Коэффициент 1.143 в подобранной модели связи интерпретируется здесь как эластичность потребления текстиля по доходу ***при неизменном значении относительной цены*** P на текстиль, а коэффициент -0.829 — как эластичность потребления текстиля по относительным ценам ***при неизменном уровне дохода***. Такие значения коэффициентов говорят в пользу того, что потребление текстиля эластично по доходам и неэластично по ценам. Вопрос о том, в какой степени можно доверять подобным заключениям, мы рассмотрим далее в контексте вероятностных моделей.

ЧАСТЬ 2. СТАТИСТИЧЕСКИЕ ВЫВОДЫ ПРИ СТАНДАРТНЫХ ПРЕДПОЛОЖЕНИЯХ О ВЕРОЯТНОСТНОЙ СТРУКТУРЕ ОШИБОК В ЛИНЕЙНОЙ МОДЕЛИ НАБЛЮДЕНИЙ

2.1. ВЕРОЯТНОСТНОЕ МОДЕЛИРОВАНИЕ ОШИБОК

Мы уже неоднократно сталкивались с вопросом о том, сколь существенно величина коэффициента корреляции (детерминации) должна отличаться от нуля, чтобы можно было говорить о действительно существующей линейной связи между исследуемыми переменными.

Если оцененное значение эластичности потребления некоторого товара оказалось несколько больше единицы, то возникает вопрос о том, сколь надежным является заключение о том, что потребление этого товара эластично по ценам.

Если мы будем использовать подобранную прямую

$$y = \hat{\alpha} + \hat{\beta}\, x$$

для прогнозирования значений y_i для новых наблюдений x_i, $t = n+1,...,n+k$, то сколь надежными будут такие прогнозы?

Если у нас нет теоретических (экономических) оснований для выбора между моделью в уровнях переменных и моделью в логарифмах уровней, то как выбрать одну из этих моделей на основании одних только наблюдений?

Ответы на эти и другие подобные вопросы невозможны, если мы не сделаем некоторых более или менее подробных предположений о структуре последовательности ошибок $\varepsilon_1, \ldots, \varepsilon_n$, участвующих в определении модели наблюдений

$$\boxed{y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \ldots, n.}$$

Базовая, и наиболее простая модель для последовательности $\varepsilon_1,\text{K},\varepsilon_n$ предполагает, что $\varepsilon_1,\text{K},\varepsilon_n$ — ***независимые случайные величины, имеющие одинаковое распределение*** **(*i. i. d. — independent, identically distributed random variables*)**.

Для нас (*пока*!) достаточно представлять случайную величину Z как переменную величину, такую, что до наблюдения ее значения невозможно предсказать это значение абсолютно точно, и, в то же время, для любого z, $-\infty \le z \le \infty$, определена ***вероятность***

$$\boxed{F(z) = \text{P}\{Z \le z\}}$$

того, что наблюдаемое значение переменной Z не превзойдет z; $0 \le F(z) \le 1$. Функция $F(z)$, $-\infty \le z \le \infty$, называется ***функцией распределения*** случайной величины Z (**c. d. f. — *cumulative distribution function***).

Говоря об ошибках $\varepsilon_1,\text{K},\varepsilon_n$ как о случайных величинах, мы, соответственно, понимаем указанную линейную модель наблюдений таким образом, что

а) существует (теоретическая, объективная или в виде тенденции) линейная зависимость значений переменной y от значений переменной x с вполне определенными, хотя обычно и не известными исследователю, значениями параметров α и β;

б) эта линейная связь для реальных статистических данных не является строгой: наблюдаемые значения y_i переменной y отклоняются от значений $\tilde{y}_i$, указываемых моделью линейной связи

$\tilde{y}_i = \alpha + \beta x_i$, $i = 1,\text{K},n$;

в) при заданных (известных) значениях x_i конкретные значения отклонений

$\varepsilon_i = y_i - \tilde{y}_i$, $i = 1,\text{K},n$,

не могут быть точно предсказаны до наблюдения значений y_i даже если значения параметров α и β известны точно;

г) для каждого z, $-\infty \le z \le \infty$, определена вероятность $F(z)$ того, что наблюдаемое значение отклонения ε_i не превзойдет z, причем эта вероятность не зависит от номера наблюдения;

д) вероятность того, что наблюдаемое значение отклонения ε_i в i-м наблюдении не превзойдет z, не зависит от того, какие именно значения принимают отклонения в остальных $n-1$ наблюдениях.

В дальнейшем, говоря о той или иной случайной величине Z, мы будем предполагать существование функции $p(z)$, $-\infty \le z \le \infty$, принимающей только неотрицательные значения и такой, что

1) площадь под кривой

$v = p(z)$

в прямоугольной системе координат zOv (точнее, площадь, ограниченная сверху этой кривой и снизу — горизонтальной осью Oz) равна 1,

2) для любой пары значений z_1, z_2 с $z_1 < z_2$, вероятность

$\mathrm{P}\{z_1 < Z \le z_2\}$

численно равна площади, ограниченной снизу осью Oz, сверху — кривой $v = p(z)$, слева — вертикальной прямой $z = z_1$, справа — вертикальной прямой $z = z_2$ (т. е. равна части площади под кривой $v = p(z)$, расположенной между точками $z = z_1$ и $z = z_2$).

3) для любого z_0, $-\infty \le z_0 \le \infty$, вероятность $F(z_0)$ того, что наблюдаемое значение Z не превзойдет z_0, равна площади, ограниченной снизу осью Oz, сверху — кривой $v = p(z)$ и

справа — вертикальной прямой $z = z_0$, т. е. равна части площади под кривой $v = p(z)$, расположенной левее точки $z = z_0$.

Заметим, что при этом выполняется следующее важное соотношение:

$$\boxed{\mathrm{P}\{z_1 < Z \le z_2\} = F(z_2) - F(z_1)\ .}$$

(Действительно, вероятность $F(z_2)$ численно равна части площади под кривой $v = p(z)$, расположенной левее точки $z = z_2$, а эта часть складывается из части площади под кривой, расположенной левее точки $z = z_1$ и части площади под кривой, расположенной между точками $z = z_1$ и $z = z_2$, так что

$$F(z_2) = F(z_1) + \mathrm{P}\{z_1 < Z \le z_2\}\ ,$$

откуда и следует заявленное соотношение.) Кроме того,

$$\boxed{\mathrm{P}\{Z > z\} = 1 - F(z)\ .}$$

(Действительно,

$$F(z) + \mathrm{P}\{Z > z\} = 1\ ,$$

поскольку слева складываются части площади под кривой $v = p(z)$, расположенные, соответственно, левее и правее точки z, так что в сумме они составляют всю площадь под этой кривой, а вся площадь под кривой $v = p(z)$ как раз и равна 1.)

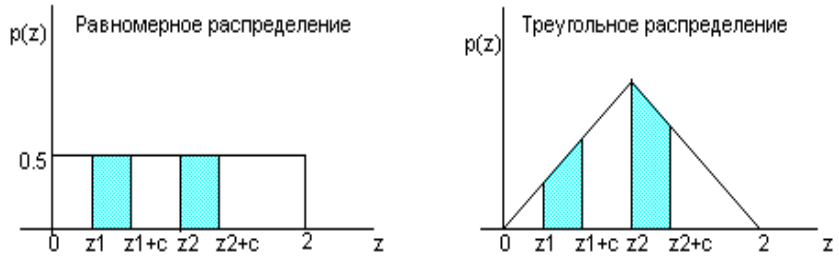
Функция $p(z)$ связана с функцией распределения случайной величины Z соотношениями

$$\boxed{p(z) = \frac{dF(z)}{dz}\ ,\quad F(z) = \int_{-\infty}^{z} p(t)dt}$$

и называется ***функцией плотности вероятности*** случайной величины Z (**p.d.f. — *probability density function***). Для краткости, мы часто будем говорить о функции $p(z)$ как о

функции плотности или о ***плотности распределения*** случайной величины Z.

Возьмем два непересекающихся интервала значений переменной z: $z_1 \le z \le z_1 + c$ и $z_2 \le z \le z_2 + c$. Рассмотрим два варианта распределения вероятности случайной величины Z: ***равномерное распределение*** на отрезке $0 \le z \le 2$ и ***треугольное распределение*** на том же отрезке. Графики функций плотности для этих двух вариантов имеют следующий вид:



Площади заштрихованных прямоугольников на первом графике численно равны вероятностям того, что случайная величина Z, имеющая равномерное распределение на отрезке $0 \le z \le 2$, примет значения в пределах $z_1 \le z \le z_1 + c$ и $z_2 \le z \le z_2 + c$, соответственно. Поскольку основания и высоты этих прямоугольников равны, то равны и их площади, т.е. равны указанные вероятности.

Площади заштрихованных трапеций на втором графике численно равны вероятностям того, что случайная величина Z, имеющая треугольное распределение на отрезке $0 \le z \le 2$, примет значения в пределах $z_1 \le z \le z_1 + c$ и $z_2 \le z \le z_2 + c$, соответственно. Высоты этих трапеций равны, однако стороны трапеции, расположенной правее, больше сторон трапеции, расположенной левее. Поэтому и площадь трапеции, расположенной правее, больше площади трапеции, расположенной

левее. А это означает, в свою очередь, что вероятность того, что случайная величина Z, имеющая треугольное распределение на отрезке $0 \le z \le 2$, примет значения в пределах $z_2 \le z \le z_2 + c$, больше вероятности того, что эта случайная величина Z примет значения в пределах $z_1 \le z \le z_1 + c$.

Таким образом, функция плотности указывает на более вероятные и менее вероятные интервалы значений случайной величины. Если случайная величина Z имеет равномерное распределение на отрезке $0 \le z \le 2$, то для нее все интервалы значений, имеющие одинаковую длину и расположенные целиком в пределах отрезка $0 \le z \le 2$, имеют одинаковые вероятности (т. е. вероятности попадания значений случайной величины на эти интервалы одинаковы). Если же случайная величина Z имеет треугольное распределение на отрезке $0 \le z \le 2$, то для нее интервалы значений, имеющие одинаковую длину и расположенные целиком в пределах отрезка $0 \le z \le 2$, имеют, вообще говоря, различные вероятности: вероятность того, что случайная величина примет значение в интервале, расположенном ближе к центральному значению $z = 2$, больше вероятности того, что случайная величина примет значение в интервале, расположенном ближе к одному из концов отрезка $0 \le z \le 2$.

Обсудим несколько более точно вопрос о том, что мы понимаем под независимостью нескольких случайных величин. Пусть мы имеем n случайных величин Z_1, Z_2, K, Z_n, имеющих одинаковую функцию распределения $F(z)$. Мы говорим, что эти случайные величины ***независимы в совокупности,*** если для любого набора пар $a_1 < b_1$, $a_2 < b_2, ..., a_n < b_n$, где a_i и b_i могут быть равны также $-\infty$ и $+\infty$,

$$\boxed{\begin{aligned}&\mathrm{P}\{a_1 < Z_1 \le b_1,\ a_2 < Z_2 \le b_2, \mathrm{K}\ ,\ a_n < Z_n \le b_n\} = \\ &\mathrm{P}\{a_1 < Z_1 \le b_1\} \cdot \mathrm{P}\{a_2 < Z_2 \le b_2\} \Lambda\ \mathrm{P}\{a_n < Z_n \le b_n\}\ .\end{aligned}}$$

При таком предположении ***условная*** вероятность того, что, например, $a_n < Z_n \le b_n$, при условии, что $a_1 < Z_1 \le b_1$, κ, $a_{n-1} < Z_{n-1} \le b_{n-1}$, равна ***безусловной*** вероятности того, что $a_n < Z_n \le b_n$, т. е. вероятности, вычисляемой без задания указанногоусловия:

$$\boxed{\begin{aligned}&\mathrm{P}\{a_n < Z_n \le b_n \mid a_1 < Z_1 \le b_1, \mathrm{K}\ ,\ a_{n-1} < Z_{n-1} \le b_{n-1}\} \\ &\qquad = \mathrm{P}\{a_n < Z_n \le b_n\}\ .\end{aligned}}$$

(Вертикальная черта в этой формуле указывает на то, что первая вероятность — условная; справа от вертикальной черты записано условие, при котором вычисляется эта вероятность.) Иначе говоря, на распределение вероятности случайной величины Z_n не влияет информация о значениях случайных величин $Z_1, Z_2, \mathrm{K}\ , Z_{n-1}$. И вообще, на распределение вероятностей случайной величины Z_j не влияет информация о значениях случайных величин Z_k с $k \ne j$.

Если случайные величины $Z_1, Z_2, \mathrm{K}\ , Z_n$ имеют одинаковое распределение F (заданное или функцией распределения или функцией плотности) и независимы в совокупности, то часто это обозначают в записи следующим образом:

$Z_1, \mathrm{K}\ , Z_n$ - $i.i.d.$, $Z_i \sim F$.

Возвращаясь к модели наблюдений

$$\boxed{y_i = \alpha + \beta x_i + \varepsilon_i\ ,\ \ i = 1, \mathrm{K}\ , n\ ,}$$

и предполагая, что $\varepsilon_1, \mathrm{K}\ , \varepsilon_n$ — ***независимые случайные величины, имеющие одинаковое распределение*** **(*i. i. d*),** мы должны теперь сделать еще и предположение о том, каким

именно является это одинаковое для всех $\varepsilon_1,\text{K},\varepsilon_n$ распределение.

2.2. ГАУССОВСКОЕ (НОРМАЛЬНОЕ) РАСПРЕДЕЛЕНИЕ ОШИБОК В ЛИНЕЙНОЙ МОДЕЛИ НАБЛЮДЕНИЙ

Итак, предположив, что в модели наблюдений

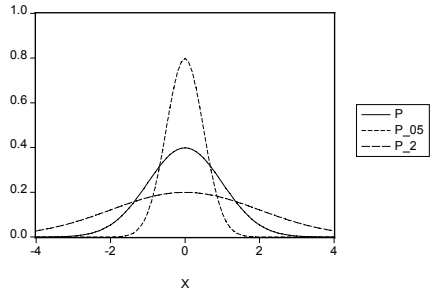
$$y_i = \alpha + \beta x_i + \varepsilon_i\ ,\ \ i = 1,\text{K}\ ,n\ ,$$

ошибки $\varepsilon_1,\text{K}\ ,\varepsilon_1$ — ***независимые случайные величины, имеющие одинаковое распределение*** **(*i. i. d*),** мы должны сделать и предположение о том, каким именно является это распределение.

Классические методы статистического анализа линейных моделей наблюдений предполагают, что таковым является ***распределение Гаусса*** **(*Gaussian distribution*)**, функция плотности которого имеет вид

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/\left(2\sigma^2\right)}\ ,\quad -\infty < x < +\infty\ .$$

График указанной функции плотности имеет колоколообразную форму



Параметр $\sigma > 0$ характеризует степень рассредоточения распределения вдоль оси абсцисс. На диаграмме представлены графики функций плотности гауссовского распределения при

трех различных значениях параметра σ: $\sigma = 1$, $\sigma = 0.5$, $\sigma = 2$. Из трех представленных функций наибольшее значение в нуле имеет функция плотности с $\sigma = 0.5$, наименьшее — функция плотности с $\sigma = 2$, а промежуточное между ними — функция плотности с $\sigma = 1$. Эти значения равны, соответственно,

$2/\sqrt{2\pi} = 0.7979$, $1/\sqrt{2\pi} = 0.3989$, $1/\left(2\sqrt{2\pi}\right) = 0.1995$.

Гауссовское распределение симметрично относительно нуля, и это предполагает, что положительные ошибки столь же вероятны, как и отрицательные; при этом, малые ошибки встречаются чаще, чем большие. Если случайная ошибка имеет гауссовское распределение с параметром σ, то с вероятностью 0.95 ее значение будет заключено в пределах от -1.96σ до $+1.96\sigma$. Соответственно, для трех рассмотренных случаев получаем: с вероятностью 0.95 значение случайной ошибки заключено в интервале

$(-0.98, 0.98)$ — при $\sigma = 0.5$, $(-1.96, 1.96)$ - при $\sigma = 1$, $(-3.92, 3.92)$ - при $\sigma = 2$.

Хотя гауссовское распределение довольно часто вполне приемлемо для описания случайных ошибок в моделях наблюдений, оно вовсе не является универсальным. Такое распределение характерно для ситуаций, когда результирующая ошибка является следствием сложения большого количества независимых случайных ошибок, каждая из которых достаточно мала.

Мы будем далее в этом параграфе предполагать, что ***процесс порождения данных*** **(*ППД***, или ***DGP- data generating process*****)** устроен следующим образом. Значения x_1, K, x_n известны точно и рассматриваются как заданные, а значения y_1, K, y_n получаются наложением на значения $\alpha + \beta x_i$ случайных ошибок ε_i.

В этом контексте, $\alpha + \beta\, x_i$ рассматриваются как некоторые постоянные (хотя и не известные наблюдателю). Напротив, значения y_i носят случайный характер, определяемый случайным характером значений ε_i. Собственно, y_i отличается от случайной величины ε_i лишь сдвигом на постоянную $\alpha + \beta\, x_i$, и потому также является случайной величиной. Мы будем обозначать ее в этом качестве как случайную величину Y_i. Функция распределения этой случайной величины имеет вид

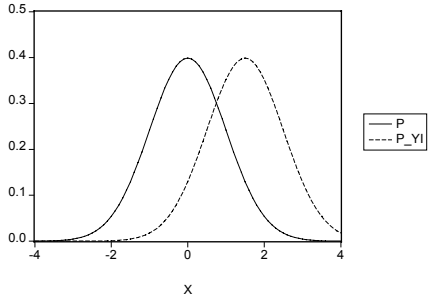
$$F_{Y_i}(y) = P\{Y_i \le y\} = P\{\alpha + \beta\, x_i + \varepsilon_i \le y\}$$
$$= P\{\varepsilon_i \le y - (\alpha + \beta\, x_i)\} = F(y - \alpha - \beta\, x_i)\,,$$

где F — функция распределения случайной величины ε_i (одинаковая для всех $\varepsilon_1, \text{K}, \varepsilon_n$). Соответственно, функция плотности распределения случайной величины Y_i имеет вид

$$p_{Y_i}(y) = \frac{dF_{Y_i}(y)}{dy} = \frac{dF(y - \alpha - \beta\, x_i)}{dy} = p(y - \alpha - \beta\, x_i)\,,$$

где p — функция плотности распределения случайной величины ε_i.

Таким образом, случайные величины Y_1, K, Y_n хотя и являются взаимно независимыми (в силу предполагаемой взаимной независимости случайных величин $\varepsilon_1, \text{K}, \varepsilon_n$), но имеют разные распределения, отличающиеся сдвигом. На следующем рисунке представлены графики функции плотности $p(x)$ распределения ε_i (гауссовское распределение с параметром $\sigma = 1$) и функции плотности $p_{Y_i}(x)$ распределения случайной величины $Y_i = \alpha + \beta\, x_i + \varepsilon_i$ при значении $\alpha + \beta\, x_i = 1.5$.



Заметим, что если случайная ошибка ε_i имеет гауссовское распределение с плотностью

$$p(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-y^2/(2\sigma^2)}, \quad -\infty < y < +\infty,$$

то отличающаяся от нее сдвигом случайная величина $Y_i = \alpha + \beta\, x_i + \varepsilon_i$ имеет функцию плотности

$$p_{Y_i}(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-\alpha-\beta x_i)^2/(2\sigma^2)}, \quad -\infty < y < +\infty.$$

Эта функция плотности принадлежит двухпараметрическому семейству функций плотности вида

$$p(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-\mu)^2/(2\sigma^2)}, \ -\infty < y < +\infty;\ \sigma > 0,\ -\infty < \mu < +\infty.$$

Функции плотности такого вида называются ***нормальными плотностями***, а определяемые ими распределения вероятностей называются ***нормальными распределениями вероятностей***. Если некоторая случайная величина Y имеет плотность распределения, заданную последним соотношением, то говорят, что ***случайная величина Y имеет нормальное распределение с параметрами μ и σ^2***. Распределение такой случайной величины симметрично относительно своего ***среднего*** значе-

ния μ. Максимальное значение функции плотности этой случайной величины достигается при $y = \mu$.

Таким образом, строго говоря, гауссовское распределение — это нормальное распределение с нулевым средним значением. Однако, в современной научной литературе термины ***нормальное распределение*** и ***гауссовское распределение*** используются как синонимы: ***нормальное распределение с параметрами μ и σ^2*** называют также ***гауссовским распределением с параметрами μ и σ^2.***

Важнейшая роль предположения о нормальном (гауссовском) распределении ошибок в линейной модели наблюдений

$$y_i = \alpha + \beta x_i + \varepsilon_i\ ,\ \ i = 1,\text{K}\ ,n\ ,$$

определяется тем обстоятельством, что при добавлении такого предположения к стандартному предположению о том, что ошибки $\varepsilon_1,\text{K}\ ,\varepsilon_n$ — независимые случайные величины, имеющие одинаковое распределение, можно легко найти точный вид распределения оценок наименьших квадратов для неизвестных значений параметров модели.

Вспомним, в этой связи, полученное ранее выражение

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

Обозначая

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

мы можем записать выражение для $\hat{\beta}$ в виде

$$\hat{\beta} = \sum_{i=1}^{n} w_i \left(y_i - \bar{y} \right) = \sum_{i=1}^{n} w_i y_i - \bar{y} \sum_{i=1}^{n} w_i$$

$$= \sum_{i=1}^{n} w_i y_i - \bar{w} \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \left(w_i - \bar{w} \right) y_i = \sum_{i=1}^{n} c_i y_i \ ,$$

где

$c_i = w_i - \bar{w}$.

Таким образом,

$$\boxed{\hat{\beta} = \sum_{i=1}^{n} c_i y_i \ ,}$$

где c_1, K, c_n — фиксированные величины, а y_1, K, y_n — ***наблюдаемые значения случайных величин*** Y_1, K, Y_n. Поэтому вычисленное по последней формуле значение $\hat{\beta}$ является ***наблюдаемым значением случайной величины***

$$\boxed{\hat{\beta} = \sum_{i=1}^{n} c_i Y_i \ ,}$$

которая является линейной комбинацией случайных величин Y_1, K, Y_n и имеет некоторое распределение вероятностей, зависящее от распределения последних.

В общем случае, аналитическое описание распределения $\hat{\beta}$ как случайной величины довольно затруднительно. Более просто эта задача решается в ситуации, когда ε_i имеет гауссовское распределение. Если ошибки $\varepsilon_1, \mathrm{K}, \varepsilon_n$ - независимые случайные величины, имеющие одинаковое нормальное распределение с нулевым средним, то тогда оценка наименьших квадратов $\hat{\beta}$ параметра β также имеет нормальное распределение. Чтобы указать параметры этого нормального распределения и иметь возможность проводить статистический анализ

подобранной модели линейной связи между переменными факторами, нам придется уделить внимание некоторым важным числовым характеристикам случайных величин и их свойствам.

2.3. ЧИСЛОВЫЕ ХАРАКТЕРИСТИКИ СЛУЧАЙНЫХ ВЕЛИЧИН И ИХ СВОЙСТВА

Случайные величины, с которыми мы имеем дело в данном курсе, полностью определяются заданием их функции плотности, указывающей на зоны более вероятных и менее вероятных значений случайной величины. Часто, однако, интересуются более сжатыми характеристиками распределений случайных величин, выраженными отдельными числами. К таким характеристикам, в первую очередь, относятся ***математическое ожидание*** и ***дисперсия*** случайной величины.

Пусть случайная величина X имеет функцию плотности $p(x)$. График функции $p(x)$ ограничивает вместе с осью абсцисс Ox полосу переменной ширины. Если рассматривать эту полосу как материальный объект определенной (постоянной) толщины, изготовленный из однородного материала и имеющий массу, равную единице, то абсцисса центра тяжести этого материального объекта называется ***математическим ожиданием*** (***expectation***) ***случайной величины X***, обозначается ***E*** **(*X*)** и вычисляется по формуле

$$\boxed{E(X) = \int_{-\infty}^{\infty} x\, p(x) dx\,.}$$

Если график функции плотности симметричен относительно оси ординат (так что $p(x)$ — четная функция), то $E(X) = 0$.

Довольно часто о $E(X)$ говорят как о ***среднем значении случайной величины X***. Это связано с тем, что если X_1,K,X_n — ***независимые копии*** случайной величины X (т. е. случайные величины X_1,K,X_n независимы в совокупности и имеют то же распределение, что и X), то тогда при больших n для наблюдаемых значений x_1,K,x_n случайных величин X_1,K,X_n имеет место приближенное равенство

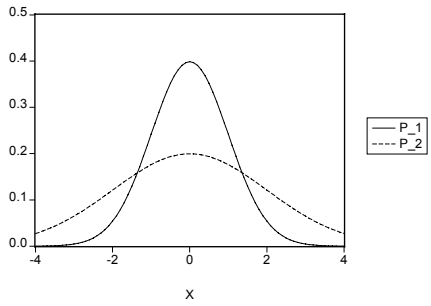
$$\boxed{\tfrac{1}{n}\left(x_1+\text{K}+x_n\right)\cong E(X)\,,}$$

тем более точное, чем больше значение n. Иными словами, с увеличением n значение $E(X)$ сколь угодно точно приближается значением среднеарифметического наблюдаемых величин x_1,K,x_n.

Обратимся опять к упомянутому ранее гауссовскому (нормальному) распределению с функцией плотности

$$\boxed{p(x)=\frac{1}{\sigma\sqrt{2\pi}}\;e^{-x^2/\left(2\sigma^2\right)}}$$

и пусть случайная величина X_1 имеет такое распределение с $\sigma=1$, а случайная величина X_2 имеет такое распределение с $\sigma=2$. Сравним графики соответствующих функций плотности (сплошной линией представлен график функции плотности случайной величины X_1):



Поскольку в обоих случаях графики симметричны относительно нуля, то

$$E(X_1) = E(X_2) = 0\,,$$

т. е. математические ожидания случайных величин X_1 и X_2 совпадают. Однако, распределение случайной величины X_2 более рассредоточено, и это означает, что для любого $a > 0$

$$P\{|X_1| > a\} < P\{|X_2| > a\}\,.$$

При этом говорят, что распределение случайной величины X_2 имеет ***более тяжелые*** (***heavy***), или ***более длинные*** (***long***) ***хвосты*** (***tails***). Соответственно,

$$P\{|X_1| \le a\} = 1 - P\{|X_1| > a\} > 1 - P\{|X_2| > a\} = P\{|X_2| \le a\}\,.$$

В рассмотренном случае в качестве числовой характеристики степени рассредоточенности распределения можно было бы принять параметр σ: чем больше значение этого параметра, тем более рассредоточено распределение. В общем случае, сравнивать степени рассредоточенности распределений случайных величин можно, привлекая для этой цели понятие дисперсии.

Дисперсией (***variance***) ***случайной величины X*** называют число

$$D(X) = E(X - E(X))^2,$$

равное математическому ожиданию квадрата отклонения случайной величины X от ее математического ожидания $E(X)$. [1] Зная функцию плотности $p(x)$ случайной величины X, дисперсию этой случайной величины можно вычислить по формуле

$$D(X) = \int_{-\infty}^{\infty} \big(x - E(X)\big)^2 p(x)dx .$$

Таким образом, математическое ожидание $E(X)$ можно интерпретировать как взвешенное среднее возможных значений x случайной величины X, с весами, пропорциональными $p(x)$, а дисперсию $D(X)$ — как взвешенное среднее (с теми же весами) квадратов отклонений возможных значений x случайной величины X от ее математического ожидания.

Если случайная величина X имеет ***нормальное распределение*** с функцией плотности

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-\mu)^2/(2\sigma^2)},$$

то для нее

$$E(X) = \mu \ , \ D(X) = \sigma^2 .$$

Таким образом, случайная величина, имеющая нормальное распределение, полностью определяется (в отношении ее распределения) заданием значений ее математического ожидания и дисперсии.

В связи с частым использованием нормально распределенных случайных величин в дальнейшем изложении, мы будем

---

[1] В литературе по эконометрике математическое ожидание случайной величины X обозначают иногда символом $M(X)$, а для дисперсии случайной величины X используют также обозначения $Var(X)$ и $V(X)$.

обозначать нормальное распределение, имеющее математическое ожидание μ и дисперсию σ^2, символом $N(\mu,\sigma^2)$. В случае, когда $\mu = 0, \sigma^2 = 1$, говорят о ***стандартном нормальном распределении*** $N(0,1)$. Имеются весьма подробные таблицы значений функции распределения и функции плотности стандартного нормального распределения.

Для дальнейшего нам, в первую очередь, понадобятся следующие простые ***свойства математического ожидания и дисперсии***.

Если a - некоторая постоянная, отличная от нуля, а X - некоторая случайная величина, то тогда сумма $X + a$ и произведение aX также являются случайными величинами; при этом,

$$\boxed{\begin{array}{l} E(X+a) = E(X) + a \\ E(aX) = aE(X) \end{array}} \quad \boxed{\begin{array}{l} D(X+a) = D(X) \\ D(aX) = a^2 D(X). \end{array}}$$

Два свойства, касающиеся математического ожидания, непосредственно следуют из определения математического ожидания. При выводе первого из них учитываем, что по самому определению функции плотности распределения,

$$\int_{-\infty}^{\infty} p(x)dx = 1 .$$

Из этих двух свойств математического ожидания легко получаем указанные два свойства дисперсии. Действительно,

$$\begin{aligned} D(X+a) &= E\Big(\big(X+a\big) - E\big(X+a\big)\Big)^2 \\ &= E(X+a-E(X)-a)^2 = E(X-E(X))^2 = D(X) , \\ D(aX) &= E\Big(aX - E\big(aX\big)\Big)^2 = E\Big(aX - aE\big(X\big)\Big)^2 \\ &= E\Big(a^2\big(X-E(X)\big)^2\Big) = a^2 E\Big(X - E\big(X\big)\Big)^2 = a^2 D(X) . \end{aligned}$$

Таким образом, изменение случайной величины на некоторую постоянную вызывает такое же изменение математического ожидания, но не отражается на дисперсии. Изменение случайной величины в a раз приводит к такому же изменению математического ожидания и изменяет значение дисперсии в a^2 раз.

В применении к линейной модели наблюдений

$$\boxed{y_i = \alpha + \beta\, x_i + \varepsilon_i\,, \quad i = 1,\text{K}\,,n,}$$

с фиксированными $x_1,\text{K}\,,x_n$ и взаимно независимыми гауссовскими ошибками $\varepsilon_1,\text{K}\,,\varepsilon_n$, мы имеем:

$$\varepsilon_i \sim N\left(0,\sigma^2\right) \;\Rightarrow\; Y_i = \alpha + \beta\, x_i + \varepsilon_i \;\sim\; N\left(\alpha + \beta\, x_i, \sigma^2\right).$$

Соответственно,

$$\boxed{E\left(\varepsilon_i\right) = 0,\; D(\varepsilon_i) = \sigma^2;\;\; E(Y_i) = \alpha + \beta\, x_i\,,\; D(\varepsilon_i) = \sigma^2.}$$

Заметим, наконец, что если $Z_1,\text{K}\,,Z_n$ — случайные величины и $Z = Z_1 + \text{K} + Z_n$, то

$$\boxed{E\left(Z\right) = E\left(Z_1\right) + \text{K} + E\left(Z_n\right)}$$

и если случайные величины $Z_1,\text{K}\,,Z_n$ ***попарно некоррелированы***, т. е.

$$\boxed{Cov\left(Z_j, Z_k\right) = E\left(\left(Z_j - E(Z_j)\right)\left(Z_k - E(Z_k)\right)\right) = 0\,,}$$

то тогда

$$\boxed{D\left(Z\right) = D\left(Z_1\right) + \text{K} + D\left(Z_n\right).}$$

В применении к последней линейной модели наблюдений это означает, что рассматриваемая как случайная величина оценка наименьших квадратов $\vec{\beta}$, которую мы представили ранее в виде

$$\boxed{\vec{\beta} = \sum_{i=1}^{n} c_i Y_i\,,}$$

где

$c_i = w_i - \overline{w}$,

$$w_i = \frac{x_i - \overline{x}}{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2} ,$$

так что c_1, K, c_n — фиксированные величины, имеет нормальное распределение с математическим ожиданием

$$\boxed{E(\overline{\beta}) = \sum_{i=1}^{n} c_i E(Y_i)}$$

и дисперсией

$$\boxed{D(\overline{\beta}) = \sum_{i=1}^{n} c_i^2 D(Y_i) .}$$

2.4. НОРМАЛЬНЫЕ ЛИНЕЙНЫЕ МОДЕЛИ С НЕСКОЛЬКИМИ ОБЪЯСНЯЮЩИМИ ПЕРЕМЕННЫМИ

Начиная с этого момента, мы будем предполагать, что

(1) ***Модель наблюдений*** имеет вид

$$\boxed{y_i = \theta_1 x_{i1} + \text{K} + \theta_p x_{ip} + \varepsilon_i , \;\; i = 1, \text{K}, n, \;\; n \geq p,}$$

где y_i - значение объясняемой переменной в i-м наблюдении;

x_{ij} - известное значение j-ой объясняющей переменной в i-м наблюдении;

θ_j - неизвестный коэффициент при j-ой объясняющей переменной;

ε_j - случайная составляющая (“ошибка“) в i-м наблюдении.

(2) $\varepsilon_1, \text{K}, \varepsilon_n$ - ***случайные величины***, ***независимые в совокупности***, имеющие ***одинаковое нормальное распределе-***

ние $N(0,\sigma^2)$ с нулевым математическим ожиданием и дисперсией $\sigma^2 > 0$.

(3) Если не оговорено противное, то в число объясняющих переменных ***включается переменная, тождественно равная единице***, которая объявляется ***первой*** объясняющей переменной, так что

$$x_{i1} \equiv 1, \quad i = 1,\text{K},n.$$

При сделанных предположениях y_1,K,y_n являются наблюдаемыми значениями нормально распределенных случайных величин Y_1,K,Y_n, которые независимы в совокупности и для которых

$$E(Y_i) = \theta_1 x_{i1} + \text{K} + \theta_p x_{ip}, \qquad D(Y_i) = \sigma^2,$$

так что

$$Y_i \sim N(\theta_1 x_{i1} + \text{K} + \theta_p x_{ip}, \sigma^2), \quad i = 1,\text{K},n.$$

В отличие от $\varepsilon_1,\text{K},\varepsilon_n$, случайные величины Y_1,K,Y_n имеют распределения, отличающиеся сдвигами.

Определенную указанным образом модель наблюдений мы будем называть ***нормальной линейной моделью*** с p объясняющими переменными. Иначе ее еще называют ***нормальной линейной моделью множественной регрессии переменной y на переменные $x_1, \ldots, x_p$***. Термин “множественная” указывает на использование в правой части модели наблюдений двух и более объясняющих переменных, отличных от постоянной. Термин “регрессия” имеет определенные исторические корни и используется лишь в силу традиции.

Оценивание неизвестных коэффициентов модели ***методом наименьших квадратов*** состоит в минимизации по всем возможным значениям $\theta_1,\text{K},\theta_p$ суммы квадратов

$$Q\left(\theta_1,\text{K}\ \theta_p\right)=\sum_{i=1}^{n}\left(y_i-\theta_1 x_{i1}-\text{K}-\theta_p x_{ip}\right)^2.$$

Минимум этой суммы достигается при некотором наборе значений коэффициентов

$$\theta_1=\hat{\theta}_1,\text{K}\ ,\theta_p=\hat{\theta}_p,$$

так что

$$Q\left(\hat{\theta}_1,\text{K}\ \hat{\theta}_p\right)=\min_{\theta_1,\text{K},\theta_p} Q\left(\theta_1,\text{K}\ \theta_p\right).$$

Это минимальное значение мы опять обозначаем RSS, так что

$$RSS=\sum_{i=1}^{n}\left(y_i-\hat{\theta}_1 x_{i1}-\text{K}-\hat{\theta}_p x_{ip}\right)^2,$$

и называем ***остаточной суммой квадратов***.

Коэффициент детерминации R^2 определяется как

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

где

$$TSS=\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2.$$

Обозначая

$$\hat{y}_i=\hat{\theta}_1 x_{i1}+\text{K}+\hat{\theta}_p x_{ip}\ ,\ \ i=1,\text{K}\ ,n,$$

(***подобранные - fitted***- значения объясняющей переменной по оцененной линейной модели связи), и определяя ***остаток*** (***residual***) ***от i-го наблюдения*** как

$$e_i=y_i-\hat{y}_i\ ,$$

мы получаем:

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

Обозначая

$$ESS = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y} \right)^2$$

- ***объясненная моделью*** (***explained***) ***сумма квадратов,*** или ***регрессионная сумма квадратов***, мы так же, как и в случае ***простой*** линейной регрессии с $p = 2$, имеем разложение

$$TSS = RSS + ESS,$$

так что

$$R^2 = \frac{ESS}{TSS}.$$

И опять, это разложение справедливо только при наличии постоянной составляющей в модели линейной связи. При этом, также, здесь

$$R^2 = r_{y,\hat{y}}^2,$$

т.е. коэффициент детерминации равен квадрату выборочного коэффициента корреляции $r_{y,\hat{y}}$ между переменными y и $\hat{y}$. Последний называется ***множественным коэффициентом корреляции*** (***multiple-R***).

Для поиска значений $\hat{\theta}_1, \text{K}, \hat{\theta}_p$**,** минимизирующих сумму

$$Q\left(\theta_1, \text{K}\ \theta_p\right) = \sum_{i=1}^{n} \left(y_i - \theta_1 x_{i1} - \text{K} - \theta_p x_{ip} \right)^2 ,$$

следует приравнять нулю частные производные этой суммы (как функции от $\theta_1, \text{K}, \theta_p$**)** по каждому из аргументов $\theta_1, \text{K}, \theta_p$. В результате получаем ***систему нормальных уравнений***

$$\sum_{i=1}^{n} 2\left(y_i - \bar{\theta}_1 x_{i1} - \text{K} - \bar{\theta}_p x_{ip}\right)\left(-x_{i1}\right) = 0,$$

$$\sum_{i=1}^{n} 2\left(y_i - \bar{\theta}_1 x_{i1} - \text{K} - \bar{\theta}_p x_{ip}\right)\left(-x_{i2}\right) = 0,$$

K

$$\sum_{i=1}^{n} 2\left(y_i - \bar{\theta}_1 x_{i1} - \text{K} - \bar{\theta}_p x_{ip}\right)\left(-x_{ip}\right) = 0,$$

или

$$\left(\sum_{i=1}^{n} x_{i1}^2\right)\cdot\bar{\theta}_1 + \left(\sum_{i=1}^{n} x_{i1}x_{i2}\right)\cdot\bar{\theta}_2 + \text{K} + \left(\sum_{i=1}^{n} x_{i1}x_{ip}\right)\cdot\bar{\theta}_p = \sum_{i=1}^{n} y_i x_{i1},$$

$$\left(\sum_{i=1}^{n} x_{i2}x_{i1}\right)\cdot\bar{\theta}_1 + \left(\sum_{i=1}^{n} x_{i2}^2\right)\cdot\bar{\theta}_2 + \text{K} + \left(\sum_{i=1}^{n} x_{i2}x_{ip}\right)\cdot\bar{\theta}_p = \sum_{i=1}^{n} y_i x_{i2},$$

K

$$\left(\sum_{i=1}^{n} x_{ip}x_{i1}\right)\cdot\bar{\theta}_1 + \left(\sum_{i=1}^{n} x_{ip}x_{i2}\right)\cdot\bar{\theta}_2 + \text{K} + \left(\sum_{i=1}^{n} x_{ip}^2\right)\cdot\bar{\theta}_p = \sum_{i=1}^{n} y_i x_{ip}.$$

Это система p линейных уравнений с p неизвестными $\bar{\theta}_1, \text{K}, \bar{\theta}_p$. Ее можно решать или методом подстановки или по правилу Крамера с использованием соответствующих определителей. В векторно-матричной форме эта система имеет вид

$$\boxed{X^T X\bar{\theta} = X^T y}$$

где

$$X = \begin{pmatrix} x_{11} & x_{12} & \text{K} & x_{1p} \\ x_{21} & x_{22} & \text{K} & x_{2p} \\ \text{M} & \text{M} & \text{K} & \text{M} \\ x_{n1} & x_{n2} & \text{K} & x_{np} \end{pmatrix}$$

- матрица значений p объясняющих переменных в n наблюдениях;

$$X^T = \begin{pmatrix} x_{11} & x_{12} & \text{K} & x_{n1} \\ x_{21} & x_{22} & \text{K} & x_{n2} \\ \text{M} & \text{M} & \text{K} & \text{M} \\ x_{1p} & x_{2p} & \text{K} & x_{np} \end{pmatrix}$$

- транспонированная матрица;

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \text{M} \\ y_n \end{pmatrix} \qquad \text{и} \qquad \bar{\theta} = \begin{pmatrix} \bar{\theta}_1 \\ \bar{\theta}_2 \\ \text{M} \\ \bar{\theta}_p \end{pmatrix}$$

соответственно, вектор-столбец значений объясняемой переменной в n наблюдениях и вектор-столбец оценок p неизвестных коэффициентов. Система нормальных уравнений имеет единственное решение, если выполнено условие

(4) ***матрица X^TX невырождена,*** т.е. ее ***определитель отличен от нуля***:

$$\det X^T X \neq 0 ,$$

которое можно заменить условием

(4') ***столбцы матрицы X линейно независимы***.

При выполнении этого условия матрица $X^T X$ (размера $p \times p$) имеет обратную к ней матрицу $(X^T X)^{-1}$. Умножая в таком случае обе части последнего уравнения слева на матрицу $(X^T X)^{-1}$, находим искомое решение системы нормальных уравнений:

$$\bar{\theta} = \left(X^T X\right)^{-1} X^T y .$$

Введем дополнительные обозначения

$$\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \text{M} \\ \theta_p \end{pmatrix}, \ \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \text{M} \\ \varepsilon_p \end{pmatrix}, \ \hat{y} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \text{M} \\ \hat{y}_n \end{pmatrix}, \quad e = \begin{pmatrix} e_1 \\ e_2 \\ \text{M} \\ e_n \end{pmatrix}.$$

Тогда модель наблюдений

$$\boxed{y_i = \theta_1 x_{i1} + \text{K} + \theta_p x_{ip} + \varepsilon_i, \ \ i = 1, \text{K}, n,}$$

можно представить в матрично-векторной форме

$$\boxed{y = X\theta + \varepsilon .}$$

Вектор подобранных значений имеет вид

$$\boxed{\hat{y} = X\hat{\theta}}$$

и вектор остатков равен

$$\boxed{e = y - \hat{y} = y - X\hat{\theta} .}$$

Определяющим для всего последующего является то обстоятельство, что в нормальной линейной модели с несколькими объясняющими переменными оценки $\hat{\theta}_1, \text{K}, \hat{\theta}_p$ коэффициентов $\theta_1, \text{K}, \theta_p$ как случайные величины имеют ***нормальные распределения*** (хотя эти случайные величины уже ***не являются независимыми в совокупности***).

Действительно, поскольку $\hat{\theta} = \left(X^T X\right)^{-1} X^T y$, то оценки $\hat{\theta}_1, \text{K}, \hat{\theta}_p$ являются линейными комбинациями значений y_1, K, y_n, т.е. имеют вид

$$\hat{\theta}_j = c_{j1} y_1 + c_{j2} y_2 + \text{K} + c_{jn} y_n ,$$

где c_{jk} - коэффициенты, определяемые значениями объясняющих переменных. Поскольку же у нас y_1, K, y_n - на-

блюдаемые значения случайных величин Y_1,K,Y_n, то $\bar{\theta}_j$ является наблюдаемым значением случайной величины $c_{j1}Y_1 + c_{j2}Y_2 + \text{K} + c_{jn}Y_n$, которую мы также будем обозначать $\bar{\theta}_j$:

$$\boxed{\bar{\theta}_j = c_{j1}Y_1 + c_{j2}Y_2 + \text{K} + c_{jn}Y_n,\ j = 1,\text{K},p.}$$

Ранее мы выяснили, что при наших предположениях

$$Y_i \sim N(\theta_1 x_{i1} + \text{K} + \theta_p x_{ip}, \sigma^2),\ i = 1,\text{K},n.$$

Поэтому случайные величины $\bar{\theta}_1,\text{K},\bar{\theta}_p$ также будут нормальными как линейные комбинации независимых нормально распределенных случайных величин.

Можно показать, что математическое ожидание случайной величины $\bar{\theta}_j$ равно

$$\boxed{E\left(\bar{\theta}_j\right) = \theta_j,\ j = 1,\text{K},p,}$$

($\bar{\theta}_j$ является ***несмещенной оценкой*** истинного значения коэффициента θ_j), а дисперсия этой случайной величины равна j-му диагональному элементу матрицы $\sigma^2 (X^T X)^{-1}$:

$$\boxed{D\left(\bar{\theta}_j\right) = \left[\sigma^2 (X^T X)^{-1}\right]_{jj}.}$$

Рассмотренная ранее модель простой линейной регрессии

$$y_i = \alpha + \beta\, x_i + \varepsilon_i,\ i = 1,\text{K},n,$$

вкладывается в модель множественной линейной регрессии с $p = 2$:

$$y=\begin{pmatrix} y_1 \\ y_2 \\ \text{M} \\ y_n \end{pmatrix}, \quad X=\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \text{M} & \text{M} \\ 1 & x_n \end{pmatrix}, \quad \theta=\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \quad \varepsilon=\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \text{M} \\ \varepsilon_n \end{pmatrix}.$$

Матрица $(X^T X)^{-1}$ имеет вид

$$\left(X^T X\right)^{-1}=\frac{1}{n\sum_{i=1}^{n} x_i^2-\left(\sum_{i=1}^{n} x_i\right)^2}\begin{pmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{pmatrix}.$$

Учитывая, что

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

находим:

$$D\left(\hat{\alpha}\right)=\left[\sigma^2\left(X^T X\right)^{-1}\right]_{11}=\frac{\sigma^2\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2},$$

$$D\left(\hat{\beta}\right)=\left[\sigma^2\left(X^T X\right)^{-1}\right]_{22}=\frac{\sigma^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}.$$

2.5. НОРМАЛЬНАЯ МНОЖЕСТВЕННАЯ РЕГРЕССИЯ: ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ ДЛЯ КОЭФФИЦИЕНТОВ

Рассматривая нормальную модель линейной множественной регрессии

$$y_i = \theta_1 x_{i1} + \mathrm{K} + \theta_p x_{ip} + \varepsilon_i , \quad i = 1, \mathrm{K}, n,$$

с $\varepsilon_i \sim i.\ i.\ d.\ N(0, \sigma^2)$, мы установили, что оценка наименьших квадратов $\bar{\theta}_j$ неизвестного истинного значения θ_j коэффициента при j — ой объясняющей переменной имеет нормальное распределение, причем

$$E(\bar{\theta}_j) = \theta_j , \quad D(\bar{\theta}_j) = \left[\sigma^2 (X^T X)^{-1}\right]_{jj} , \ j = 1, \mathrm{K}, n .$$

Рассмотрим теперь случайную величину

$$\frac{\bar{\theta}_j - \theta_j}{\sqrt{D(\bar{\theta}_j)}} ,$$

получаемую путем вычитания из случайной величины $\bar{\theta}_j$ ее математического ожидания и деления полученной разности на корень из дисперсии $\bar{\theta}_j$ (т. е. путем ***центрирования*** и ***нормирования*** случайной величины $\bar{\theta}_j$). При совершении этих двух действий мы не выходим из семейства нормальных случайных величин, получая опять же нормальную случайную величину, но только уже с другими математическим ожиданием и дисперсией. Используя упомянутые ранее свойства математического ожидания и дисперсии, находим:

$$E\left(\frac{\bar{\theta}_j - \theta_j}{\sqrt{D\left(\bar{\theta}_j\right)}}\right) = \frac{1}{\sqrt{D\left(\bar{\theta}_j\right)}}\left(E(\bar{\theta}_j) - \theta_j\right) = 0\,,$$

$$D\left(\frac{\bar{\theta}_j - \theta_j}{\sqrt{D\left(\bar{\theta}_j\right)}}\right) = \frac{1}{D\left(\bar{\theta}_j\right)} D\left(\bar{\theta}_j - \theta_j\right) = 1\,,$$

так что

$$\frac{\bar{\theta}_j - \theta_j}{\sqrt{D\left(\bar{\theta}_j\right)}} \sim N(0,1)\,,\ j = 1,\text{K}\,,p\,.$$

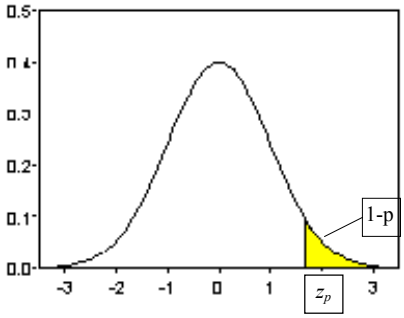
Иными словами, в результате центрирования и нормирования случайной величины $\bar{\theta}_j$ мы получили случайную величину, имеющую ***стандартное нормальное распределение***, т. е. ***нормальное распределение с нулевым математическим ожиданием и единичной дисперсией***. Функцию распределения и функцию плотности распределения такой случайной величины обозначают, соответственно, как $\Phi(x)$ и $\varphi(x)$:

$$\boxed{\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2},\ \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt.}$$

Для каждого значения p, $0 < p < 1$, определим символом z_p число, для которого $\Phi(z_p) = p$, так что если случайная величина Z имеет стандартное нормальное распределение, то тогда

$$\boxed{P\left\{Z \le z_p\right\} = p\,.}$$

Такое число называется ***квантилью уровня p*** стандартного нормального распределения.



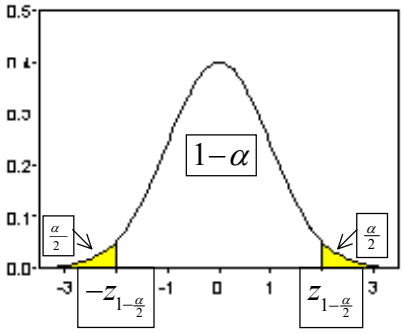
Заштрихованная площадь под графиком плотности стандартного нормального распределения находится правее квантили z_p уровня 0.95;

эта квантиль равна $z_{0.95} = 1.645$. Поэтому площадь под кривой, лежащая левее точки $z = 1.645$, равна 0.95, а заштрихованная площадь равна $1 - 0.95 = 0.05$. Последняя величина есть вероятность того,что случайная величина Z, имеющая стандартное нормальное распределение, примет значение, превышающее 1.645.

Если мы возьмем какое-нибудь число α в пределах от 0.5 до 1, $0.5 < \alpha < 1$, и выделим интервал

$$\left(-z_{1-\frac{\alpha}{2}}\ ,\ z_{1-\frac{\alpha}{2}}\right),$$

то получим следующую картину:



Из симметрии функции плотности нормального распределения вытекает равенство площадей областей, заштрихованных на последнем рисунке. Но площадь правой заштрихованной области равна $1-\left(1-\frac{\alpha}{2}\right)=\frac{\alpha}{2}$; следовательно, такова же и площадь левой заштрихованной области. Это, в частности, означает, что вероятность того, что случайная величина Z примет значение, не превышающее $-z_{1-\frac{\alpha}{2}}$, равна $\frac{\alpha}{2}$, так что

$$\boxed{-z_{1-\frac{\alpha}{2}}=z_{\frac{\alpha}{2}}}.$$

Часть площади под кривой стандартной нормальной плотности, лежащая в пределах выделенного интервала, меньше единицы на сумму площадей заштрихованных областей («хвостов»), т. е. равна

$$1-\left(\frac{\alpha}{2}+\frac{\alpha}{2}\right)=1-\alpha.$$

Эта величина равна вероятности того, что случайная величина Z, имеющая стандартное нормальное распределение, примет значение в пределах указанного интервала[2]:

[2] Заметим, что в этом и других подобных выражениях знак $\leq$ можно свободно заменять знаком $<$, а знак $\geq$ знаком $>$ (и обратно), поскольку мы всегда предполагаем существование функции плотности распределений рассматриваемых случайных величин.

$$P\left\{-z_{1-\frac{\alpha}{2}} \le Z \le z_{1-\frac{\alpha}{2}}\right\} = 1-\alpha \ .$$

Но ранее мы установили, что стандартное нормальное распределение имеет случайная величина

$$\frac{\bar{\theta}_j - \theta_j}{\sqrt{D\left(\bar{\theta}_j\right)}} .$$

Поэтому для этой случайной величины справедливо соотношение

$$P\left\{-z_{1-\frac{\alpha}{2}} \le \frac{\bar{\theta}_j - \theta_j}{\sqrt{D\left(\bar{\theta}_j\right)}} \le z_{1-\frac{\alpha}{2}}\right\} = 1-\alpha \ ,$$

так что с вероятностью, равной $1-\alpha$, выполняется двойное неравенство

$$-z_{1-\frac{\alpha}{2}} \le \frac{\bar{\theta}_j - \theta_j}{\sqrt{D\left(\bar{\theta}_j\right)}} \le z_{1-\frac{\alpha}{2}} ,$$

т. е.

$$\bar{\theta}_j - z_{1-\frac{\alpha}{2}}\sqrt{D\left(\bar{\theta}_j\right)} \le \theta_j \le \bar{\theta}_j + z_{1-\frac{\alpha}{2}}\sqrt{D\left(\bar{\theta}_j\right)} .$$

Иными словами, ***с вероятностью, равной 1–α, случайный интервал***

$$\left[\bar{\theta}_j - z_{1-\frac{\alpha}{2}}\sqrt{D\left(\bar{\theta}_j\right)} \ , \ \bar{\theta}_j + z_{1-\frac{\alpha}{2}}\sqrt{D\left(\bar{\theta}_j\right)}\right]$$

накрывает истинное значение коэффициента θ_j. Такой интервал называется ***доверительным интервалом для*** θ_j ***с уровнем доверия (доверительной вероятностью) 1–α,*** или ***(1–α)-доверительным интервалом,*** или ***100(1–α)-процентным доверительным интервалом для*** θ_j.

Последний рисунок был получен при значении $\alpha = 0.05$. Поэтому площади заштрихованных областей («***хвосты***») равны $\frac{\alpha}{2} = 0.025$, сумма этих площадей равна 0.05 , и площадь области под кривой в пределах интервала $\left(-z_{1-\frac{\alpha}{2}}\ ,\ z_{1-\frac{\alpha}{2}}\right)$ равна $1-0.05 = 0.95$. Остается заметить, что

$$z_{0.95} = 1.960\ ,$$

так что случайный интервал

$$\left[\bar{\theta}_j - 1.96\sqrt{D\left(\bar{\theta}_j\right)}\ ,\ \bar{\theta}_j + 1.96\sqrt{D\left(\bar{\theta}_j\right)}\right]$$

является ***95%-доверительным интервалом для*** θ_j. Его ***длина***

$$2 \cdot 1.96\sqrt{D\left(\bar{\theta}_j\right)}$$

пропорциональна $\sqrt{D\left(\bar{\theta}_j\right)}$ — ***среднеквадратической ошибке (среднеквадратическому отклонению)*** оценки коэффициента θ_j.

Хотелось бы, конечно, прямо сейчас построить доверительные интервалы для коэффициентов линейной модели по каким-нибудь реальным статистическим данным. Однако этому препятствует то обстоятельство, что в выражения для дисперсий

$$D\left(\bar{\theta}_j\right) = \left[\sigma^2 (X^T X)^{-1}\right]_{jj}\ ,\ \ i = 1,\text{K}\ ,n\ ,$$

входит ***не известное нам*** значение σ^2.

2.6. ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ ДЛЯ КОЭФФИЦИЕНТОВ: РЕАЛЬНЫЕ СТАТИСТИЧЕСКИЕ ДАННЫЕ

Итак, практическому построению доверительных интервалов для коэффициентов θ_j нормальной модели линейной множественной регрессии

$$y_i = \theta_1 x_{i1} + \mathrm{K} + \theta_p x_{ip} + \varepsilon_i, \quad i = 1, \mathrm{K}, n,$$

с $\varepsilon_i \sim i.\,i.\,d.\ N(0, \sigma^2)$ препятствует вхождение в выражения для дисперсий

$$D(\bar{\theta}_j) = \left[\sigma^2 (X^T X)^{-1}\right]_{jj}, \quad i = 1, \mathrm{K}, n,$$

неизвестного значения σ^2.

Единственный выход из этого положения — ***заменить неизвестное значение*** σ^2 какой-нибудь подходящей его ***оценкой*** (***estimate***), которую можно было бы ***вычислить на основании имеющихся статистических данных***. Такого рода оценки принято называть ***статистиками*** (***statistics***).

В данной ситуации такой подходящей оценкой для неизвестного значения σ^2 является статистика

$$S^2 = \frac{RSS}{n - p}.$$

Поскольку сумма $RSS = \sum_{i=1}^{n} \left(y_i - \bar{y}_i\right)^2$ является ***квадратичной функцией от случайных величин*** $\varepsilon_1, \mathrm{K}, \varepsilon_n$, то она является случайной величиной, а следовательно, ***случайной величиной*** является и ***статистика*** $\boldsymbol{S^2}$. Математическое ожидание этой случайной величины равно σ^2:

$$E(S^2) = \sigma^2,$$

т. е. S^2 — ***несмещенная оценка*** для σ^2.

Замечание. В частном случае $p = 1$ модель наблюдений принимает вид

$$y_i = \theta_1 + \varepsilon_i, \quad i = 1,\text{K},n,$$

(случайная выборка из распределения N (θ_1,σ^2)***)***. Несмещенной оценкой для σ^2 служит

$$S^2 = \frac{RSS}{n-1}.$$

Оценкой наименьших квадратов для параметра θ_1 является $\hat{\theta}_1 = \bar{y}$, так что $RSS = \sum_{i=1}^{n}(y_i - \bar{y})^2 = TSS$, и

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

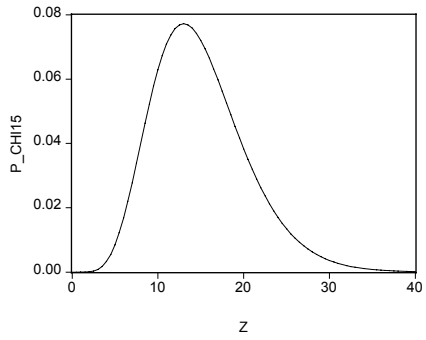
Таким образом, выборочная дисперсия $Var(y)$ переменной y, получаемая делением TSS именно на $n-1$ (а не на n), является несмещенной оценкой для σ^2 в модели случайной выборки из нормального распределения, имеющего дисперсию σ^2. Этим и объясняется сделанный нами выбор нормировки при определении выборочных дисперсий и ковариаций.

При выполнении стандартных предположений отношение

$$\frac{(n-p)S^2}{\sigma^2} = \frac{RSS}{\sigma^2}$$

имеет стандартное распределение, называемое ***распределением хи-квадрат с (n-p) степенями свободы***. Такое же распределение имеет сумма квадратов $n-p$ случайных величин, ***независимых в совокупности и имеющих одинаковое***

стандартное нормальное распределение. При $n - p = 15$ график функции плотности этого распределения имеет вид



Для обозначения распределения хи-квадрат с K степенями свободы используют символ $\chi^2(K)$.

Итак, мы не знаем истинного значения σ^2 и поэтому в попытке построить доверительный интервал для θ_j вынуждены заменить неизвестное нам значение $D\left(\bar{\theta}_j\right) = \left[\sigma^2 (X^T X)^{-1}\right]_{jj}$ на его несмещенную оценку

$$\boxed{s^2_{\bar{\theta}_j} = S^2 (X^T X)^{-1}_{jj} \ .}$$

Соответственно, вместо отношения

$$\frac{\bar{\theta}_j - \theta_j}{\sqrt{D\left(\bar{\theta}_j\right)}}$$

приходится использовать отношение

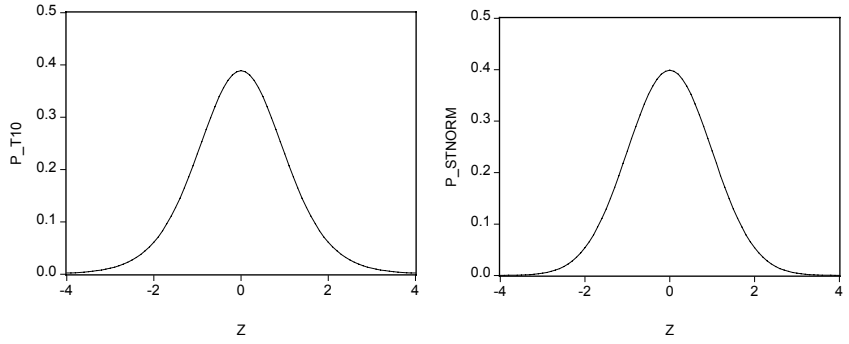
$$\boxed{\frac{\bar{\theta}_j - \theta_j}{s_{\bar{\theta}_j}} \ .}$$

Однако последнее отношение как случайная величина уже ***не имеет стандартного нормального распределения***, по-

скольку в знаменателе теперь стоит не постоянная, а случайная величина.

Тем не менее, распределение последнего отношения также относят к стандартным, и оно известно под названием ***t-распределения Стьюдента с (n-p) степенями свободы***.

Для распределения Стьюдента с K степенями свободы принято обозначение ***t (K)***. Квантиль уровня p такого распределения будем обозначать символом $\boldsymbol{t_p\ (K)}$. График функции плотности распределения Стьюдента симметричен относительно нуля и похож на график функции плотности нормального распределения. Например, при K=10 он имеет следующий вид (левый график).



Для сравнения, справа приведен график функции стандартного нормального распределения. Отличие графиков столь невелико, что визуально они почти неразличимы. Квантили этих двух распределений различаются более ощутимо:

$$z_{0.95} = 1.645\ ,\ \ t_{0.95}(10)\ = 1.812\ ;$$

$$z_{0.975} = 1.960\ ,\ t_{0.975}(10) = 2.228;$$

$$z_{0.99}\ = 2.326,\ \ t_{0.99}(10)\ = 2.764;$$

$$z_{0.995} = 2.576,\ \ t_{0.995}(10) = 3.169.$$

Распределение Стьюдента имеет более тяжелые хвосты. Из приведенных значений квантилей следует, например, что случайная величина, имеющая стандартное нормальное распределение, может превысить значение 1.645 лишь с вероятностью 0.05. В то же самое время, с такой же вероятностью 0.05 случайная величина, имеющая распределение Стьюдента с 10 степенями свободы, принимает значения, большие, чем 1.812.

Впрочем, для значений $K > 30$ квантили распределения Стьюдента $t(K)$ практически совпадают с соответствующими квантилями стандартного нормального распределения $N(0,1)$.

Итак,

$$\frac{\bar{\theta}_j - \theta_j}{s_{\bar{\theta}_j}} \sim t(n-p).$$

Поэтому для этой случайной величины выполняется соотношение

$$P\left\{-t_{1-\frac{\alpha}{2}}(n-p) \le \frac{\bar{\theta}_j - \theta_j}{s_{\bar{\theta}_j}} \le t_{1-\frac{\alpha}{2}}(n-p)\right\} = 1-\alpha ,$$

так что с вероятностью, равной $1-\alpha$, выполняется двойное неравенство

$$-t_{1-\frac{\alpha}{2}}(n-p) \le \frac{\bar{\theta}_j - \theta_j}{s_{\bar{\theta}_j}} \le t_{1-\frac{\alpha}{2}}(n-p),$$

т. е.

$$\boxed{\bar{\theta}_j - t_{1-\frac{\alpha}{2}}(n-p)\, s_{\bar{\theta}_j} \le \theta_j \le \bar{\theta}_j + t_{1-\frac{\alpha}{2}}(n-p)\, s_{\bar{\theta}_j}}.$$

Иными словами, ***с вероятностью, равной 1–α, случайный интервал***

$$\left[\bar{\theta}_j - t_{1-\frac{\alpha}{2}}(n-p)\, s_{\bar{\theta}_j} \;,\; \bar{\theta}_j + t_{1-\frac{\alpha}{2}}(n-p)\, s_{\bar{\theta}_j}\right]$$

накрывает истинное значение коэффициента θ_j, т. е. является ***95%- доверительным интервалом для*** θ_j в случае, ***когда не известно истинное значение*** σ^2 дисперсии случайных ошибок $\varepsilon_1, \text{K}, \varepsilon_n$. В среднем, длина такого интервала больше, чем длина доверительного интервала с тем же уровнем доверия, построенного при известном значении σ^2.

Замечание. Выбор конкретного значения α определяет компромисс между желанием получить более короткий доверительный интервал и желанием обеспечить более высокий уровень доверия.

Попытка повысить уровень доверия $1-\alpha$, выраженная в выборе меньшего значения α, приводит к квантили $t_{1-\frac{\alpha}{2}}(n-p)$ с более высоким значением $1-\frac{\alpha}{2}$, т. е. к большему значению $t_{1-\frac{\alpha}{2}}(n-p)$. Но длина доверительного интервала пропорциональна $t_{1-\frac{\alpha}{2}}(n-p)$. Следовательно, ***увеличение уровня доверия сопровождается увеличением ширины доверительного интервала*** (при тех же статистических данных).

Так, для $n-p>30$ можно приближенно считать, что

$t_{1-\frac{\alpha}{2}} \cong z_{1-\frac{\alpha}{2}}$,

где z_p — квантиль уровня p стандартного нормального распределения. Соответственно, выбирая уровень доверия $1-\alpha$ равным 0.9, 0.95 или 0.99, мы получаем для $t_{1-\frac{\alpha}{2}}(n-p)$ значения, приблизительно равные $z_{0.95}=1.64$, $z_{0.975}=1.96$, $z_{0.995}=2.58$. Это означает, что переход от уровня доверия 0.9 к уровню доверия 0.95 сопровождается увеличением длины доверительного интервала приблизитель-

но в 1.2 раза, а дополнительное повышение уровня доверия до 0.99 увеличивает длину доверительного интервала еще примерно в 1.3 раза.

Теперь мы в состоянии перейти к построению интервальных оценок параметров моделей линейной регрессии для различного рода социально-экономических факторов на основании соответствующих статистических данных.

Пример. Вернемся к модели зависимости уровня безработицы среди белого населения США от уровня безработицы среди цветного населения. Запишем линейную модель наблюдений в виде

$$\boxed{BEL_i = \theta_1 + \theta_2\, ZVET_i + \varepsilon_i\,,\ \ i = 1,\text{K}\,,n\,.}$$

Получаем: $S^2 = RSS/(n-2) = 0.161231/(17-2) = 0.010749$. Коэффициент θ_2 оценивается величиной $\bar{\theta}_2 = 0.125265$; дисперсия $D(\bar{\theta}_2)$ оценивается величиной $s^2_{\bar{\theta}_2} = (0.062286)^2$. Для построения 95% — доверительного интервала для θ_2 остается найти квантиль уровня $1 - \frac{0.05}{2} = 0.975$ распределения Стьюдента с $n - p = 17 - 2 = 15$ степенями свободы. Используя, например, Таблицу А.2 из книги Доугерти (стр.368), находим: $t_{0.975}(15) = 2.131$. Соответственно, получаем 95%-доверительный интервал для θ_2 в виде

$$\bar{\theta}_2 - t_{0.975}(15)\, s_{\bar{\theta}_2} \le \theta_2 \le \bar{\theta}_2 + t_{0.975}(15)\, s_{\bar{\theta}_2}\,,$$

т. е.

$$\boxed{-0.0075 \le \theta_2 \le 0.2580\,.}$$

Для θ_1 имеем $\bar{\theta}_1 = 2.293843$, $s_{\bar{\theta}_1} = 0.410396$; 95%-доверительный интервал для θ_1 имеет вид

$$\bar{\theta}_1 - t_{0.975}(15)\, s_{\bar{\theta}_1} \le \theta_1 \le \bar{\theta}_1 + t_{0.975}(15)\, s_{\bar{\theta}_1}\,,$$

т. е.

$$\boxed{1.4193 \le \theta_1 \le 3.1684}\ .$$

В связи с этим примером, отметим два обстоятельства.

(а) Доверительный интервал для коэффициента θ_2 допускает как положительные, так и отрицательные значения этого коэффициента.

(б) Каждый из двух построенных интервалов имеет уровень доверия 0.95; однако это не означает, что с той же вероятностью 0.95 сразу оба интервала накрывают истинные значения параметров θ_1, θ_2.

Справиться с первым затруднением в данном примере можно, понизив уровень доверия до 0.90. В этом случае в выражении для доверительного интервала квантиль $t_{0.975}(15) = 2.131$ заменяется на квантиль $t_{0.95}(15) = 1.753$, так что левая граница доверительного интервала для θ_2 становится положительной и равной 0.0164. Однако это достигается ценой того, что новый доверительный интервал будет накрывать истинное значение параметра θ_2 в среднем только в 90 случаев из 100, а не в 95 из100 случаев.

Что касается второго затруднения, то наиболее простой путь взятия под контроль вероятности одновременного накрытия доверительными интервалами для θ_1, θ_2 истинных значений этих параметров связан с тем, что

$$P\{\ \text{оба интервала накрывают } \theta_1 \text{ и } \theta_2\text{, соответственно}\ \} =$$

$$1 - P\{\ \text{хотя бы один из них не накрывает соответствующее } \theta_j\ \} =$$

$$1 - \big[\ P\{\ \text{доверительный интервал для } \theta_1 \text{ не накрывает } \theta_1\ \} +$$

$$P\{\ \text{доверительный интервал для } \theta_2 \text{ не накрывает } \theta_2\ \} -$$

$$P\{\ \text{оба интервала не накрывают свои } \theta_j\ \}\ \big] =$$

$1-\left[\alpha+\alpha-P\left\{\ \underline{\text{оба}}\ \text{интервала не накрывают свои}\ \theta_j\ \right\}\ \right]\geq$
$1-\alpha\ -\alpha=1-2\alpha$.

Следовательно, если построить доверительный интервал для θ_1 и доверительный интервал для θ_2 с уровнями доверия каждого, равными $\alpha^*=\alpha/2$, то тогда правая часть полученной цепочки соотношений будет равна $1-2\alpha^*=1-\alpha$.

Это означает, что в нашем примере мы можем гарантировать, что вероятность <u>одновременного накрытия</u> истинных значений θ_1,θ_2 соответствующими доверительными интервалами будет <u>не менее</u> 0.95, если возьмем $\alpha^*=0.025$. Но тогда при построении этих интервалов придется использовать вместо значения

$t_{1-\frac{\alpha}{2}}(15)=t_{0.975}(15)=2.131$

значение

$t_{1-\frac{\alpha^*}{2}}(15)=t_{1-\frac{0.025}{2}}(15)=t_{0.9875}(15)=2.49$,

так что каждый из исходных интервалов <u>увеличится</u> в $2.9/2.131\cong 1.17$ раза. Это, конечно, приводит к еще более неопределенным выводам относительно истинных значений параметров θ_1,θ_2.

2.7. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ О ЗНАЧЕНИЯХ КОЭФФИЦИЕНТОВ

В только что рассмотренном примере мы построили 95% — доверительный интервал для параметра θ_2 в виде

$\bar{\theta}_2\ -t_{0.975}(15)\,s_{\bar{\theta}_2}\leq\theta_2\leq\bar{\theta}_2+t_{0.975}(15)\,s_{\bar{\theta}_2}$,

т. е.

$\boxed{-0.0075\leq\theta_2\leq 0.2580\ .}$

Существенно, что при любом истинном значении параметра θ_2 вероятность накрытия этого значения построенным доверительным интервалом равна 0.95.

Рассмотрим значение $\theta_2 = 1$; построенный интервал его не накрывает. Однако если θ_2 действительно равняется 1, то вероятность такого ненакрытия равна $1 - 0.95 = 0.05$. Таким образом, факт ненакрытия значения $\theta_2 = 1$ построенным интервалом представляет (в случае, когда $\theta_2 = 1$) осуществление довольно редкого события, имеющего малую вероятность 0.05, и это дает нам основания сомневаться в том, что в действительности $\theta_2 = 1$.

То же самое относится и к любому другому фиксированному значению θ_2^0, не принадлежащему указанному 95%-доверительному интервалу: предположение о том, что в действительности $\theta_2 = \theta_2^0$, представляется маловероятным.

Подобного рода предположения называют в этом контексте ***статистическими гипотезами (statistical hypothesis)***. О ***проверяемой гипотезе*** говорят как об ***исходной — «нулевой» (maintained, null)*** гипотезе

и обозначают такую гипотезу символом H_0, так что в последнем случае мы имеем дело с гипотезой

$$\boxed{H_0 : \theta_2 = \theta_2^0}$$

В соответствии со сказанным выше, такую гипотезу естественно ***отвергать (отклонять)***, если значение θ_2^0 ***не принадлежит*** 95%-доверительному интервалу для θ_2, т. е. интервалу

$$\left[-0.0075,\ 0.2580\right].$$

Вспоминая, как этот интервал строился, мы замечаем, что θ_2^0 не принадлежит этому интервалу тогда и только тогда, когда

$$\left|\frac{\bar{\theta}_2 - \theta_2^0}{s_{\bar{\theta}_2}}\right| > t_{0.975}(15),$$

т. е. когда наблюдаемое значение отношения

$$\frac{\bar{\theta}_2 - \theta_2^0}{s_{\bar{\theta}_2}}$$

«слишком велико» по абсолютной величине. Последнее означает «слишком большое» отклонение оценки $\bar{\theta}_2$ от гипотетического значения θ_2^0 параметра θ_2, в сравнении с оценкой $s_{\bar{\theta}_2}$ значения $\sqrt{D\left(\bar{\theta}_2\right)}$ корня из дисперсии оценки этого параметра.

Итак, если

$$\left|\frac{\bar{\theta}_2 - \theta_2^0}{s_{\bar{\theta}_2}}\right| > t_{0.975}(15),$$

мы отвергаем гипотезу $H_0: \theta_2 = \theta_2^0$. Однако выполнение этого неравенства для некоторого значения θ_2^0 ***вовсе не означает***, что гипотеза $H_0: \theta_2 = \theta_2^0$ обязательно не верна. Если в действительности $\theta_2 = \theta_2^0$, то все же имеется вероятность $1 - 0.95 = 0.05$ того, что это неравенство будет выполнено.

В последнем случае, в соответствии с выбранным правилом, мы все же отвергнем гипотезу H_0, допустив при этом «***ошибку 1-го рода***». Такая ошибка происходит в среднем в 5 случаях из ста.

Если бы мы выбрали произвольный доверительный уровень $1-\alpha$, то тогда мы отвергали бы гипотезу $H_0: \theta_2 = \theta_2^0$ при выполнении неравенства

$$\left| \frac{\bar{\theta}_2 - \theta_2^0}{s_{\bar{\theta}_2}} \right| > t_{1-\frac{\alpha}{2}}(15),$$

и ошибка 1-го рода происходила в среднем в 100α случаев из 100. Точнее, вероятность ошибки 1-го рода была бы равна α:

$$P\{ \ H_0 \text{ отвергается} \mid \ H_0 \text{ верна} \ \} = \alpha .$$

Само ***правило*** решения вопроса об отклонении или неотклонении статистической гипотезы H_0 называется ***статистическим критерием проверки гипотезы H_0***, а выбранное при формулировании этого правила значение α называется ***уровнем значимости*** критерия.

Выбор большего или меньшего значения α определяется степенью значимости для исследователя исходной гипотезы H_0. Скажем, выбор между значениями $\alpha = 0.05$ и $\alpha = 0.01$ в пользу $\alpha = 0.01$ означает, что исследователь заранее настроен в пользу гипотезы H_0 и ему требуются очень весомые аргументы, свидетельствующие против этой гипотезы, чтобы все же отказаться от нее. Выбор же в пользу уровня значимости $\alpha = 0.05$ означает, что исследователь не столь сильно отстаивает гипотезу H_0 и готов отказаться от нее и при менее убедительной аргументации против этой гипотезы.

Всякий статистический критерий основывается на использовании той или иной ***статистики (статистики критерия)***, т. е. случайной величины, значения которой ***могут быть вычислены*** (по крайней мере, теоретически) на основании

имеющихся статистических данных и распределение которой ***известно*** (хотя бы приближенно).

В нашем примере критерий проверки гипотезы $H_0: \theta_2 = \theta_2^0$ основывался на использовании ***t-статистики***

$$\frac{\bar{\theta}_2 - \theta_2^0}{s_{\bar{\theta}_2}},$$

значение которой можно вычислить по данным наблюдений, поскольку θ_2^0 — известное (заданное) число, а $\bar{\theta}_2$ и $s^2_{\bar{\theta}_2}$ вычисляются на основании данных наблюдений.

Каждому статистическому критерию соответствует ***критическое множество R*** значений статистики критерия, при которых гипотеза H_0 отвергается в соответствии с принятым правилом. В нашем примере таковым является множество значений указанной t-статистики, превышающих по абсолютной величине значение $t_{1-\frac{\alpha}{2}}(15)$.

Итак, ***статистический критерий*** определяется заданием

a. ***статистической гипотезы H_0***;
b. ***уровня значимости α***;
c. ***статистики критерия***;
d. ***критического множества R***.

Можно подумать, что пункты b) и d) дублируют друг друга, поскольку в нашем примере критическое множество R однозначно определяется по заданному уровню значимости α. Однако, как мы увидим в дальнейшем, одному и тому же уровню значимости можно сопоставить различные критические множества, что дает возможность выбирать множество R наиболее рациональным образом, в зависимости от выбора гипотезы H_0 (выбор ***наиболее мощного*** критерия).

Компьютерные ***пакеты программ статистического анализа данных*** первоочередное внимание уделяют проверке гипотезы

$$H_0 : \theta_j = 0$$

в рамках нормальной модели множественной линейной регрессии

$$y_i = \theta_1 x_{i1} + \mathrm{K} + \theta_p x_{ip} + \varepsilon_i , \quad i = 1, \mathrm{K}, n,$$

с $\varepsilon_i \sim i.\,i.\,d.\ N\left(0, \sigma^2\right)$. Эта гипотеза соответствует предположению исследователя о том, что j-я объясняющая переменная ***не имеет существенного значения с точки зрения объяснения изменчивости значений объясняемой переменной*** y***, так что она может быть исключена из модели.***

Для соответствующего критерия

a. $H_0 : \theta_j = 0$;

b. уровень значимости α по умолчанию обычно выбирается равным 0.05;

c. статистика критерия имеет вид

$$\frac{\bar{\theta}_j - \theta_j^0}{s_{\bar{\theta}_j}} = \frac{\bar{\theta}_j}{s_{\bar{\theta}_j}} ;$$

если гипотеза $H_0 : \theta_j = 0$ ***верна***, то эта статистика имеет t ***- распределение Стьюдента*** с $n - p$ степенями свободы,

$$\frac{\bar{\theta}_j}{s_{\bar{\theta}_j}} \sim t(n - p),$$

в связи с чем ее обычно называют ***t-статистикой (t-statistic)*** или

t-отношением (t-ratio);

d) критическое множество имеет вид

$$\left| \frac{\bar{\theta}_j}{s_{\bar{\theta}_j}} \right| > t_{1-\frac{\alpha}{2}}(n-p) .$$

При этом, в распечатках результатов ***регрессионного анализа*** (т. е. статистического анализа модели линейной регрессии) сообщаются:

- значение оценки $\bar{\theta}_j$ параметра θ_j в графе ***Коэффициенты (Coefficient)***;
- значение $s_{\bar{\theta}_j}$ знаменателя *t*-статистики в графе ***Стандартная ошибка (Std. Error)***;
- значение отношения $\bar{\theta}_j \big/ s_{\bar{\theta}_j}$ в графе ***t-статистика (t-statistic)***.

Кроме того, сообщается также

- вероятность того, что случайная величина, имеющая распределение Стьюдента с $n-p$ степенями свободы, примет значение, не меньшее по абсолютной величине, чем наблюденное значение $\left| \bar{\theta}_j \big/ s_{\bar{\theta}_j} \right|$ — в графе ***P-значение (P-value*** или ***Probability)***.

В отношении полученного при анализе *P*-значения возможны следующие варианты.

Если указываемое *P*-значение меньше выбранного уровня значимости α, то это равносильно тому, что значение *t*-статистики $\bar{\theta}_j \big/ s_{\bar{\theta}_j}$ попало в область отвержения гипотезы H_0, т. е. $\left| \bar{\theta}_j \big/ s_{\bar{\theta}_j} \right| > t_{1-\frac{\alpha}{2}}(n-p)$.В этом случае гипотеза H_0 ***отвергается***.

Если указываемое *P*-значение больше выбранного уровня значимости α, то это равносильно тому, что значение *t*-

статистики $\bar{\theta}_j \big/ s_{\bar{\theta}_j}$ не попало в область отвержения гипотезы $H_0: \theta_j = 0$, т. е. $\left| \bar{\theta}_j \big/ s_{\bar{\theta}_j} \right| < t_{1-\frac{\alpha}{2}}(n-p)$. В этом случае гипотеза H_0 ***не отвергается***.

Если (в пределах округления) указываемое *P*-значение равно выбранному уровню значимости α, то в отношении гипотезы $H_0: \theta_j = 0$ можно принять ***любое*** из двух возможных решений.

В случае, когда гипотеза $H_0: \theta_j = 0$ отвергается (*вариант 1*), говорят, что параметр θ_j ***статистически значим (statistically significant)***; это соответствует признанию того, что наличие *j*-й объясняющей переменной в правой части модели существенно для объяснения наблюдаемой изменчивости объясняемой переменной.

Напротив, в случае, когда гипотеза $H_0: \theta_j = 0$ не отвергается (*вариант 2*), говорят, что параметр θ_j ***статистически незначим (statistically unsignificant)***. В этом случае в рамках используемого статистического критерия мы не получаем убедительных аргументов против предположения о том, что $\theta_j = 0$. Это соответствует признанию того, что наличие *j*-й объясняющей переменной в правой части модели не существенно для объяснения наблюдаемой изменчивости объясняемой переменной, а следовательно, можно обойтись и без включения этой переменной в модель регрессии.

Впрочем, выводы о статистической значимости (или незначимости) того или иного параметра модели зависят от выбранного уровня значимости α: решение в пользу статистической значимости параметра может измениться на противоположное при уменьшении α, а решение в пользу ста-

тистической незначимости параметра может измениться на противоположное при уменьшении значения α.

Пример. В уже рассматривавшемся выше примере с уровнями безработицы в США получаем в распечатке $R^2 = 0.212375$ и следующую таблицу:

Переменная	**Коэф-т**	**Ст. ошибка**	**t-статист.**	**P-знач.**
1	2.294	0.410	5.589	0.0001
ZVET	0.125	0.062	2.011	0.0626

Соответственно, при выборе уровня значимости $\alpha = 0.05$ коэффициент при переменной *ZVET* признается статистически незначимым (P-значение больше уровня значимости). Однако, если выбрать $\alpha = 0.10$, то P-значение меньше уровня значимости, и коэффициент при переменной *ZVET* придется признать статистически значимым.

Пример. При исследовании зависимости спроса на куриные яйца от цены (данные были приведены ранее) получаем в распечатке $R^2 = 0.513548$ и следующую таблицу:

Переменная	Коэф-т	Ст. ошибка	t-статист.	P-знач.
1	21.100	2.304	9.158	0.0000
CENA	–18.559	5.010	-3.705	0.0026

Здесь коэффициент при объясняющей переменной *CENA* статистически значим даже при выборе $\alpha = 0.01$, так что цена является существенной объясняющей переменной.

Пример. Регрессионный анализ потребления свинины на душу населения США в зависимости от оптовых цен на свинину (данные были приведены ранее) дает значения $R^2 = 0.054483$ и

Переменная	Коэф-т	Ст. ошибка	t-статист.	P-знач.
1	77.484	13.921	5.566	0.0001
Цена	-24.775	29.794	-0.832	0.4219

В этом примере коэффициент при переменной *Цена* оказывается статистически незначимым при любом разумном выборе уровня значимости α $(\alpha = 0.01,\ \alpha = 0.05,\ \alpha = 0.10)$.

Замечание. Мы уже отмечали ранее возможность ***ложной*** корреляции между двумя переменными и, соответственно, возможность ***ложного использования*** одной из переменных в качестве объясняющей для описания изменчивости другой переменной. Проиллюстрируем такую ситуацию на основе рассмотренных нами методов регрессионного анализа.

Пример. В числе прочих подобных примеров мы получили модель линейной связи между мировым рекордом по прыжкам в высоту с шестом среди мужчин (H, в *см*) и суммарным производством электроэнергии в США (E, в *млрд. квт-час*). Мы уже указывали на высокое значение коэффициента детерминации для этой модели: $R^2 = 0.900$. Теперь мы можем привести результаты регрессионного анализа:

Переменная	Коэф-т	Ст. ошибка	t-статист.	P-знач.
1	-2625.497	420.840	-6.234	0.0000
H	7.131	0.841	8.483	0.0000

Формально, переменная H признается существенной для объяснения изменчивости переменной E, так что здесь мы сталкиваемся с ложной (паразитной) регрессией переменной E на переменную H, обусловленной наличием выраженного (линейного) тренда обеих переменных во времени.

2.8. ПРОВЕРКА ЗНАЧИМОСТИ ПАРАМЕТРОВ ЛИНЕЙНОЙ РЕГРЕССИИ И ПОДБОР МОДЕЛИ С ИСПОЛЬЗОВАНИЕМ *F*-КРИТЕРИЕВ

Приводимая ниже таблица содержит ежегодные данные о следующих показателях экономики Франции за период с 1949 по 1960 годы (млрд. франков, в ценах 1959 г.):

***Y* —** объем импорта товаров и услуг во Францию;

X_2 – валовой национальный продукт;

X_3 — потребление семей;

obs	Y	X2	X3	X4	obs	Y	X2	X3	X4
1949	15.9	149.3	4.2	108.1	1955	22.7	202.1	2.1	146.0
1950	16.4	161.2	4.1	114.8	1956	26.5	212.4	5.6	154.1
1951	19.0	171.5	3.1	123.2	1957	28.1	226.1	5.0	162.3
1952	19.1	175.5	3.1	126.9	1958	27.6	231.9	5.1	164.3
1953	18.8	180.8	1.1	132.1	1959	26.3	239	0.7	167.6
1954	20.4	190.7	2.2	137.7	1960	31.1	258	5.6	176.8

Выберем модель наблюдений в виде

$$y_i = \theta_1 x_{i1} + \theta_2 x_{i2} + \theta_3 x_{i3} + \varepsilon_i, \quad i = 1,\text{K},12,$$

где x_{ij} — значение показателя X_j в *i*-м наблюдении (*i*-му наблюдению соответствует $(1948+i)$ год, и $x_{i1} \equiv 1$ (значения «переменной» X_1, тождественно равной единице). Будем, как обычно, предполагать что $\varepsilon_1,\text{K},\varepsilon_{12} \sim$ *i. i. d.* $N(0,\sigma^2)$ и что значение σ^2 нам не известно. Регрессионный анализ дает следующие результаты: $R^2 = 0.9560$ и

Переменная	Коэф-т	Ст. ошибка	t-статист.	P-знач.
X_1	–8.570	2.869	-2.988	0.0153
X_2	0.029	0.110	0.267	**0.7953**
X_3	0.177	0.166	1.067	**0.3136**

Обращают на себя внимание выделенные P- значения. В соответствии с ними, проверка каждой отдельной гипотезы $H_0: \theta_2 = 0$, $H_0: \theta_3 = 0$ (даже при уровне значимости 0.10) приводит к решению о ее неотклонении. Соответственно, при реализации каждой из этих двух процедур проверки соответствующий параметр (θ_2 или θ_3) признается статистически незначимым. И это выглядит противоречащим весьма высокому значению коэффициента детерминации.

По-существу, вопрос стоит таким образом: необходимо построить статистическую процедуру для проверки гипотезы

$$H_0 : \theta_2 = \theta_3 = 0 ,$$

конкретизирующей значения не какого-то одного, а сразу двух коэффициентов.

И вообще, как проверить гипотезу

$$H_0 : \theta_2 = \theta_3 = \mathrm{K} = \theta_p = 0$$

(гипотеза ***значимости регрессии***) в рамках нормальной линейной модели множественной регрессии

$$y_i = \theta_1 x_{i1} + \mathrm{K} + \theta_p x_{ip} + \varepsilon_i , \quad i = 1, \mathrm{K} , n ,$$

с $x_{i1} \equiv 1$?

Соответствующий статистический критерий основывается на так называемой ***F-статистике***

$$F = \frac{\left(RSS_{H_0} - RSS \right) / \left(p - 1 \right)}{RSS / \left(n - p \right)} .$$

Здесь RSS — остаточная сумма квадратов, получаемая при оценивании полной модели (с p объясняющими переменными, включая тождественную единицу), а RSS_{H_0} — остаточная сумма квадратов, получаемая при оценивании модели с наложенными гипотезой H_0 ограничениями на параметры. Но последняя (***редуцированная***) модель имеет вид

$$y_i = \theta_1 + \varepsilon_i , \quad i = 1, \mathrm{K} , n ,$$

и применение к ней метода наименьших квадратов приводит к оценке

$$\hat{\theta}_1 = \bar{y} ,$$

так что

$$RSS_{H_0} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 = \sum_{i=1}^{n} \left(y_i - \bar{y}_i \right)^2 = TSS .$$

Следовательно,

$$\boxed{F = \frac{\left(TSS - RSS\right)/\left(p-1\right)}{RSS/\left(n-p\right)} = \frac{ESS/\left(p-1\right)}{RSS/\left(n-p\right)}}\ .$$

В некоторых пакетах статистического анализа (например, в ***EXCEL***) в распечатках результатов приводятся значения числителя и знаменателя этой статистики (в графе ***Средние квадраты — Mean Squares***).

Если $\varepsilon_1, \text{K}, \varepsilon_n \sim$ *i. i. d.* $N\left(0, \sigma^2\right)$, то указанная F-статистика, рассматриваемая как случайная величина, имеет ***при гипотезе*** $\boldsymbol{H_0}$ (т. е. ***когда действительно*** $\boldsymbol{\theta_2 = \ldots = \theta_p = 0}$) стандартное распределение $\boxed{F\left(p-1,\ n-p\right)}$, называемое ***F-распределением Фишера с (p-1) и (n-p) степенями свободы***.

Чем больше отношение ESS/RSS, тем больше есть оснований говорить о том, что совокупность переменных X_2, K, X_p действительно помогает в объяснении изменчивости объясняемой переменной Y.

В соответствии с этим, гипотеза

$$\boxed{H_0 :\ \theta_2 = \theta_3 = \text{K} = \theta_p = 0}$$

отвергается при «слишком больших» значениях F, скорее указывающих на невыполнение этой гипотезы. Соответствующее пороговое значение определяется как квантиль уровня $\left(1-\alpha\right)$ распределения $F\left(p-1,\ n-p\right)$, обозначаемая символом $\boxed{F_{1-\alpha}\left(p-1,\ n-p\right)}$.

Итак, ***гипотеза*** $\boldsymbol{H_0}$ ***отвергается***, если выполняется неравенство

$$\boxed{F = \frac{ESS/\left(p-1\right)}{RSS/\left(n-p\right)} > F_{1-\alpha}\left(p-1,\ n-p\right).}$$

При этом, вероятность ошибочного отвержения гипотезы H_0 равна α.

Статистические пакеты, выполняющие регрессионный анализ, приводят среди прочих результатов такого анализа также значение F указанной F-статистики и соответствующее ему ***P-значение (P-value)***, т. е. вероятность

$$\boxed{P\left\{F(p-1,\ n-p) > F\right\}.}$$

В частности, в рассмотренном выше примере с импортом товаров и услуг во Францию вычисленное (наблюдаемое) значение F-статистики равно $F = 97.75$, в то время как критическое значение

$$\boxed{F_{0.95}(2,\ 9) = 4.26\ .}$$

Соответственно, P-значение крайне мало — в распечатке результатов приведено значение 0.000000. Значит, здесь нет практически никаких оснований принимать составную гипотезу $H_0: \theta_2 = \theta_3 = 0$, хотя каждая из частных гипотез

$$\boxed{H_{02}:\ \theta_2 = 0} \text{ и } \boxed{H_{03}:\ \theta_3 = 0},$$

рассматриваемая сама по себе, в отрыве от второй, не отвергается.

Подобное положение встречается не так уж и редко и связано с проблемой ***мультиколлинеарности данных***. Далее мы уделим этой проблеме определенное внимание.

Что касается рассмотренных до этого примеров, то для них результаты использования F-статистики таковы.

Пример. Анализ данных об уровнях безработицы среди белого и цветного населения США приводит к следующим результатам:

$R^2 = 0.212$, $F = 4.0446$, P-значение $= 0.0626$, так что при выборе $\alpha = 0.05$ гипотеза H_0 не отвергается, а при выборе $\alpha = 0.10$ отвергается.

Пример. Анализ зависимости спроса на куриные яйца от цены приводит к значениям

$R^2 = 0.513$, $F = 13.7241$, P-значение $= 0.0026$, так что гипотеза H_0 отвергается, а регрессия признается статистически значимой.

Пример. Зависимость производства электроэнергии в США от мирового рекорда по прыжкам в высоту с шестом:

$R^2 = 0.900$, $F = 71.96$, P-значение $= 0.0000$, регрессия признается статистически значимой.

Пример. Потребление свинины в США в зависимости от оптовых цен:

$R^2 = 0.054$, $F = 0.6915$, P-значение $= 0.4219$, так что гипотеза H_0 не отвергается даже при выборе $\alpha = 0.10$.

Отметим, наконец, еще одно обстоятельство. Во всех четырех рассмотренных примерах регрессионного анализа модели ***простой (парной) линейной регрессии (p=2)*** вычисленные P-значения F-статистик совпадают с P-значениями t-статистик, используемых для проверки гипотезы $\theta_2 = 0$. Факт такого совпадения отнюдь не случаен и может быть доказан с использованием преобразований, приведенных, например, в книге Доугерти (параграф 3.11).

Применение критериев, основанных на статистиках, имеющих при нулевой гипотезе F-распределение Фишера (***F-критерии***), отнюдь не ограничивается только что рассмотренным анализом статистической значимости регрессии. Такие критерии широко применяются в процессе ***подбора модели***.

Пусть мы находимся в рамках множественной линейной модели регрессии

$$\mathrm{M}_p: \quad y_i = \theta_1 x_{i1} + \mathrm{K} + \theta_{p-q} x_{i,p-q} + \mathrm{K} + \theta_p x_{ip} + \varepsilon_i, \quad i = 1, \mathrm{K}, n,$$

с p объясняющими переменными, и гипотеза H_0 состоит в том, что в модели M_p последние q коэффициентов равны нулю, т. е.

$$H_0 : \theta_p = \theta_{p-1} = \mathrm{K} = \theta_{p-q+1} = 0 .$$

Тогда при гипотезе H_0 (т. е. в случае, когда она верна) мы имеем ***редуцированную модель***

$$\mathrm{M}_{p-q}: \quad y_i = \theta_1 x_{i1} + \mathrm{K} + \theta_{p-q} x_{i,p-q} + \varepsilon_i, \quad i = 1, \mathrm{K}, n,$$

уже с $p - q$ объясняющими переменными.

Пусть RSS - остаточная сумма квадратов в полной модели M_p, а RSS_{H_0} — остаточная сумма квадратов в редуцированной модели M_{p-q}. ***Если гипотеза H_0 верна и выполнены стандартные предположения о модели*** (в частности, $\varepsilon_1, \mathrm{K}, \varepsilon_n \sim i.\,i.\,d.\ N(0, \sigma^2)$), то тогда F-статистика

$$F = \frac{\left(RSS_{H_0} - RSS\right)/q}{RSS/(n-p)},$$

рассматриваемая как случайная величина, имеет ***при гипотезе H_0*** (т. е. ***когда действительно $\theta_p = \theta_{p-1} = \ldots = \theta_{p-q+1} = 0$***) ***F-распределение Фишера F (q, n-p) с q и (n-p) степенями свободы***.

В рассмотренном ранее случае проверки ***значимости регрессии в целом*** мы имели $q = 1$, и при этом там имело равенство $RSS_{H_0} - RSS = ESS$, которое ***не выполняется в общем случае***.

Пусть

$ESS = TSS - RSS$ — сумма квадратов, объясняемая полной моделью M_p,

$\boxed{ESS_{H_0} = TSS - RSS_{H_0}}$ — сумма квадратов, объясняемая редуцированной моделью M_{p-q}.

Тогда

$$\boxed{ESS - ESS_{H_0} = RSS_{H_0} - RSS,}$$

так что F-статистику можно записать в виде

$$\boxed{F = \frac{\left(ESS - ESS_{H_0}\right)/q}{RSS/(n-p)},}$$

из которого следует,что ***F-статистика измеряет***, в соответствующем масштабе, ***возрастание объясненной суммы квадратов вследствие включения в модель дополнительного количества объясняющих переменных.***

Естественно считать, что включение дополнительных переменных ***существенно***, если указанное возрастание объясненной суммы квадратов ***достаточно велико***. Это приводит нас к ***критерию проверки гипотезы***

$$\boxed{H_0 : \theta_p = \theta_{p-1} = \mathrm{K} = \theta_{p-q+1} = 0,}$$

основанному на ***F-статистике***

$$\boxed{F = \frac{\left(RSS_{H_0} - RSS\right)/q}{RSS/(n-p)} = \frac{\left(ESS - ESS_{H_0}\right)/q}{RSS/(n-p)}}$$

и ***отвергающему гипотезу*** H_0, когда ***наблюдаемое значение*** F этой статистики удовлетворяет неравенству

$$\boxed{F > F_{1-\alpha}(p-1,\ n-p),}$$

где α — выбранный уровень значимости критерия (вероятность ошибки 1-го рода).

Пример. В следующей таблице приведены данные по США о следующих макроэкономических показателях:

DPI — годовой совокупный располагаемый личный доход;

C — годовые совокупные потребительские расходы;

A — финансовые активы населения на начало календарного года

(все показатели указаны в млрд. долларов, в ценах 1982 г.).

obs	C82	DPI82	A82	1971	1540.3	1730.1	1902.8
1966	1300.5	1433.0	1641.6	1972	1622.3	1797.9	2011.4
1967	1339.4	1494.9	1675.2	1973	1687.9	1914.9	2190.6
1968	1405.9	1551.1	1772.6	1974	1672.4	1894.9	2301.8
1969	1458.3	1601.7	1854.7	1975	1710.8	1930.4	2279.6
1970	1491.8	1668.1	1862.2	1976	1804.0	2001.0	2308.4

Рассмотрим модель наблюдений

$$\mathrm{M}_1:\ C_t = \theta_1 + \theta_2 DPI_t + \theta_3 A_t + \theta_4 DPI_{t-1} + \varepsilon_t,\ \ t = 1,\mathrm{K},11,$$

где индексу t соответствует $(1965 + t)$ год. Это модель с 4 объясняющими переменными:

$$X_1 \equiv 1,\ X_2 = DPI,\ X_3 = A,\ X_4 = DPI(-1);$$

символ $DPI(-1)$ обозначает переменную, значения которой ***запаздывают на одну единицу времени*** относительно значений переменной, $DPI_0 = 1367{,}4$. Оценивание этой модели дает следующие результаты:

$\hat{\theta}_2 = \ 0.904$, $\quad P\text{-}value = 0.0028$;

$\hat{\theta}_3 = -0.029$, $\quad P\text{-}value = 0.8387$;

$\hat{\theta}_4 = -0.024$, $\quad P\text{-}value = 0.9337$;

$RSS = 2095.3,\ TSS = 268835,\ R^2 = 1 - (RSS/TSS) = 0.9922$;

F — статистика критерия проверки значимости регрессии в целом

$F = 297.04, \quad P\text{-}value = 0.0000.$

Регрессия имеет очень высокую статистическую значимость. Вместе с тем, каждый из коэффициентов при двух последних переменных статистически незначим, так что, в част-

ности, не следует придавать особого значения отрицательности оценок этих коэффициентов.

Используя t — критерий, мы могли бы попробовать удалить из модели какую-нибудь одну из двух последних переменных, и если оставшиеся переменные окажутся значимыми, то остановиться на модели с 3 объясняющими переменными; если же и в новой модели окажутся статистически незначимые переменные, то произвести еще одну редукцию модели.

Рассмотрим, в этой связи, модель

$$\boxed{\mathrm{M}_2:\ C_t = \theta_1 + \theta_2 DPI_t + \theta_3 A_t + \varepsilon_t,\ \ t = 1,\mathrm{K},11,}$$

с удаленной переменной $DPI(-1)$. Для нее получаем:

$\hat{\theta}_2 = 0.893$, $\quad P\text{-}value = 0.0001$;

$\hat{\theta}_3 = -0.039$, $\quad P\text{-}value = 0.6486$;

$RSS = 2098.31,\ R^2 = 0.9922$;

F-статистика критерия проверки значимости регрессии в этой модели

$F = 508.47,\quad P\text{-}value = 0.0000.$

Поскольку эдесь остается статистически незначимым коэффициент при переменной A, можно произвести дальнейшую редукцию, переходя к модели

$$\boxed{\mathrm{M}_3:\ C_t = \theta_1 + \theta_2 DPI_t + \varepsilon_t,\ \ t = 1,\mathrm{K},11.}$$

Для этой модели

$\hat{\theta}_2 = 0.843$, $\quad P\text{-}value = 0.0000$;

$RSS = 2143.57,\ R^2 = 0.9920$;

F-статистика критерия проверки значимости регрессии в этой модели

$F = 1119.7,\quad P\text{-}value = 0.0000,$

и эту модель в данном контексте можно принять за окончательную.

С другой стороны, обнаружив при анализе модели M_1 (посредством применения ***t-критериев***) статистическую незначимость коэффициентов при двух последних переменных, мы можем попробовать выяснить возможность ***одновременного исключения*** из этой модели указанных объясняющих переменных, опираясь на использование соответствующего ***F-критерия***.

Исключение двух последних переменных из модели M_1 соответствует гипотезе

$$\boxed{H_0 : \theta_3 = \theta_4 = 0,}$$

при которой модель M_1 редуцируется сразу к модели M_3. Критерий проверки гипотезы H_0 основывается на статистике

$$\boxed{F = \frac{\left(RSS_{H_0} - RSS\right)/q}{RSS/(n-p)},}$$

где RSS — остаточная сумма квадратов в модели M_1, RSS_{H_0} — остаточная сумма квадратов в модели M_3, $q = 2$ — количество зануляемых параметров, $n - p = 11 - 4 = 7$.

Для наших данных получаем значение

$$F = \frac{(2143.57 - 2095.3)/2}{2095.3/7} = 0.08,$$

которое следует сравнить с критическим значением $F_{0.95}(2,7) = 4.74$. Поскольку $F < F_{0.95}(2,7)$, мы не отвергаем гипотезу $H_0 : \theta_3 = \theta_4 = 0$ и можем сразу перейти от модели M_1 к модели M_3.

Замечание. В рассмотренном примере мы действовали двумя способами:

Дважды использовали t -критерии, сначала приняв (не отвергнув) гипотезу $\boxed{H_0\colon\ \theta_4 = 0}$ в рамках модели M_1, а затем приняв гипотезу $\boxed{H_0\colon\ \theta_3 = 0}$ в рамках модели M_2.

Однократно использовали F-критерий, приняв гипотезу $\boxed{H_0\colon \theta_3 = \theta_4 = 0}$ в рамках модели M_1.

Выводы при этих двух альтернативных подходах оказались одинаковыми. Однако, из выбора модели M_3 в подобной последовательной процедуре, ***вообще говоря, не следует*** что такой же выбор будет обязательно сделан и при применении F -критерия, сравнивающего первую и последнюю модели.

2.9. ПРОВЕРКА ЗНАЧИМОСТИ И ПОДБОР МОДЕЛИ С ИСПОЛЬЗОВАНИЕМ КОЭФФИЦИЕНТОВ ДЕТЕРМИНАЦИИ. ИНФОРМАЦИОННЫЕ КРИТЕРИИ

Ранее мы неоднократно задавались вопросом о том, как следует интерпретировать значения коэффициента детерминации R^2 с точки зрения их близости к нулю или, напротив, их близости к единице.

Естественным было бы построение статистической ***процедуры проверки значимости линейной связи*** между переменными, ***основанной на значениях коэффициента детерминации*** R^2 — ведь R^2 является статистикой, поскольку значения этой случайной величины вычисляются по данным наблюдений. Теперь мы в состоянии построить такую статистическую процедуру.

Представим F - статистику критерия проверки значимости регрессии в целом в виде

$$\boxed{F = \frac{ESS/(p-1)}{RSS/(n-p)} = \frac{ESS/TSS}{RSS/TSS}\cdot\frac{n-p}{p-1} = \frac{R^2}{1-R^2}\cdot\frac{n-p}{p-1}\ .}$$

Отсюда находим:

$(p-1)F\cdot(1-R^2)=(n-p)R^2$, $(p-1)F=\big((p-1)F+(n-p)\big)R^2$,

$$R^2=\frac{(p-1)F}{(p-1)F+(n-p)}=\frac{1}{1+\dfrac{(n-p)}{(p-1)F}}\ .$$

Большим значениям статистики F соответствуют и большие значения статистики R^2, так что гипотеза $H_0:\ \theta_2=\theta_3=\text{K}=\theta_p=0$, отвергаемая при $F>F_{crit}=F_{1-\alpha}(p-1,\ n-p)$, должна отвергаться при выполнении неравенства $R^2>R^2_{crit}$, где

$$R^2_{crit}=\frac{1}{1+\dfrac{(n-p)}{(p-1)F_{crit}}}\ .$$

При этом, вероятность ошибочного отклонения гипотезы H_0 по-прежнему равна α.

Интересно вычислить ***критические значения*** R^2_{crit} при $\alpha=0.05$ для различного количества наблюдений.

Ограничимся здесь ***простой*** линейной регрессией $(p=2)$, так что

$$R^2_{crit}=\frac{1}{1+\dfrac{(n-2)}{F_{crit}}}\ ,\quad F_{crit}=F_{0.95}(1,n-2)\ .$$

В зависимости от количества наблюдений n, получаем следующие критические значения R^2_{crit}:

n	3	4	10	20	30	40	60	120	500
R^2_{crit}	0.910	0.720	0.383	0.200	0.130	0.097	0.065	0.032	0.008

Иначе говоря, при большом количестве наблюдений даже весьма малые отклонения наблюдаемого значения R^2 от нуля оказываются достаточными для того, чтобы признать значимость регрессии, т. е. статистическую значимость коэффициента при содержательной объясняющей переменной.

Поскольку же значение R^2 равно при $p = 2$ квадрату выборочного коэффициента корреляции между объясняемой и (нетривиальной) объясняющей переменными, то аналогичный вывод справедлив и в отношении величины этого коэффициента корреляции, только получаемые результаты еще более впечатляющи:

n	**3**	**4**	**10**	**20**	**30**	**40**	**60**	**120**	**500**
$\|r_{xy}\|_{crit}$	0.953	0.848	0.618	0.447	0.360	0.311	0.254	0.179	0.089

Если сравнивать модели по величине коэффициента детерминации $\boldsymbol{R^2}$, то с этой точки зрения полная модель всегда лучше (точнее, не хуже) редуцированной — значение ***R^2 в полной модели всегда не меньше***, чем в редуцированной, просто потому, что ***в полной модели остаточная сумма квадратов не может быть больше, чем в редуцированной***.

Действительно, в полной модели с p объясняющими переменными минимизируется сумма

$$\sum_{i=1}^{n}\left(y_i - \theta_1 x_{i1} - \text{K} - \theta_p x_{ip}\right)^2$$

по всем возможным значениям коэффициентов $\theta_1, \text{K}, \theta_p$. Если мы рассмотрим редуцированную модель, например, без p-ой объясняющей переменной, то в этом случае минимизируется сумма

$$\sum_{i=1}^{n}\left(y_i - \theta_1 x_{i1} - \text{K} - \theta_{p-1} x_{i,p-1}\right)^2$$

по всем возможным значениям коэффициентов $\theta_1, \text{K}, \theta_{p-1}$, что равносильно минимизации первой суммы по всем возможным значениям $\theta_1, \text{K}, \theta_{p-1}$ при фиксированном значении $\theta_p = 0$. Но получаемый при этом минимум не может быть больше чем минимум, получаемый при минимизации первой суммы по всем возможным значениям $\theta_1, \text{K}, \theta_p$, включая и все возможные значения θ_p. Последнее означает, что RSS в полной модели не может быть меньше, чем в редуцированной модели. Поскольку же полная сумма квадратов в обеих моделях одна и та же, отсюда и вытекает заявленное выше свойство коэффициента R^2.

Чтобы сделать процедуру выбора модели с использованием R^2 более приемлемой, было предложено использовать вместо R^2 его ***скорректированный (adjusted) вариант***

$$\boxed{R^2_{adj} = 1 - \frac{RSS/(n-p)}{TSS/(n-1)}},$$

в который по-существу вводится ***штраф*** за увеличение количества объясняющих переменных. При этом,

$$R^2_{adj} = 1 - \frac{RSS}{TSS}\left(\frac{n-1}{n-p}\right) = \left(1 - \frac{RSS}{TSS}\right) + \left(\frac{RSS}{TSS} - \frac{RSS}{TSS}\left(\frac{n-1}{n-p}\right)\right)$$

$$= R^2 - \frac{RSS}{TSS}\left(\frac{n-1}{n-p} - 1\right) = R^2 - \frac{(p-1)}{(n-p)}\frac{RSS}{TSS},$$

так что

$$\boxed{R^2_{adj} < R^2}$$

при $n > p$ и $p > 1$.

При использовании коэффициента R^2_{adj} для выбора между конкурирующими моделями, лучшей признается та, для которой этот коэффициент принимает ***максимальное*** значение.

Замечание. Если при сравнении полной и редуцированных моделей оценивание каждой из альтернативных моделей производится с использованием одного и того же количества наблюдений, то тогда, как следует из формулы, определяющей R^2_{adj}, сравнение моделей по величине R^2_{adj} равносильно сравнению этих моделей по величине $S^2 = RSS/(n-p)$ или по величине $S = \sqrt{RSS/(n-p)}$. Только в последних двух случаях выбирается модель с ***минимальным*** значением S^2 (или S).

Пример. Продолжая последний пример, находим значения коэффициента R^2_{adj} при подборе моделей $\mathrm{M}_1, \mathrm{M}_2, \mathrm{M}_3$:

для M_1 — $R^2_{adj} = 0.9889$,

для M_2 — $R^2_{adj} = 0.9902$,

для M_3 — $R^2_{adj} = 0.9911$.

Таким образом, выбирая модель по максимуму R^2_{adj}, мы выберем из этих трех моделей именно модель M_3, к которой мы уже пришли до этого, пользуясь t - и F -критериями.

В этом конкретном случае сравнение всех трех моделей по величине R^2_{adj} не равносильно сравнению их по величине S^2 (или S), если модели $\mathrm{M}_2, \mathrm{M}_3$ оцениваются по всем 11 наблюдениям, представленным в таблице данных, тогда как модель M_1 оценивается только по 10 наблюдениям (одно наблюдение теряется из-за отсутствия в таблице запаздывающего значения DPI_0, соответствующего 1965 году).

Наряду со скорректированным коэффициентом детерминации, для выбора между несколькими альтернативными моделями часто используют так называемые ***информационные критерии: критерий Акаике*** и ***критерий Шварца***, также «штрафующие» за увеличение количества объясняющих переменных в модели, но несколько отличными способами.

Критерий Акаике (Akaike's information criterion — AIC). При использовании этого критерия, линейной модели с p объясняющими переменными, оцененной по n наблюдениям, сопоставляется значение

$$AIC = \ln\left(\frac{RSS_p}{n}\right) + \frac{2p}{n} + 1 + \ln 2\pi$$

где RSS_p - остаточная сумма квадратов, полученная при оценивании коэффициентов модели методом наименьших квадратов. При увеличении количества объясняющих переменных первое слагаемое в правой части уменьшается, а второе увеличивается. Среди нескольких альтернативных моделей (полной и редуцированных) предпочтение отдается модели с наименьшим значением AIC, в которой достигается определенный компромисс между величиной остаточной суммы квадратов и количеством объясняющих переменных.

Критерий Шварца (Schwarz's information criterion — SC, SIC). При использовании этого критерия, линейной модели с p объясняющими переменными, оцененной по n наблюдениям, сопоставляется значение

$$SC = \ln\left(\frac{RSS_p}{n}\right) + \frac{p\ln n}{n} + 1 + \ln 2\pi.$$

И здесь при увеличении количества объясняющих переменных первое слагаемое в правой части уменьшается, а вто-

рое увеличивается. Среди нескольких альтернативных моделей (полной и редуцированных) предпочтение отдается модели с наименьшим значением SC.

Пример. В последнем примере получаем для полной модели M_1 и редуцированных моделей M_2 и M_3 следующие значения AIC и SC.

	AIC	***SC***
M1	8.8147	8.9594
M2	8.6343	8.7428
M3	8.4738	8.5462

Предпочтительной по обоим критериям оказывается опять модель M_3.

Замечание. В рассмотренном примере все три критерия R^2_{adj}, AIC и SC выбирают одну и ту же модель. В общем случае подобное совпадение результатов выбора вовсе не обязательно.

Включение в модель большого количества объясняющих переменных часто приводит к ситуации, которую называют ***мультиколлинеарностью***.

Мы обещали ранее коснуться ***проблемы мультиколлинеарности*** и сейчас выполним это обещание. Прежде всего напомним наше предположение

(4) *матрица X^TX невырождена,* т. е. ее ***определитель отличен от нуля***:

$$\det X^T X \neq 0 ,$$

которое можно заменить условием

(4') ***столбцы матрицы X линейно независимы***.

Полная мультиколлинеарность соответствует случаю, когда предположение (4) нарушается, т. е. когда столбцы матрицы X линейно зависимы, например,

$$x_{ip} = \gamma_1 x_{i1} + \gamma_2 x_{i2} + \text{K} + \gamma_{p-1} x_{i,p-1,} \quad i = 1,\text{K} \ , n$$

(p-й столбец является линейной комбинацией остальных столбцов матрицы X). При наличии чистой мультиколлинеарности система нормальных уравнений не имеет единственного решения, так что оценка наименьших квадратов для вектора параметров (коэффициентов) попросту не определена однозначным образом.

На практике, указывая на ***наличие мультиколлинеарности***, имеют в виду осложнения со статистическими выводами в ситуациях, когда формально условие (4) выполняется, но при этом определитель матрицы $\boldsymbol{X^TX}$ близок к нулю. Указанием на то, что p-я объясняющая переменная «почти является» линейной комбинацией остальных объясняющих переменных, служит большое значение ***коэффициента возрастания дисперсии***

$$\boxed{(VIF)_p = \frac{1}{1 - R_p^2}}$$

оценки коэффициента при этой переменной вследствие наличия такой «почти линейной» зависимости между этой и остальными объясняющими переменными. Здесь R_p^2 - коэффициент детерминации при оценивании методом наименьших квадратов модели

$$\boxed{x_{ip} = \gamma_1 x_{i1} + \gamma_2 x_{i2} + \text{K} + \gamma_{p-1} x_{i,p-1} + \nu_i, \;\; i = 1, \text{K}\;, n.}$$

Если $R_p^2 = 0$, то $(VIF)_p = 1$, и это соответствует некоррелированности p-ой переменной с остальными переменными. Если же $R_p^2 \neq 0$, то тогда $(VIF)_p > 1$, и чем больше корреляция p-ой переменной с остальными переменными, тем в большей мере возрастает дисперсия оценки коэффициента при p-ой переменной по сравнению с минимально возможной величиной этой оценки.

Мы можем аналогично определить коэффициент возрастания дисперсии $(VIF)_j$ оценки коэффициента при j-ой объясняющей переменной для каждого $j = 1,\text{K}\ ,p$:

$$\boxed{(VIF)_j = \frac{1}{1 - R_j^2}\ .}$$

Здесь R_j^2 — коэффициент детерминации при оценивании методом наименьших квадратов модели линейной регрессии j-ой объясняющей переменной на остальные объясняющие переменные. Слишком большие значения коэффицентов возрастания дисперсии указывают на то, что статистические выводы для соответствующих объясняющих переменных могут быть весьма неопределенными: доверительные интервалы для коэффициентов могут быть слишком широкими и включать в себя как положительные, так и отрицательные значения, что ведет в конечном счете к признанию коэффициентов при этих переменных статистически незначимыми при использовании t - критериев.

Пример. Обращаясь опять к данным об импорте товаров и услуг во Францию, находим:

$$(VIF)_2 = (VIF)_3 = \frac{1}{1 - 0.9909} = 109.89\ .$$

Коэффициенты возрастания дисперсии для переменных X_2 и X_3 совпадают вследствие совпадения коэффициентов детерминации регрессии переменной X_2 на переменные X_1 и X_3 и регресии переменной X_3 на переменные X_1 и X_2 (взаимно обратные регрессии).

Полученные значения коэффициентов возрастания дисперсий отражают очень сильную коррелированность переменных

X_2 и X_3. (Выборочный коэффициент корреляции между этими переменными равен $Corr(X_2, X_3) = 0.995$.)

При наличии мультиколлинеарности может оказаться невозможным правильное разделение влияния отдельных объясняющих переменных. Удаление одной из переменных может привести к хорошо оцениваемой модели. Однако оставшиеся переменные примут на себя дополнительную нагрузку, так что коэффициент при каждой из этих переменных измеряет уже не собственно влияние этой переменной на объясняемую переменную, а учитывает также и часть влияния исключенных переменных, коррелированных с данной переменной.

Пример. Продолжая последний пример, рассмотрим редуцированные модели, получамые исключением из числа объясняющих переменных переменной X_2 или переменной X_3. Оценивание этих моделей приводит к следующим результатам:

$Y = -6.507 + 0.146X_2$

с $R^2 = 0.9504$ и $P - value = 0.0000$ для коэффициента при X_2;

$Y = -9.030 + 0.222X_3$

с $R^2 = 0.9556$ и $P - value = 0.0000$ для коэффициента при X_3.

В каждой из этих двух моделей коэффициенты при X_2 и X_3 имеют очень высокую статистическую значимость. В первой модели изменчивость переменной X_2 объясняет 95.04% изменчивости переменной Y; во второй модели изменчивость переменной X_3 объясняет 95.56% изменчивости переменной Y. С этой точки зрения, переменные X_2 и X_3 вполне заменяют друг друга, так что дополнение каждой из редуцированных моделей недостающей объясняющей переменной практически

ничего не добавляя к объяснению изменчивости Y (в полной модели объясняется 95.60% изменчивости переменной Y), в то же время приводит к неопределенности в оценивании коэффициентов при X_2 и X_3.

Но коэффициент при X_2 в полной модели соответствует связи между переменными X_2 и Y, очищенными от влияния переменной X_3, тогда как коэффициент при X_3 в полной модели соответствует связи между переменными X_3 и Y, очищенными от влияния переменной X_2. Поэтому неопределенность в оценивании коэффициентов при X_2 и X_3 в полной модели по-существу означает невозможность разделения эффектов влияния переменных X_2 и X_3 на переменную Y.

Приведем значения R^2_{adj}, S, AIC и SC для всех трех моделей.

	R^2_{adj}	S	AIC	SC
Полная	0.9702	1.1324	3.274	3.411
Без X_3	0.9704	1.1286	3.211	3.303
Без X_2	0.9719	1.0991	3.158	3.250

Все четыре критерия выбирают в качестве наилучшей модель с исключенной переменной X_2.

Мы не будем далее углубляться в проблему мультиколлинеарности, обсуждать другие ее последствия и возможные способы преодоления затруднений, связанных с мультиколлинеарностью. Заинтересованный читатель может обратиться по этому вопросу к более полным руководствам по эконометрике.

2.10. ПРОВЕРКА ГИПОТЕЗ О ЗНАЧЕНИЯХ КОЭФФИЦИЕНТОВ: ОДНОСТОРОННИЕ КРИТЕРИИ

Вспомним пример с потреблением текстиля. Мы подобрали линейную модель в логарифмах (с постоянными эластичностями)

$$\lg T = 1.3739 - 0.8289 \lg P + 1.1432 \lg DPI$$

(здесь T — расходы на личное потребление текстиля, P — относительная цена текстиля, DPI - располагаемый доход). В рамках этой модели представляют интерес гипотезы $\boxed{H_0: \theta_2 = -1}$ и $\boxed{H_0: \theta_3 = 1}$ о «единичной эластичности» расходов на потребление текстиля как по доходам, так и по ценам.

Построить критерии с уровнем значимости α для проверки этих гипотез можно по той же схеме, по которой строятся критерии проверки гипотез $H_0: \theta_j = 0$, только теперь для проверки гипотезы $H_0: \theta_2 = -1$ следует использовать t - статистику

$$\frac{\bar{\theta}_2 - (-1)}{s_{\bar{\theta}_2}} = \frac{\bar{\theta}_2 + 1}{s_{\bar{\theta}_2}},$$

а для проверки гипотезы $H_0: \theta_3 = 1$ — t - статистику

$$\frac{\bar{\theta}_3 - 1}{s_{\bar{\theta}_3}}.$$

Каждая из этих статистик, в случае справедливости соответствующей нулевой гипотезы, имеет распределение $t(n - p) = t(14)$. Нулевая гипотеза отвергается, если значение t - статистики превышает по абсолютной величине значение $t_{1-\frac{\alpha}{2}}(14) = t_{0.975}(14) = 2.145$.

В нашем примере

$$\frac{\bar{\theta}_2 + 1}{s_{\bar{\theta}_2}} = \frac{-0.8289 + 1}{0.0361} = 4.740 > 2.145 ,$$

$$\frac{\bar{\theta}_3 - 1}{s_{\bar{\theta}_3}} = \frac{1.1432 - 1}{0.1560} = 0.918 < 2.145 .$$

Таким образом, отклонение значения $\bar{\theta}_2$ от гипотетического значения $\theta_2 = -1$ статистически значимо — гипотеза $H_0 : \theta_2 = -1$ отвергается. В то же время, отклонение значения $\bar{\theta}_3$ от гипотетического значения $\theta_3 = 1$ не является статистически значимым, и гипотеза $H_0 : \theta_3 = 1$ не отвергается.

Замечание. Из проведенного рассмотрения видна важность не только абсолютных отклонений оценок $\bar{\theta}_j$ от гипотетических значений параметров θ_j, но и точностей оценок $\bar{\theta}_j$, измеряемых дисперсиями $D\left(\bar{\theta}_j\right)$ и оцениваемых величинами $s_{\bar{\theta}_j}$. Действительно, абсолютные величины отклонений в рассмотренном примере равны

$\left|-0.8289 + 1\right| = 0.1711$ и $\left|1.1432 - 1\right| = 0.1432$,

соответственно, т. е. отличаются не очень существенно. Однако $s_{\bar{\theta}_2}$ примерно в 4.3 раза меньше, чем $s_{\bar{\theta}_3}$, и именно такое большое отличие $s_{\bar{\theta}_2}$ и $s_{\bar{\theta}_3}$ и приводит, в конечном счете, к противоположным решениям в отношении гипотез $H_0 : \theta_2 = -1$ и $H_0 : \theta_3 = 1$.

Итак, на основании построенной процедуры гипотеза $H_0 : \theta_2 = -1$ отвергается. А что же тогда принимается?

Формально, альтернативой для $H_0 : \theta_2 = -1$ в построенном критерии является гипотеза $H_0 : \theta_2 \neq -1$, поскольку критиче-

ское множество содержит в равной степени как большие положительные, так и большие (по абсолютной величине) отрицательные значения t - статистики $\left(\bar{\theta}_2 + 1\right)\Big/ s_{\bar{\theta}_2}$. В то же время, значение $\left(\bar{\theta}_2 + 1\right)\Big/ s_{\bar{\theta}_2} = 4.740$, соответствующее отклонению $\bar{\theta}_2 - (-1) = 0.1711$, скорее говорит в пользу того, что в действительности $\theta_2 > -1$.

В этой связи, естественным представляется более определенный выбор альтернативной гипотезы, а именно, сопоставление нулевой гипотезе $H_0: \theta_2 = -1$ ***односторонней альтернативы*** $\boxed{H_A: \theta_2 > -1}$ ***(односторонняя альтернатива*** — в отличие от ***двухсторонней альтернативы*** $\boxed{H_0: \theta_2 \neq -1}$***)***. При такой постановке задачи отвержение нулевой гипотезы $H_0: \theta_2 = -1$ в пользу альтернативы $H_A: \theta_2 > -1$ производится только при больших положительных отклонениях $\bar{\theta}_2 - (-1)$, т. е. ***при больших положительных значениях*** t -статистики. Если мы отнесем к последним значения, превышающие $t_{1-\alpha}(14) = t_{0.95}(14) = 1.761$, то получим статистический критерий, у которого ошибка первого рода (уровень значимости) равна 0.05. Его критическое множество определяется соотношением

$$\frac{\bar{\theta}_2 + 1}{s_{\bar{\theta}_2}} > 1.761;$$

справа стоит теперь значение 1.761, а не 2.145, как это было при двухсторонней альтернативе. Поскольку у нас $\left(\bar{\theta}_2 + 1\right)\Big/ s_{\bar{\theta}_2} = 4.740$, мы отвергаем гипотезу $H_0: \theta_2 = -1$ в пользу гипотезы $H_A: \theta_2 > -1$.

Построим аналогичную процедуру для параметра θ_3. Именно, построим критерий уровня 0.05 для проверки гипотезы $H_0: \theta_3 = 1$ против односторонней альтернативы $H_A: \theta_3 > 1$. Критическое множество такого критерия должно состоять из значений t-статистики, превышающих $t_{0.95}(14) = 1.761$. У нас значение

$$\frac{\bar{\theta}_3 - 1}{s_{\bar{\theta}_3}} = 0.918 < 1.761$$

опять меньше порогового, так что гипотеза $H_0: \theta_3 = 1$ не отвергается в пользу $H_A: \theta_3 > 1$.

Обратим теперь внимание на то, что при рассмотрении пары конкурирующих гипотез

$$\boxed{H_0: \theta_3 = 1,\ H_A: \theta_3 > 1}$$

мы выделяем в гипотезу H_0 только одно частное значение $\theta_3 = 1$, хотя по-существу дела проблема состоит скорее в выборе между гипотезами

$$\boxed{H_0:\ 0 \le \theta_3 \le 1,\ H_A: \theta_3 > 1.}$$

Последняя ситуация коренным образом отличается от предыдущей: H_0 оказывается ***сложной гипотезой***, т. е. гипотезой, допускающей ***более одного значения параметра***, в данном случае даже бесконечно много значений параметра θ_3. В противоположность этому, в предыдущей ситуации гипотеза была H_0 ***простой***.

Какие осложнения возникают при использовании сложной нулевой гипотезы?

Возьмем, для примера, частную гипотезу $\boxed{H_0: \theta_3 = 0.5}$. Мы отвергли бы ее в пользу $H_A: \theta_3 > 1$ при

$$\frac{\bar{\theta}_3 - 0.5}{s_{\bar{\theta}_3}} > t_{0.95}(14) = 1.761 .$$

В то же время, частную гипотезу $\boxed{H_0: \theta_3 = 1}$ мы отвергаем в пользу той же $H_A: \theta_3 > 1$ при

$$\frac{\bar{\theta}_3 - 1}{s_{\bar{\theta}_3}} > t_{0.95}(14) = 1.761 .$$

Иначе говоря, при различных частных гипотезах, входящих в состав сложной нулевой гипотезы H_0: $0 \le \theta_3 \le 1$, мы получаем различные критические множества, обеспечивающие заданный уровень значимости (ошибку 1-го рода) 0.05. Построение каждого такого множества непосредственно использует конкретное гипотетическое значение $\boxed{\theta_3 = \theta_3^0}$, тогда как в рамках гипотезы H_0: $0 \le \theta_3 \le 1$ отдельное гипотетическое значение параметра θ_3 не конкретизируется.

Возникающее затруднение преодолевается, исходя из следующих соображений. Коль скоро мы не в состоянии построить единое для всех $0 \le \theta_3 \le 1$ критическое множество, вероятность попадания в которое равна $\alpha = 0.05$ при справедливости каждой отдельной частной гипотезы, следует попытаться построить единое для всех $0 \le \theta_3 \le 1$ критическое множество, вероятность попадания в которое при выполнении каждой отдельной частной гипотезы была бы не больше $\alpha = 0.05$. Такая задача реализуется путем использования критического множества, соответствующего граничному значению односторонней гипотезы, в данном случае $\theta_3 = 1$.

Действительно, пусть мы берем критическое множество $\frac{\bar{\theta}_3 - 1}{s_{\bar{\theta}_3}} > 1.761$, соответствующее граничной частной гипотезе $\theta_3 = 1$, так что

$$P\left\{\frac{\bar{\theta}_3 - 1}{s_{\bar{\theta}_3}} > 1.761\right\} = 0.05 .$$

Тогда, если в действительности верна частная гипотеза $\theta_3 = 0.5$, то

$$P\left\{\frac{\bar{\theta}_3 - 1}{s_{\bar{\theta}_3}} > 1.761 \mid \theta_3 = 0.5\right\}$$

$$= P\left\{\frac{\bar{\theta}_3 - 0.5}{s_{\bar{\theta}_3}} > 1.761 + \frac{0.5}{s_{\bar{\theta}_3}} \mid \theta_3 = 0.5\right\}$$

$$< P\left\{\frac{\bar{\theta}_3 - 0.5}{s_{\bar{\theta}_3}} > 1.761 \mid \theta_3 = 0.5\right\} = 0.05.$$

Вообще, какая бы частная гипотеза $\theta_3 = \theta_3^0 \ \left(0 \le \theta_3^0 \le 1\right)$ ни была верна, вероятность отвергнуть ее в рамках указанной процедуры не превысит 0.05.

В этом контексте, $\alpha = 0.05$ по-прежнему называется ***уровнем значимости*** критерия, тогда как понятие ошибки 1-го рода уже теряет смысл для критерия в целом. Уровень значимости ограничивает сверху ошибки 1-го рода, соответствующие частным гипотезам, входящим в состав сложной нулевой гипотезы.

Основной вывод из сказанного: при указанном подходе к построению критериев проверки сложных нулевых гипотез вида

H_0: $\theta_j < -1$ (эластичность при $\theta_j \le 0$),

H_0: $-1 \le \theta_j \le 0$ (неэластичность при $\theta_j \le 0$),

H_0: $0 \le \theta_j \le 1$ (неэластичность при $\theta_j \ge 0$),

H_0: $\theta_j > 1$ (эластичность при $\theta_j \ge 0$)

против соответствующих односторонних альтернатив можно пользоваться критериями уровня α, построенными для работы с теми же альтернативами, но при простых гипотезах $\theta_j = -1$, $\theta_j = -1$, $\theta_j = 1$, $\theta_j = 1$, соответственно.

Замечание. То же относится и к другим аналогичным парам гипотез, в которых вместо значения 1 берутся другие фиксированные граничные значения.

2.11. НЕКОТОРЫЕ ПРОБЛЕМЫ, СВЯЗАННЫЕ С ПРОВЕРКОЙ ГИПОТЕЗ О ЗНАЧЕНИЯХ КОЭФФИЦИЕНТОВ

Итак, фактически, мы уже построили критерий проверки гипотезы

$\boxed{H_0: \theta_2 < -1}$

против альтернативы

$\boxed{H_A: -1 \le \theta_2 \le 0.}$

Это тот же критерий с уровнем значимости 0.05, который был предназначен для проверки гипотезы H_0: $\theta_2 = -1$ против альтернативы H_A: $\theta_2 > -1$. Такой критерий отвергает гипотезу H_0 при

$$\frac{\bar{\theta}_2 + 1}{s_{\bar{\theta}_2}} > 1.761 ,$$

что и имеет место в нашем примере. Соответственно, нулевая гипотеза эластичности потребления текстиля по цене отвергается.

Мы также фактически построили критерий проверки гипотезы

$$H_0 :\ 0 \le \theta_3 \le 1$$

против альтернативы

$$H_A :\ \theta_3 > 1 .$$

Это тот же критерий с уровнем значимости 0.05, который был предназначен для проверки гипотезы $H_0 :\ \theta_3 = 1$ против альтернативы $H_A :\ \theta_3 > 1.$ Такой критерий отвергает гипотезу H_0 при

$$\frac{\bar{\theta}_3 - 1}{s_{\bar{\theta}_3}} > 1.761 ,$$

что не выполняется в нашем примере. Соответственно, нулевая гипотеза неэластичности потребления текстиля по доходу отвергается.

Представляет, однако, интерес то, какие решения будут приняты, если поменять местами нулевую и альтернативную гипотезы.

В отношении эластичности по цене возьмем теперь пару гипотез

$$H_0 :\ -1 \le \theta_2 \le 0 \qquad H_A :\ \theta_2 < -1 .$$

При построении соответствующего критерия достаточно обратиться к критерию для пары

$$H_0 :\ \theta_2 = -1 \qquad H_A :\ \theta_2 < -1 ,$$

который отвергает гипотезу H_0 при

$$\frac{\bar{\theta}_2 + 1}{s_{\bar{\theta}_2}} < t_\alpha(14) = t_{0.05}(14) = -1.761$$

(на левом хвосте распределения $t(14)$). Но у нас

$$\frac{\bar{\theta}_2 + 1}{s_{\bar{\theta}_2}} > 0 ,$$

так что гипотеза H_0: $\theta_2 = -1$, а значит, и H_0: $-1 \le \theta_2 \le 0$ не отвергаются в пользу H_A: $\theta_2 < -1$.

Итак, здесь нулевая гипотеза о неэластичности потребления по цене не отвергается, и это решение согласуется с отклонением нулевой гипотезы об эластичности потребления по цене.

Рассмотрим, наконец, пару гипотез

$\boxed{H_0: \theta_3 > 1 ,}$ $\boxed{H_A: 0 \le \theta_3 \le 1 .}$

Здесь мы исходим из критерия, предназначенного для пары

$\boxed{H_0: \theta_3 = 1 ,}$ $\boxed{H_A: \theta_3 < 1 ,}$

и, с учетом использования знаков равенства в этих парах, отвергаем гипотезу $H_0: \theta_3 > 1$ при

$$\frac{\bar{\theta}_3 - 1}{s_{\bar{\theta}_3}} \le t_\alpha(14) = t_{0.05}(14) = -1.761 .$$

В нашем случае

$$\frac{\bar{\theta}_3 - 1}{s_{\bar{\theta}_3}} = 0.918 > -1.761 ,$$

так что гипотеза $H_0: \theta_3 > 1$ не отвергается.

Итак, здесь нулевая гипотеза эластичности потребления по доходу не отвергается. Но ранее мы установили, что и нулевая

гипотеза неэластичности потребления по доходу также не отвергается.

Из рассмотренного примера мы должны сделать важнейший вывод:

Решения об отклонении или неотклонении одной из двух соперничающих гипотез могут быть различными, в зависимости от того, какая из двух гипотез принимается за основную (нулевую).

При решении вопроса о характере зависимости потребления текстиля от его относительной цены оба варианта выбора нулевой гипотезы дали согласованные результаты: основная гипотеза неэластичности не отвергается, а основная гипотеза эластичности отвергается.

Однако при решении вопроса о характере зависимости потребления текстиля от располагаемого дохода не отвергаются ни основная гипотеза эластичности ни основная гипотеза неэластичности. В такой ситуации каждый из исследователей, придерживающихся противоположных априорных позиций относительно эластичности или неэластичности потребления текстиля по доходу, может считать, что имеющиеся статистические данные «подтверждают» именно его гипотезу, хотя правильнее заключить, что имеющиеся статистические данные «не противоречат» его гипотезе в рамках соответствующего статистического критерия.

Мы должны теперь сделать еще одно важнейшее замечание. Пусть

$\boxed{H_0:\ \theta_j \le \theta_0}$ $\boxed{H_A:\ \theta_j > \theta_0}$.

Тогда t — статистика критерия равна

$$t = \frac{\theta_j - \theta_0}{s_{\bar{\theta}_j}}.$$

Гипотеза H_0 отвергается в пользу H_A, если

$$\frac{\bar{\theta}_j - \theta_0}{s_{\bar{\theta}_j}} > t_{1-\alpha}(n-p).$$

Но $t_{1-\alpha}(n-p) > 0$ при $\alpha < 0.5$, и это означает, что если $\bar{\theta}_j \le \theta_0$, то гипотеза H_0 не может быть отвергнута в пользу H_A.

Следовательно, если мы сначала оценим по имеющимся статистическим данным коэффициент θ_j, и только после этого выберем указанную пару гипотез для некоторого значения $\boxed{\theta_0 \ge \bar{\theta}_j}$, то в такой ситуации построенный по ***тем же*** данным указанный t-критерий ***никогда не отвергнет гипотезу*** H_0 ***в пользу*** H_A.

Аналогично, если мы, оценив θ_j, формулируем пару гипотез

$\boxed{H_0: \theta_j \ge \theta_0}$ $\boxed{H_A: \theta_j < \theta_0}$

для некоторого $\boxed{\theta_0 \le \bar{\theta}_j}$, то тогда соответствующий односторонний t-критерий, построенный по ***тем же*** данным, ***никогда не отвергнет гипотезу*** H_0 ***в пользу*** H_A.

В случае двухстороннего t-критерия

$$\left| \frac{\bar{\theta}_j - \theta_0}{s_{\bar{\theta}_j}} \right| > t_{1-\alpha}(n-p)$$

формулирование гипотезы $\boxed{H_0: \theta_j = \theta_0}$ с $\boxed{\theta_0 = \bar{\theta}_j}$, где $\bar{\theta}_j$ — оцененное значение параметра θ_j, приводит к тому, что эта гипотеза ***заведомо не будет отвергнута*** (t-статистика принимает ***нулевое значение***).

Логическая ошибка в последних трех случаях состоит в том, что теория статистических критериев строится в предположении, что гипотезы H_0 и H_A ***фиксируются до*** обращения к статистической обработке данных.

В последней ситуации априори нельзя абсолютно точно сказать, будет ли значение $\bar{\theta}_j$ больше или меньше ***заранее выбранного*** гипотетического значения θ_0.

Пример. Пусть C - совокупные расходы на личное потребление в США, Y - совокупный располагаемый доход (1970—1979 г. г., млрд. долларов в ценах 1972 г.).

Подобранная модель

$C = -67.655 + 0.979 \cdot Y$.

Уже зная, что $\bar{\theta}_2 = 0.979$, бессмысленно (или нечестно) ставить задачу проверки гипотезы $\boxed{H_0\colon\ \theta_2 < 1}$ против альтернативы $\boxed{H_A\colon\ \theta_2 \geq 1}$, поскольку на основании имеющихся наблюдений гипотеза H_0 заведомо не будет отвергнута. Она отвергается лишь при больших положительных значениях t - статистики

$$\frac{\bar{\theta}_2 - 1}{s_{\bar{\theta}_2}},$$

а у нас числитель последнего отношения принимает отрицательное значение. Другое дело, что сформулировать такую гипотезу еще до анализа статистических данных вполне разумно. Впрочем, последнее вовсе не означает, что $\bar{\theta}_2$ будет всегда меньше единицы, даже если истинное $\theta_2 < 1$.

Проверим теперь гипотезу $\boxed{H_0\colon\ \theta_2 = 0.9}$ против односторонней альтернативы $\boxed{H_A\colon\ \theta_2 > 0.9}$ в той же ситуации, но на основании данных за период с 1970 по 1981 г., $n = 12$ лет.

В этом случае $\bar{\theta}_2 = 0.952,\ s_{\bar{\theta}_2} = 0.0261$, так что t-статистика

$$t = \frac{\bar{\theta}_2 - 0.9}{s_{\bar{\theta}_2}} = \frac{0.052}{0.0261} = 1.99 .$$

Если мы используем для проверки гипотезы H_0 двусторонний t-критерий с уровнем значимости $\alpha = 0.05$, то будем отвергать H_0, когда

$$| t | > t_{crit} = t_{0.975}(10) = 2.228 .$$

Если же использовать односторонний t-критерий с уровнем значимости $\alpha = 0.05$, то будем отвергать H_0, когда

$$t > t_{crit} = t_{0.95}(10) = 1.812 .$$

В обоих случаях вероятность ошибочного отклонения гипотезы H_0 равна 0.05.

Представим теперь, что в действительности $\theta_2 = 0.95$. Тогда распределение Стьюдента $t(10)$ имеет статистика

$$\frac{\bar{\theta}_2 - 0.95}{s_{\bar{\theta}_2}} .$$

Какова вероятность того, что гипотеза H_0 будет отвергнута?

При использовании двустороннего критерия

$$P\left\{ | t | > 2.228 \,|\theta_2 = 0.95 \right\} = P\left\{ \left| \frac{\bar{\theta}_2 - 0.9}{s_{\bar{\theta}_2}} \right| > 2.228 \,|\theta_2 = 0.95 \right\}$$

$$= P\left\{ \left| \bar{\theta}_2 - 0.9 \right| > 2.228 \cdot s_{\bar{\theta}_2} \mid \theta_2 = 0.95 \right\}$$

$$= P\Big\{ \bar{\theta}_2 - 0.9 < -2.228 \cdot s_{\bar{\theta}_2} \text{ или } \bar{\theta}_2 - 0.9 > 2.228 \cdot s_{\bar{\theta}_2} \mid \theta_2 = 0.95 \Big\}$$

$$= P\Big\{ \bar{\theta}_2 - 0.95 + 0.05 < -2.228 \cdot s_{\bar{\theta}_2} \text{ или } \bar{\theta}_2 - 0.95 + 0.05 > 2.228 \cdot s_{\bar{\theta}_2} \mid \theta_2 = 0.95 \Big\}$$

$$= P\left\{ \frac{\bar{\theta}_2 - 0.95}{s_{\bar{\theta}_2}} < -2.228 - \frac{0.05}{s_{\bar{\theta}_2}} \text{ или } \frac{\bar{\theta}_2 - 0.95}{s_{\bar{\theta}_2}} > 2.228 - \frac{0.05}{s_{\bar{\theta}_2}} \mid \theta_2 = 0.95 \right\}$$

$$= P\{t(10) < -4.14 \text{ или } t(10) > 0.312\}$$

$$= P\{ t(10) < -4.14 \} + P\{ t(10) > 0.312\}$$

$$= 0.001006 + (1 - 0.619276) = 0.3817 .$$

А при использовании <u>одностороннего</u> критерия эта вероятность будет равна

$$P\{ t > 1.812 \mid \theta_2 = 0.95 \} = P\left\{ \frac{\bar{\theta}_2 - 0.9}{s_{\bar{\theta}_2}} > 1.812 \mid \theta_2 = 0.95 \right\}$$

$$= P\left\{ \frac{\bar{\theta}_2 - 0.95}{s_{\bar{\theta}_2}} > 1.812 - \frac{0.05}{s_{\bar{\theta}_2}} \mid \theta_2 = 0.95 \right\} = P\{ t(10) > -0.104 \}$$

$$= 1 - P\{ t(10) \le -0.104 \} = 1 - 0.4596 = 0.5404 .$$

Таким образом, вероятность отвергнуть ошибочную гипотезу $\boxed{H_0:\ \theta_2 = 0.9}$ в случае, когда <u>в действительности</u> $\boxed{\theta_2 = 0.95}$, равна

0.3817 — при использовании двухстороннего критерия,

0.5404 — при использовании одностороннего критерия;

две последние величины представляют собой ***мощности*** соответствующих критериев при частной альтернативе $\theta_2 = 0.95$.

Односторонний критерий имеет ***более высокую мощность*** — 0.5404 против 0.3817 у двухстороннего критерия — ***при той же вероятности ошибочного отклонения нулевой гипотезы***, равной 0.05. Такое же положение будет, если в действительности $\boxed{\theta_2 = \theta_2^0}$ и значение θ_2^0 входит в множество значений параметра θ_2, составляющих альтернативную гипотезу $\boxed{H_A\colon\ \theta_2 > 0.9}$ (т. е. $\theta_2^0 > 0$). Это говорит о ***предпочтительности одностороннего критерия*** по сравнению с двухсторонним ***при использовании в качестве альтернативной гипотезы*** $\boxed{H_A\colon\ \theta_2 > 0.9}$.

2.12. ИСПОЛЬЗОВАНИЕ ОЦЕНЕННОЙ МОДЕЛИ ДЛЯ ПРОГНОЗИРОВАНИЯ

Пусть мы имеем модель наблюдений в виде модели простой линейной регрессии

$$\boxed{y_i = \alpha + \beta x_i + \varepsilon_i\ ,\ \ i = 1,\text{K}\ ,n,}$$

и хотим дать прогноз, каким будет значение объясняемой переменной y при некотором выбранном (фиксированном) значении x^* объясняющей переменной x, если мы будем продолжать наблюдения.

Мы умеем оценивать коэффициенты α и β методом наименьших квадратов, и естественно использовать для целей прогнозирования получаемую в результате такого оценивания (подобранную) модель линейной связи

$$\boxed{y = \hat{\alpha} + \hat{\beta} x},$$

что приводит к ***прогнозируемому значению*** объясняемой переменной, равному

$$\boxed{\hat{y}^* = \hat{\alpha} + \hat{\beta} x^*},$$

Вопрос только в том, сколь надежным является выбор такого значения в качестве прогнозного. И здесь надо иметь в виду следующее.

Поскольку мы используем для прогноза оценки, полученные, исходя из модели наблюдений $y_i = \alpha + \beta x_i + \varepsilon_i$, $i = 1,\text{K},n$, то для того, чтобы этот прогноз был осмысленным, нам по необходимости приходится предполагать, что структура модели наблюдений и ее параметры не изменятся при переходе к новому наблюдению, так что соответствующее x^* значение $y = y^*$ должно описываться тем же линейным соотношением $y^* = \alpha + \beta x^* + \varepsilon^*$. В таком случае, мы по-существу имеем дело с расширенной линейной моделью с $n+1$ наблюдениями, в которой дополнительное наблюдение удовлетворяет соотношению

$y_{n+1} = y^*$, $x_{n+1} = x^*$.

При этом, случайная величина ε^* должна иметь то же распределение, что и случайные величины ε_i, $i = 1,\text{K},n$, и должна образовывать вместе с ними множество случайных величин, независимых в совокупности.

Итак, мы договорились, что в расширенной модели

$$\boxed{y^* = \alpha + \beta x^* + \varepsilon^*}.$$

Выбирая в качестве прогноза для y^* значение $\hat{y}^* = \hat{\alpha} + \hat{\beta} x^*$, мы тем самым допускаем ***ошибку прогноза***, равную

$$\hat{y}^* - y^* = \left(\hat{\alpha} + \hat{\beta} x^*\right) - \left(\alpha + \beta x^* + \varepsilon^*\right) = \left(\hat{\alpha} - \alpha\right) + \left(\hat{\beta} - \beta\right) x^* - \varepsilon^* .$$

Поскольку вычисленные оценки $\hat{\alpha}, \hat{\beta}$ являются (как мы уже выяснили выше) реализациями случайных величин, наблюдаемая ошибка прогноза также является реализацией случайной величины $\hat{Y}^* - Y^*$ и включает два источника неопределенности:

- неопределенность, связанную с отклонением вычисленных значений случайных величин $\hat{\alpha}, \hat{\beta}$ от истинных значений параметров α, β ;
- неопределенность, связанную со случайной ошибкой ε^* в $(n+1)$ - м наблюдении.

При наших стандартных предположениях о линейной модели наблюдений ошибка прогноза является случайной величиной $\hat{Y}^* - Y^*$, имеющей математическое ожидание

$$E\left(\hat{Y}^* - Y^*\right) = E\left(\hat{\alpha} - \alpha\right) + x^* E\left(\hat{\beta} - \beta\right) - E\left(\varepsilon^*\right) = 0 .$$

(Мы использовали здесь справедливые при выполнении стандартных предположений соотношения $E\left(\hat{\alpha}\right) = \alpha,\ E\left(\hat{\beta}\right) = \beta,\ E\left(\varepsilon^*\right) = 0$.)

Точность прогноза характеризуется дисперсией ошибки прогноза

$$D\left(\hat{Y}^* - Y^*\right) = D\left(\hat{\alpha} + \hat{\beta} x^* - \alpha - \beta x^* - \varepsilon^*\right) = D\left(\hat{\alpha} + \hat{\beta} x^* - \varepsilon^*\right) .$$

Здесь использован тот факт, что сумма $\alpha + \beta x^*$ неслучайна (хотя ее точное значение и не известно). Далее, из предположенной независимости случайных ошибок ε_i , $i = 1, \text{K}, n$, и ε^* вытекает независимость случайных величин $\hat{Y}^* = \hat{\alpha} + \hat{\beta} x^*$ (эта величина зависит от случайных ошибок ε_i , $i = 1, \text{K}, n$) и

ε^* (последняя не зависит от случайных ошибок ε_i , $i = 1,\text{K}\ ,n$). В силу же независимости $\hat{Y}^* = \hat{\alpha} + \hat{\beta} x^*$ и ε^*,

$$D\left(\hat{\alpha} + \hat{\beta} x^* - \varepsilon^*\right) = D\left(\hat{\alpha} + \hat{\beta} x^*\right) + D\left(\varepsilon^*\right)$$

(использовано правило сложения дисперсий). Остается заметить, что

$$\sigma^2_{\hat{Y}^*} = D\left(\hat{Y}^*\right) = D\left(\hat{\alpha} + \hat{\beta} x^*\right) = \sigma^2 \left[\frac{1}{n} + \frac{\left(x^* - \bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\right],$$

где, как обычно, $\bar{x} = \left(\sum_{i=1}^{n} x_i\right) \Big/ n$. (Мы не будем выводить эту формулу.) Таким образом,

$$\boxed{\sigma^2_{\hat{Y}^* - Y^*} = D\left(\hat{Y}^* - Y^*\right) = \sigma^2 \left[1 + \frac{1}{n} + \frac{\left(x^* - \bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\right].}$$

Если случайные ошибки ε_i , $i = 1,\text{K}\ ,n$, имеют нормальное распределение, то тогда случайные величины $\hat{Y}^* = \hat{\alpha} + \hat{\beta} x^*$ и $\hat{Y}^* - Y^*$

также имеют нормальные распределения. При этом, ошибка прогноза $\hat{Y}^* - Y^*$ имеет нормальное распределение с нулевым математическим ожиданием и дисперсией, вычисляемой по последней формуле.

Разделив разность $\hat{Y}^* - Y^*$ на квадратный корень из ее дисперсии, получаем случайную величину

$$\frac{\hat{Y}^* - Y^*}{\sigma_{\hat{Y}^* - Y^*}},$$

имеющую стандартное нормальное распределение $N(0,1)$. Заменяя в правой части выражения для $\sigma^2_{\hat{Y}^* - Y^*}$ неизвестное значение σ^2 его несмещенной оценкой $S^2 = RSS/(n-2)$, получаем оценку дисперсии $D(\hat{Y}^* - Y^*)$ в виде

$$s^2_{\hat{Y}^* - Y^*} = S^2 \left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right].$$

Заменяя, наконец, в знаменателе отношения, имеющего стандартное нормальное распределение, неизвестное значение $\sigma_{\hat{Y}^* - Y^*}$ его оценкой $s_{\hat{Y}^* - Y^*}$, приходим к t-статистике (t-отношению)

$$t = \frac{\hat{Y}^* - Y^*}{s_{\hat{Y}^* - Y^*}},$$

имеющей при выполнении сделанных предположений о модели наблюдений t-распределение Стьюдента $t(n-2)$ с $(n-2)$ степенями свободы.

Последний факт дает возможность построения $100(1-\alpha)$-процентного доверительного интервала для значения $(\hat{Y}^* - Y^*)/s_{\hat{Y}^* - Y^*}$,

а именно,

$$t_{\frac{\alpha}{2}}(n-2) \le (\hat{Y}^* - Y^*)/s_{\hat{Y}^* - Y^*} \le t_{1-\frac{\alpha}{2}}(n-2),$$

на основании которого получаем $100(1-\alpha)$-процентный доверительный интервал для Y^*:

$$\hat{Y}^* - t_{1-\frac{\alpha}{2}}(n-2)\cdot s_{\hat{Y}^*-Y^*} \le Y^* \le \hat{Y}^* + t_{1-\frac{\alpha}{2}}(n-2)\cdot s_{\hat{Y}^*-Y^*}$$

— здесь мы использовали то, что в силу симметрии распределения Стьюдента, $t_{\frac{\alpha}{2}}(K) = -t_{1-\frac{\alpha}{2}}(K)$.

Заметим, что при заданных значениях (y_i, x_i), $i = 1,\text{K},n$, (по которым строится прогноз) доверительный интервал для Y^* будет тем длинее, чем больше значение $s_{\hat{Y}^*-Y^*}$. Последнее же равно $S^2\left[1+(1/n)\right]$ при $x^* = \bar{x}$ и возрастает с ростом $\left(x^* - \bar{x}\right)^2$. Это означает, что длина доверительного интервала возрастает при удалении значения x^*, при котором строится прогноз, от среднего арифметического значений x_1,K,x_n.

Таким образом, прогнозы для значений x^*, далеко отстоящих от $\bar{x}$, становятся менее определенными, поскольку длина соответствующих доверительных интервалов для значений объясняемой переменной возрастает.

Пример. Для данных о размерах совокупного располагаемого дохода и совокупных расходах на личное потребление в США в период с 1970 по 1979 год (в млрд. долларов, в ценах 1972 года), оцененная модель линейной связи имеет вид $C = -66.595 + 0.978\cdot DPI$.

Представим себе, что мы находимся в 1979 году и ожидаем увеличения в 1980 году совокупного располагаемого дохода (в тех же ценах) до $DPI^* = 1030$ млрд. долларов. Тогда прогнозируемый по подобранной модели объем совокупных расходов на личное потребление в 1980 году равен

$\hat{C}_{1980} = -66.595 + 0.978 * 1030 = 940.75$,

так что если выбрать уровень доверия 0.95, то

$t_{crit} = t_{1-\frac{0.05}{2}}(n-2) = t_{0.975}(8) = 2.306$

и доверительный интервал для соответствующего $DPI^* = 1030$ значения $\hat{C}_{1980}$ имеет вид

$940.75 - 2.306 * 9.8228 \le \hat{C}_{1980} \le 940.75 + 2.306 * 9.8228$,

т. е.

$940.75 - 22.651 \le \hat{C}_{1980} \le 940.75 + 22.651$,

или

$$\boxed{918.099 \le \hat{C}_{1980} \le 963.401}.$$

Заметим, что интервал достаточно широк и его нижняя граница допускает даже возможность некоторого снижения уровня потребления по сравнению с предыдущим годом.

В действительности, в 1980 г. совокупный располагаемый доход достиг 1021 млрд. долларов, а совокупное потребление — 931.8 млрд. долларов. Тем самым, ошибка прогноза составила

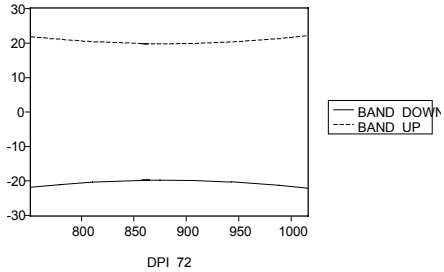
$$\frac{|940.75 - 931.8|}{931.8} \cdot 100 = 0.96\%.$$

Если бы мы исходили при прогнозе из действительного значения $DPI_{1980} = 1021$, а не из $DPI^* = 1030$, то прогнозируемое значение для C_{1980} равнялось бы 931.94 и ошибка прогноза составила всего лишь

$$\frac{|931.94 - 931.8|}{931.8} \cdot 100 = 0.015\%.$$

Проиллюстрируем, наконец, как изменяется в этом примере длина 95%-доверительных интервалов в интервале наблюдавшихся значений объясняющей переменной DPI. На гра-

фике приведены отклонения нижней и верхней границ таких интервалов от центра интервала:



В случае модели ***множественной линейной регрессии***

$$y_i = \sum_{j=1}^{p} \theta_j x_{ij} + \varepsilon_i, \quad i = 1,\text{K},n,$$

точечный прогноз значения $y^* = \sum_{j=1}^{p} \theta_j x_j^* + \varepsilon^*$, соответствующего фиксированному набору $x^* = \left(x_1^*,\text{K},x_p^*\right)$ значений объясняющих переменных, дается формулой

$$\hat{Y}^* = \sum_{j=1}^{p} \hat{\theta}_j x_j^*,$$

где $\hat{\theta}_1,\text{K},\hat{\theta}_p$ — оценки наименьших квадратов параметров $\theta_1,\text{K},\theta_p$. ***Интервальный прогноз*** имеет вид

$$\hat{Y}^* - t_{1-\frac{\alpha}{2}}(n-p)\cdot s_{\hat{Y}^*-Y^*} \le Y^* \le \hat{Y}^* + t_{1-\frac{\alpha}{2}}(n-p)\cdot s_{\hat{Y}^*-Y^*}$$

где

$$s^2_{\hat{Y}^*-Y^*} = S^2\left(1 + x^*\left(X^T X\right)^{-1}\left(x^*\right)^T\right)$$

— оценка дисперсии ошибки прогноза, а $S^2 = RSS/(n-p)$ - несмещенная оценка дисперсии σ^2 случайных ошибок.

ЧАСТЬ 3. ПРОВЕРКА ВЫПОЛНЕНИЯ СТАНДАРТНЫХ ПРЕДПОЛОЖЕНИЙ ОБ ОШИБКАХ В ЛИНЕЙНОЙ МОДЕЛИ НАБЛЮДЕНИЙ. КОРРЕКЦИЯ СТАТИСТИЧЕСКИХ ВЫВОДОВ ПРИ НАРУШЕНИИ СТАНДАРТНЫХ ПРЕДПОЛОЖЕНИЙ ОБ ОШИБКАХ

3.1. ПРОВЕРКА АДЕКВАТНОСТИ ПОДОБРАННОЙ МОДЕЛИ ИМЕЮЩИМСЯ СТАТИСТИЧЕСКИМ ДАННЫМ: ГРАФИЧЕСКИЕ МЕТОДЫ

Весь рассмотренный нами комплекс процедур получения статистических выводов для линейной модели регрессии (простой или множественной) опирается на вполне определенные предположения о модели наблюдений.

В связи с этим, большие значения коэффициента детерминации R^2 (близкие к 1) или статистическая значимость коэффициентов вовсе не обязательно говорят о том, что подобранная модель действительно хорошо ***соответствует характеру статистических данных*** (***адекватна*** статистическим данным).

В этом отношении весьма поучителен искусственный пример с четырьмя различными множествами данных, которые имеют качественно различные диаграммы рассеяния и в то же время приводят при использовании модели наблюдений

$$\boxed{y_i = \alpha + \beta x_i + \varepsilon_i,\ i = 1,\text{K},n,}$$

к одним и тем же (в пределах двух знаков после запятой) оценкам параметров, значениям коэффициента R^2 и t - статистик. Эти множества данных приведены в следующей таблице.

	Множество 1		Множество 2		Множество 3		Множество 4	
i	x	y	x	y	x	y	x	y
1	20	16.06	20	18.28	20	14.92	16	13.16
2	16	13.90	16	16.28	16	13.54	16	11.52
3	26	15.16	26	17.48	26	25.48	16	15.42
4	18	17.62	18	17.54	18	14.22	16	17.68
5	22	16.66	22	18.52	22	15.62	16	17.94
6	28	19.92	28	16.20	28	17.68	16	14.08
7	12	14.48	12	12.26	12	12.16	16	10.50
8	8	8.52	8	6.20	8	10.78	38	25.00
9	24	21.68	24	18.26	24	16.30	16	11.12
10	14	9.64	14	14.52	14	12.84	16	15.82
11	10	11.36	10	9.48	10	11.46	16	17.98

Для всех четырех множеств

подобранная модель линейной связи имеет вид $\boxed{y = 6.00 + 0.50\,x}$,

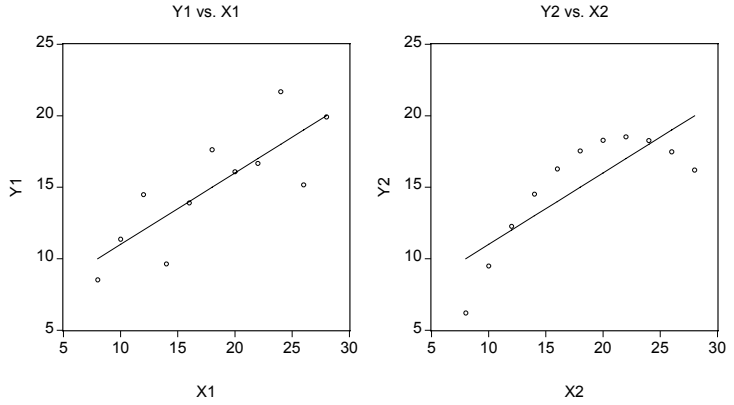
$\hat{\alpha}$ имеет (оцененную) стандартную ошибку $s_{\hat{\alpha}} = 1.12$,

$\hat{\beta}$ имеет (оцененную) стандартную ошибку $s_{\hat{\beta}} = 0.12$,

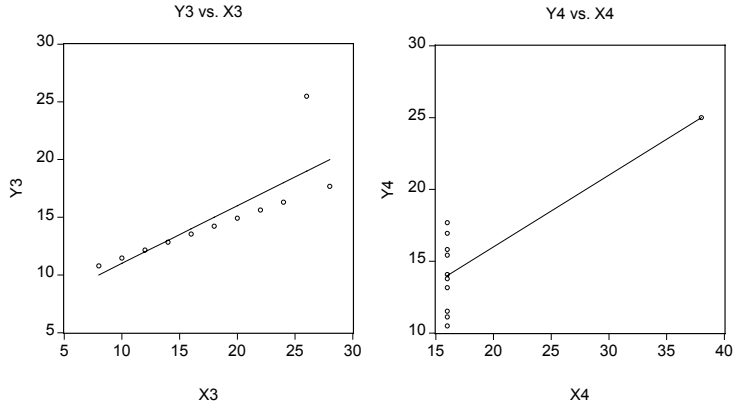
t -статистика для проверки нулевой гипотезы $\boxed{H_0\colon\ \alpha = 0}$ равна 2.67, что соответствует P -значению 0.026,

t -статистика для проверки нулевой гипотезы $\boxed{H_0\colon\ \beta = 0}$ равна 4.24, что соответствует P -значению 0.002,

$R^2 = 0.67$.



Однако диаграммы рассеяния различаются коренным образом:



Уже чисто визуальный анализ четырех диаграмм рассеяния показывает, что

только первое множество данных можно признать удовлетворительно описываемым линейной моделью наблюдений

$$y_i = \alpha + \beta x_i + \varepsilon_i \ , \ i = 1,\text{K} \ ,n \ .$$

Для второго множества более подходящей представляется модель

$$y_i = \alpha + \beta x_i + \gamma x_i^2 + \varepsilon_i,\ i = 1,\text{K},n.$$

В третьем множестве выделяется одна точка (3-е наблюдение), которая существенно влияет на наклон и положение подбираемой прямой.

Четвертое множество совершенно непригодно для подбора линейной зависимости, поскольку подобранная прямая фактически определяется наличием одного выпадающего наблюдения

Метод наименьших квадратов достаточно устойчив к малым отклонениям от стандартных предположений, в том смысле, что при таких малых отклонениях статистические выводы на основе анализа модели в основном сохраняются. Однако существенные отклонения от стандартных предположений могут серьезно исказить выводы на основе статистического анализа модели. В связи с этим необходимо

иметь возможность обнаружения отклонений от стандартных предположений,

иметь инструментарий для коррекции выявленных отклонений от стандартных предположений, позволяющий проводить строгий и информативный анализ статистических данных.

Эффективным средством обнаружения отклонений от стандартных предположений о линейной модели наблюдений

$$y_i = \theta_1 x_{i1} + \text{K} + \theta_p x_{ip} + \varepsilon_i,\ i = 1,\text{K},n,$$

является ***анализ остатков***, т. е. анализ разностей

$$e_i = y_i - \hat{y}_i,\ i = 1,\text{K},n.$$

Наблюдаемые разности $y_i - \hat{y}_i$ мы, в силу случайности значений ε_i в модели наблюдений, можем рассматривать как значения соответствующих случайных величин $Y_i - \hat{Y}_i$, за которыми сохраним те же обозначения e_i.

Если выполнены наши стандартные предположения о модели наблюдений, то остатки e_i, рассматриваемые как случайные величины $e_i = Y_i - \hat{Y}_i$, имеют нулевые математические ожидания

$$E(e_i) = 0\ ,\quad i = 1,\text{K}\ ,n\ ,$$

и дисперсии

$$D(e_i) = \sigma^2\ (1 - p_{i\,i}),\qquad i = 1,\text{K}\ ,n\ ,$$

где $p_{i\,i}$ — i-й диагональный элемент квадратной $(n \times n)$-матрицы

$$P = X\left(X^T X\right)^{-1} X^T\ .$$

Таким образом, несмотря на то, что дисперсии ошибок ε_i равны между собой при наших предположениях (все они равны σ^2), дисперсии остатков, вообще говоря, различны.

Для выравнивания дисперсий можно перейти к рассмотрению нормированных остатков

$$\frac{e_i}{\sqrt{D(e_i)}} = \frac{e_i}{\sigma\sqrt{1 - p_{i\,i}}}\ ,\quad i = 1,\text{K}\ ,n\ ,$$

для которых

$$D\left(\frac{e_i}{\sqrt{D(e_i)}}\right) = 1\ ,\quad i = 1,\text{K}\ ,n\ .$$

Поскольку значение σ^2 опять не известно, вместо нормированных остатков приходится использовать «***стьюдентизированные*» *остатки***

$$d_i = \frac{e_i}{S\sqrt{1 - p_{i\,i}}}\ ,\quad i = 1,\text{K}\ ,n\ ,$$

где, как обычно, $S^2 = RSS/(n-p)$.

Во многих пакетах программ величины p_{ii} в знаменателе правой части выражения для d_i игнорируются, что приводит к так называемым «***стандартизованным*» *остаткам***

$$c_i = \frac{e_i}{S}, \quad i = 1,\text{K},n;$$

так сделано, например, в пакете EXCEL. Практический анализ показывает, что графики остатков d_i и c_i обычно мало отличаются по характеру поведения. Поэтому для предварительного ***графического анализа*** адекватности вполне можно удовлетвориться значениями $c_i, i=1,\text{K},n$. К тому же, можно показать, что

$$\sum_{i=1}^{n} p_{ii} = p$$

(p — количество объясняющих переменных), так что если $p \ll n$ (p много меньше n), то «в среднем» значения p_{ii} достаточно малы.

Графики стандартизованных (стьюдентизированных) остатков позволяют выявлять ***типичные отклонения*** от стандартных предположений о модели наблюдений по характеру поведения остатков. При этом имеется в виду, что, по крайней мере при большом количестве наблюдений, поведение остатков $e_i, i=1,\text{K},n$, должно имитировать поведение ошибок $\varepsilon_i, i=1,\text{K},n$. Иначе говоря, поскольку мы предполагаем, что ошибки $\varepsilon_i, i=1,\text{K},n$ — независимые в совокупности случайные величины, имеющие одинаковое нормальное распределение $N(0,\sigma^2)$, то ожидаем, что поведение последовательности остатков $e_i, i=1,\text{K},n$ должно имитировать поведение последова-

тельности независимых в совокупности случайных величин, имеющих одинаковое нормальное распределение $N\left(0,\sigma^2\right)$. Соответственно, от стандартизованных остатков можно было бы ожидать поведения, похожего на поведение последовательности независимых в совокупности случайных величин, имеющих одинаковое стандартное нормальное распределение $N(0,1)$.

Строго говоря, последнее ожидание не вполне верно. Именно, хотя стандартизованные остатки и имеют распределения, близкие (хотя бы при больших n) к стандартному нормальному, они не являются взаимно независимыми случайными величинами. Это можно понять хотя бы из того, что (как мы помним) при использовании оценок наименьших квадратов алгебраическая сумма остатков равна нулю, так что каждый остаток линейно выражается через остальные остатки. Тем не менее при большом количестве наблюдений наличие такого соотношения между остатками практически не делает картину поведения стандартизованных остатков сколь-нибудь существенно отличной от поведения последовательности независимых в совокупности случайных величин, имеющих одинаковое стандартное нормальное распределение $N(0,1)$.

Наиболее часто для ***диагностики*** (проверки на наличие) типичных отклонений используют ***графики зависимости стандартизованных остатков*** (как ординат) от

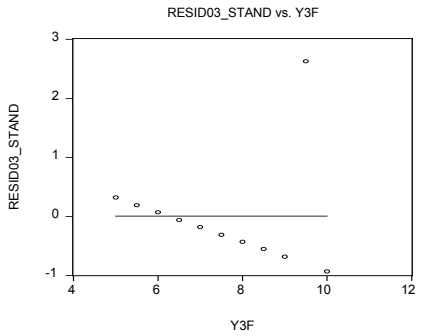
оцененных значений $\hat{y}_i = \hat{\theta}_1 x_{i1} + \mathrm{K} + \hat{\theta}_p x_{ip}$;

отдельных объясняющих переменных;

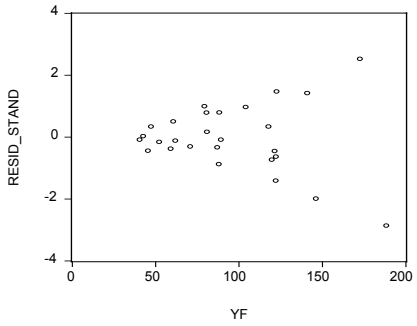
номера наблюдения, если наблюдения производятся в последовательные моменты времени с равными интервалами.

График зависимости c_i ***от*** $\hat{y}_i = \hat{\theta}_1 x_{i1} + \mathrm{K} + \hat{\theta}_p x_{ip}$ позволяет выявлять три довольно распространенных дефекта модели:

Выделяющиеся наблюдения (outliers) — наличие отдельных наблюдений, для которых либо математическое ожидание ошибки $E\left(\varepsilon_i\right)$ существенно отличается от нуля либо дисперсия ошибки $D\left(\varepsilon_i\right)$ существенно превышает величину σ^2 дисперсий остальных ошибок. Подобные наблюдения могут обнаруживать себя на указанном графике как наблюдения со «слишком большими» по абсолютной величине остатками. Такая ситуация возникает, например, при подборе прямой по третьему (из четырех рассматривавшихся выше) множеству данных:



Неоднородность дисперсий (heteroscedasticity), например, в форме той или иной функциональной зависимости $D\left(\varepsilon_i\right)$ от величины $\theta_1 x_{i1} + \mathrm{K} + \theta_p x_{ip}$. Так, если рассматриваемый график имеет вид



то это скорее всего отражает возрастание дисперсий ошибок с ростом значений $\theta_1 x_{i1} + \mathrm{K} + \theta_p x_{ip}$.

Неправильная спецификация модели в отношении множества объясняющих переменных, приводящая к нарушению соотношения $E(\varepsilon_i) \equiv 0$, так что $E(Y_i) \neq \theta_1 x_{i1} + \mathrm{K} + \theta_p x_{ip}$. Такая ситуация возникает, например, при оценивании второго множества данных из четырех рассматривавшихся выше:

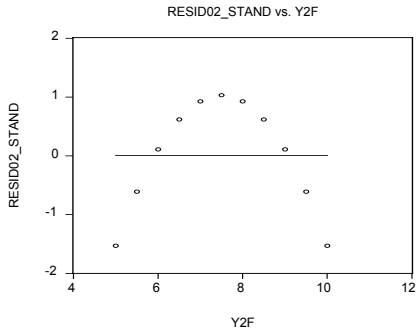


График зависимости c_i ***от значений*** x_{ij} j-й ***объясняющей переменной*** полезен для выявления ***нелинейной зависимости*** y ***от*** j-й ***объясняющей переменной***. Например, для второго из четырех искусственных множеств данных имеем

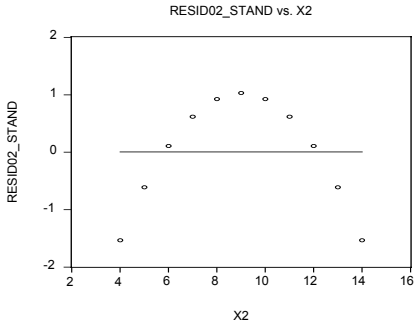
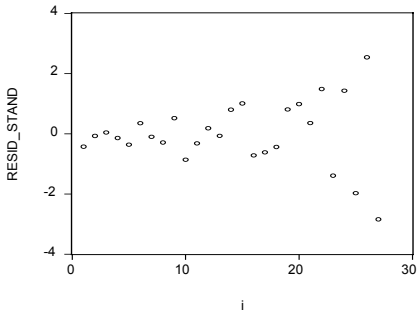
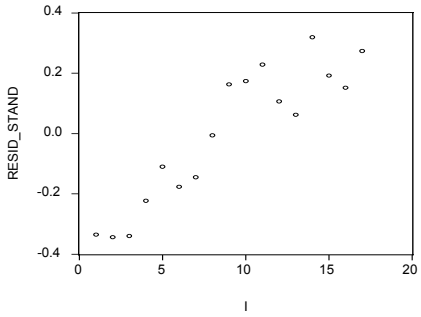


График зависимости остатков от номера наблюдения полезен в случае, когда наблюдения производятся последовательно во времени (через равные интервалы времени). По такому графику можно обнаружить

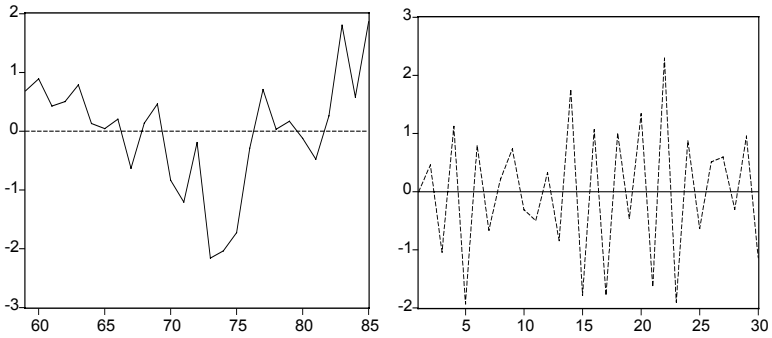
Изменение дисперсии ошибок с течением времени



Невключение в модель переменных, зависящих от времени и существенно влияющих на объясняемую переменную:



Невыполнение условия независимости в совокупности случайных ошибок $\varepsilon_i, i = \overline{1,n}$ в форме их ***автокоррелированности***. Более подробно о такой форме статистической зависимости между случайными ошибками мы поговорим позднее, а сейчас продемонстрируем, как выглядят графики остатков в случае ***положительной автокоррелированности*** (левый график) и в случае ***отрицательной автокоррелированности*** (правый график):



В первом случае проявляется ***тенденция сохранения знака*** остатка при переходе к следующему наблюдению (за положительным остатком скорее следует также положительный остаток, а за отрицательным — отрицательный). Во втором случае проявляется ***тенденция смены знака*** остатка при пере-

ходе к следующему наблюдению (за положительным остатком скорее следует отрицательный остаток, а за отрицательным — положительный).

Отдельную группу составляют графические методы проверки ***предположения о нормальности*** распределения случайных составляющих ε_i, $i = 1, \text{K}, n$.

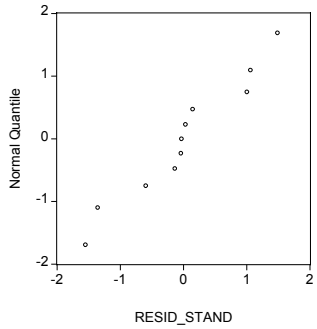
Диаграмма «квантиль-квантиль» (Q-Q plot). Для построения этой диаграммы значения стандартизованных остатков c_i, $i = 1, \text{K}, n$ упорядочивают в порядке возрастания; упорядоченные значения образуют ряд

$$c_{(1)} < c_{(2)} < \Lambda < c_{(n)}.$$

Если теперь для каждого $k = 1, \text{K}, n$ нанести в прямоугольной системе координат на плоскости точку с абсциссой $c_{(k)}$ и ординатой

$$Q_k = \Phi^{-1}\left(\frac{k - \frac{1}{2}}{n}\right)$$

(Q_k — квантиль уровня уровня $(2k-1)/(2n)$ стандартного нормального распределения), то полученные n точек $\left(c_{(k)},\ Q_k\right)$, $k = 1, \text{K}, n$, ***в случае нормальности распределения ошибок*** должны располагаться вдоль прямой, имеющей угловой коэффициент, близкий к единице. Подобное расположение имеют точки на диаграмме, построенной указанным способом по первому из четырех множеств искусственных данных:



Замечание. Если в последней процедуре не проводить стандартизацию остатков, а использовать непосредственно остатки e_i, $i=1,\text{K},n$, то полученные точки $\left(e_{(k)}, Q_k\right)$, $k=1,\text{K},n$, также будут располагаться (при нормальном распределении ошибок) вдоль некоторой прямой, но уже имеющей угловой коэффициент, не обязательно близкий к единице.

Указанное свойство диаграммы «квантиль-квантиль» основано на том, что при больших значениях n имеет место приближенное равенство

$$c_{(k)} \approx \Phi^{-1}\left(\frac{k-\frac{1}{2}}{n}\right).$$

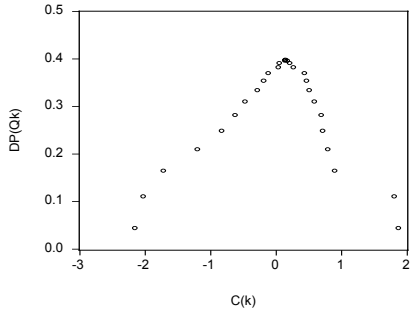
Последнему соответствует приближенное равенство

$$\Phi\left(c_{(k)}\right) \approx \frac{k-\frac{1}{2}}{n}$$

— соотношение, используемое для проверки нормальности ошибок в пакете EXCEL.

Диграмма плотности ***(DP-plot, DPP)*** отличается от диаграммы «квантиль-квантиль» тем, что по оси ординат вместо значений квантилей Q_k откладываются значения ***функции плотности стандартного нормального распределения***

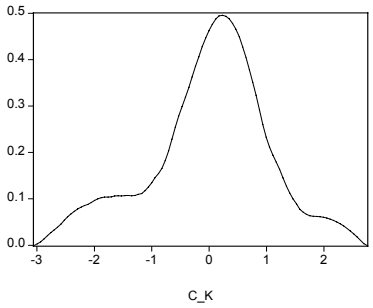
$\phi\left(c_{(k)}\right)$. Такая диаграмма дает возможность при достаточном количестве наблюдений не только проверить согласие с предположением о нормальном распределении ошибок, но и выявить характер альтернативного распределения в случае отклонения распределения ошибок от нормального. В качестве примера приведем диаграмму плотности, построенную по остаткам, полученным в результате подбора модели линейной зависимости совокупных расходов на личное потребление от совокупного располагаемого личного дохода (данные по США в млрд. долларов 1982 г., за период с 1959 по 1985 г.):



На этой диаграмме обнаруживается определенная асимметрия, что представляется не вполне согласующимся с предположением о нормальности ошибок. Однако сразу делать на этом основании вывод о нарушении такого предположения не следует. Дело в том, что при небольшом количестве наблюдений структура подобной диаграммы весьма неустойчива. Поэтому даже при заведомо нормальном распределении ошибок мы редко увидим вполне симметричную картину расположения точек на диаграмме при малом количестве наблюдений.

Ядерные (kernel) оценки плотности — еще один метод получения суждений о форме функции плотности, позволяющий, в отличие от двух предыдущих, получать график в виде

непрерывной кривой. Существует много разных вариантов таких оценок, в детали которых мы вдаваться не будем, а отметим только, что в пакете EVIEWS предлагается на выбор 8 вариантов, в рамках которых имеется еще и возможность варьирования параметров. Вариант, применяемый по умолчанию, дает для только что рассмотренных данных следующую оценку плотности распределения ошибок:



Как видим, и такой подход дает график, не очень похожий на график функции плотности стандартного нормального распределения, но это опять может быть вызвано малым количеством наблюдений (27).

3.2. ПРОВЕРКА АДЕКВАТНОСТИ ПОДОБРАННОЙ МОДЕЛИ ИМЕЮЩИМСЯ СТАТИСТИЧЕСКИМ ДАННЫМ: ФОРМАЛЬНЫЕ СТАТИСТИЧЕСКИЕ ПРОЦЕДУРЫ

Помимо графических, существует довольно много процедур, предназначенных для проверки выполнения стандартных предположений о линейной модели наблюдений, использующих ***статистические критерии проверки гипотез***. Мы остановимся только на нескольких таких процедурах. В каждой из этих процедур в качестве нулевой гипотезы берется гипотеза

$$H_0 : \varepsilon_1, \text{K}, \varepsilon_n \sim i.i.d.\ N\left(0, \sigma^2\right).$$

Однако приспособлены соответствующие критерии для выявления специфических нарушений стандартных предположений, что делает каждый из критериев особо чувствительным именно к тем нарушениям, на которые он «настроен».

Критерий Голдфелда-Квандта (Goldfeld-Quandt). Если графический анализ остатков указывает на возможную неоднородность дисперсий ошибок $D(\varepsilon_i)$, то

наблюдения, насколько это возможно, упорядочивают в порядке предполагаемого возрастания дисперсий случайных ошибок;

отбрасывают r центральных наблюдений (для более надежного разделения групп с малыми и большими дисперсиями случайных ошибок), так что для дальнейшего анализа остается $n-r$ наблюдений;

производят оценивание выбранной модели отдельно по первым $(n-r)/2$ и по последним $(n-r)/2$ наблюдениям;

вычисляют отношение $F = RSS_2/RSS_1$ остаточных сумм квадратов, полученных при подборе модели по последним $(n-r)/2$ (остаточная сумма квадратов RSS_2) и по первым $(n-r)/2$ (остаточная сумма квадратов RSS_1) наблюдениям.

При принятии решения учитывают, что если все же $D(\varepsilon_i) = \sigma^2$, $i = 1,\text{K},n$, (***дисперсии однородны***) и выполнены остальные стандартные предположения о модели наблюдений, включая предположение о нормальности ошибок, то тогда отношение

$$\boxed{F = RSS_2/RSS_1}$$

имеет F — распределение Фишера $F\left(\frac{n-r}{2}-p,\ \frac{n-r}{2}-p\right)$ с

$\left(\frac{n-r}{2}-p\right)$ и $\left(\frac{n-r}{2}-p\right)$ степенями свободы.

Гипотеза

$H_0: D(\varepsilon_i)=\sigma^2,\ i=1,\text{K},n$, (***дисперсии однородны***)

отвергается, если вычисленное значение F-отношения «слишком велико», т. е. превышает критический уровень

$$F_{1-\alpha}\left(\frac{n-r}{2}-p,\ \frac{n-r}{2}-p\right),$$

соответствующий выбранному уровню значимости α.

Критерий Дарбина-Уотсона (Durbin-Watson). Этот критерий применяется, когда наблюдения производятся последовательно во времени, с равными интервалами, и график изменения остатков во времени указывает на наличие автокоррелированности случайных составляющих ε_i модели наблюдений. Предполагается, что эта автокоррелированность определяется соотношением

$$\varepsilon_i=\rho\,\varepsilon_{i-1}+\delta_i,\ i=1,\text{K},n,$$

где $|\rho|<1$, а δ_i, $i=1,\text{K},n,$ — независимые в совокупности случайные величины, имеющие одинаковое нормальное распределение $N\left(0,\sigma_\delta^2\right)$, причем δ_i не зависит статистически от ε_{i-s} для $s>0$.

Статистика Дарбина-Уотсона определяется соотношением

$$DW=\frac{\sum_{i=2}^{n}\left(e_i-e_{i-1}\right)^2}{\sum_{i=1}^{n}e_i^2},$$

где e_1, K, e_n — остатки, получаемые при оценивании линейной модели наблюдений.

В качестве нулевой гипотезы здесь берется гипотеза

$$\boxed{H_0 : \rho = 0,}$$

соответствующая (при нашем предположении о нормальности распределения случайных ошибок) ***независимости в совокупности случайных величин*** $\varepsilon_1, \text{K}, \varepsilon_n$. В качестве альтернативной при анализе экономических данных чаще всего используют гипотезу

$$\boxed{H_A : \rho > 0,}$$

соответствующую ***положительной автокоррелированности случайных величин*** $\varepsilon_1, \text{K}, \varepsilon_n$ (т. е. тенденции преимущественного сохранения знака случайной ошибки при переходе от i-го наблюдения к $i+1$-му).

Статистика DW принимает значения в интервале от 0 до 4. Рассматриваемая как случайная величина она имеет при гипотезе $H_0 : \rho = 0$ (т. е. если эта гипотеза верна) функцию плотности $p(x)$, симметричную относительно точки $x = 2$ — середины этого интервала. Если в действительности $\rho = \rho^* > 0$, то тогда значения статистики DW тяготеют к левой границе интервала. Поэтому, в соответствии с общим подходом к построению односторонних статистических критериев, мы должны были бы для выбранного нами уровня значимости α найти соответствующее ему критическое значение d_α $\left(0 < d_\alpha < 2\right)$ и отвергать гипотезу $H_0 : \rho = 0$ в пользу $H_A : \rho > 0$ при выполнении неравенства $DW < d_\alpha$.

Однако распределение статистики Дарбина-Уотсона зависит не только от n и p, но также и от ***конкретных значений*** $x_{ij}, j = 1, \text{K}, p, \quad i = 1, \text{K}, n,$ ***объясняющих переменных***, что де-

лает неосуществимым построение таблиц критических значений этого распределения. Дарбин и Уотсон преодолели это затруднение следующим образом. Они нашли (при различных значениях n и p) ***нижнюю*** $d_{L\alpha}$ ***и верхнюю*** $d_{U\alpha}$ ***границы*** интервала, в котором только и могут находиться критические значения d_α статистики Дарбина-Уотсона, независимо от того, каковы конкретные значения $x_{ij}, j = 1,\text{K}\ ,p,\ i = 1,\text{K}\ ,n$. Иными словами,

$$\boxed{0 < d_{L\alpha} < d_\alpha < d_{U\alpha} < 2,}$$

где $d_{L\alpha}$ и $d_{U\alpha}$ ***не зависят от конкретных значений*** $x_{ij}, j = 1,\text{K}\ ,p,\ \ i = 1,\text{K}\ ,n$, а определяются только количеством наблюдений, количеством объясняющих переменных и установленным уровнем значимости критерия.

Гипотеза $\boxed{H_0 : \rho = 0}$

отвергается в пользу гипотезы $\boxed{H_A : \rho > 0}$, если $\boxed{DW < d_{L\alpha}}$;

не отвергается, если $\boxed{DW > d_{U\alpha}}$.

Если же

$$\boxed{d_{L\alpha} < DW < d_{U\alpha}\ ,}$$

то никакого вывода относительно справедливости или несправедливости гипотезы $H_0 : \rho = 0$ не делается.

При соблюдении этих правил ***вероятность ошибочного отвержения*** гипотезы $H_0 : \rho = 0$ ***не превосходит*** заданного уровня значимости α.

<u>Критерий Жарка-Бера (Jarque-Bera)</u>. Этот критерий используется в ряде пакетов статистического анализа данных (например, в EVIEWS) для ***проверки гипотезы*** H_0 ***нормальности ошибок*** в модели наблюдений, точнее,

$H_0 : \varepsilon_1, \text{K}, \varepsilon_n \sim i.i.d.\ N(0, \sigma^2)$

(значение σ^2 не конкретизируется). Если эта гипотеза верна,то при большом количестве наблюдений n статистика

$$JB = n\left[\frac{(\text{sample skewness})^2}{6} + \frac{(\text{sample kurtosis} - 3)^2}{24}\right]$$

имеет распределение, близкое к распределению хи-квадрат с двумя степенями свободы $\chi^2(2)$, функция плотности которого имеет вид

$p(x) = \frac{1}{2}e^{-x/2}$, $x > 0$.

Здесь «sample skewness» — ***выборочный коэффициент асимметрии***,

$$\text{sample skewness} = \frac{m_3}{(m_2)^{3/2}},$$

«sample kurtosis» — ***выборочный коэффициент эксцесса***,

$$\text{sample kurtosis} = \frac{m_4}{m_2^2},$$

где

$$m_k = \frac{1}{n}\sum_{i=1}^{n} e_i^k$$

и e_1, K, e_n — остатки, полученные при оценивании модели.

Если распределение ошибок действительно является нормальным, то значения выборочного коэффициента асимметрии близки к нулю, а значения выборочного коэффициента эксцесса близки к 3.

Существенное отличие выборочного коэффициента асимметрии от нуля указывает на несимметричность (относительно нуля) графика функции плотности распределения ошибок

(«скошенность» распределения). Существенное отличие от 3 выборочного коэффициента эксцесса указывает на не характерные для нормального распределения «островершинность» (при значении этого коэффициента, большем трех) или излишнюю «сглаженность» (при значении этого коэффициента, меньшем трех) графика функции плотности распределения ошибок.

При нарушении условия нормальности распределения ошибок значения статистики JB имеют тенденцию к возрастанию. Поэтому гипотеза нормальности ошибок ***отвергается, если*** значения этой статистики «слишком велики», а именно, если

$$\boxed{JB > \chi^2_{1-\alpha}(2)},$$

где $\chi^2_{1-\alpha}(2)$ — квантиль распределения $\chi^2(2)$, соответствующая уровню $1-\alpha$.

Замечание. Критерии Дарбина-Уотсона и Голдфелда-Квандта являются ***точными***, в том смысле, что они непосредственно учитывают количество наблюдений n. В противоположность этому, критерий Жарка-Бера является ***асимптотическим критерием***: распределение статистики JB хорошо приближается распределением $\chi^2(2)$ только при большом количестве наблюдений. Поэтому вполне полагаться на результаты применения критерия Жарка-Бера можно только в таких ситуациях. Помимо критерия Жарка-Бера в специализированные пакеты программ статистического анализа данных часто встраиваются и другие асимптотические критерии, например, критерии Уайта и Бройша-Годфри, которые рассматриваются ниже.

Критерий Бройша-Годфри (Breusch-Godfrey). Этот критерий используется в ряде пакетов статистического анализа

данных (например, в EVIEWS) для ***проверки гипотезы некоррелированности ошибок*** в модели наблюдений

$y_i = \theta_1 x_{i1} + \text{K} + \theta_p x_{ip} + \varepsilon_i$, $i = 1,\text{K},n.$

При наших предположениях это соответствует гипотезе независимости в совокупности случайных величин ε_i, $i = 1,\text{K},n.$ Напомним, что критерий Дарбина — Уотсона основан на рассмотрении модели наблюдений, в которой случайные составляющие ε_i связаны соотношением

$$\boxed{\varepsilon_i = \rho\,\varepsilon_{i-1} + \delta_i,\ i = 1,\text{K},n,}$$

где $|\rho| < 1$, а δ_i, $i = 1,\text{K},n,$ — независимые в совокупности случайные величины, имеющие одинаковое нормальное распределение $N\left(0,\sigma_\delta^2\right)$. В такой модели наблюдений случайные составляющие ε_i, разделенные двумя или более периодами времени и очищенные от влияния промежуточных ε_j, оказываются независимыми.

Критерий Бройша-Годфри допускает зависимость случайных составляющих ε_i, разделенных K периодами времени и также очищенных от влияния промежуточных ε_j; соответствующая модель зависимости имеет вид

$$\boxed{\varepsilon_i = a_1\varepsilon_{i-1} + \text{K} + a_K\varepsilon_{i-K} + \delta_i.}$$

Статистика этого критерия равна $\boxed{nR^2}$, где R^2 - коэффициент детерминации, получаемый при оценивании модели

$$\boxed{e_i = \gamma_1 x_{i1} + \text{K} + \gamma_p x_{ip} + \alpha_1 e_{i-1} + \text{K} + \alpha_K e_{i-K} + \nu_i,\ i = 1,\text{K},n,}$$

а e_1,K,e_n - остатки, полученные при оценивании основной модели наблюдений. (Недостающие значения e_0,K,e_{1-K} заменяются нулями.)

В рамках последней модели проверяется гипотеза

$$\boxed{H_0:\ \alpha_1 = \Lambda = \alpha_K = 0.}$$

Если эта гипотеза верна, то при большом количестве наблюдений n статистика критерия имеет распределение, близкое к распределению хи-квадрат с K степенями свободы. Гипотеза H_0 отвергается при заданном уровне значимости α, если вычисленное значение nR^2 превышает критическое значение, равное квантили уровня $1-\alpha$ указанного распределения, т. е. если

$$nR^2 > \left(nR^2\right)_{crit} = \chi^2_{1-\alpha}(K).$$

Конечно, при интерпретации результатов применения критерия Бройша-Годфри следует помнить, что этот критерий асимптотический, тогда как критерий Дарбина-Уотсона точный. Однако возможность применения критерия Дарбина-Уотсона ограничивается тем, что

он допускает зависимость «очищенных» случайных ошибок только на один шаг, т. е. $K = 1$;

он неприменим в ситуациях, когда в число объясняющих переменных включаются запаздывающие значения объясняемой переменной.

Критерий же Бройша-Годфри свободен от этих ограничений.

Критерий Уайта (White). Этот критерий используется в ряде пакетов статистического анализа данных (например, в EVIEWS) для ***проверки однородности дисперсий ошибок*** в модели наблюдений

$$y_i = \theta_1 x_{i1} + \mathrm{K} + \theta_p x_{ip} + \varepsilon_i,\ i = 1,\mathrm{K},n.$$

Критерий имеет два варианта.

Вариант I. В рамках модели

$$e_i^2 = \alpha_1 + \sum_{j=2}^{p} \alpha_j x_{ij} + \sum_{j=2}^{p} \beta_j x_{ij}^2 + \nu_i,\ i = 1,\mathrm{K},n,$$

где e_1,K,e_n - остатки, полученные при оценивании основной модели наблюдений, проверяется гипотеза

$$H_0:\ \alpha_j = \beta_j = 0,\ j = 2,\text{K},p.$$

Статистика критерия равна nR^2, где R^2 - коэффициент детерминации, получаемый при оценивании последней модели.

Если указанная гипотеза верна, то при большом количестве наблюдений n статистика критерия имеет распределение, близкое к распределению хи-квадрат с $(2p-2)$ степенями свободы. Гипотеза H_0 отвергается при заданном уровне значимости α, если вычисленное значение nR^2 превышает критическое значение, равное квантили уровня $1-\alpha$ указанного распределения, т. е. если

$$nR^2 > \left(nR^2\right)_{crit} = \chi^2_{1-\alpha}(2p-2).$$

Вариант II. В рамках модели

$$e_i^2 = \alpha_1 + \sum_{j=2}^{p}\alpha_j x_{ij} + \sum_{j=2}^{p}\sum_{k=2}^{p}\beta_{jk}x_{ij}x_{ik} + \nu_i,\ i = 1,\text{K},n,$$

где e_1,K,e_n - остатки, полученные при оценивании основной модели наблюдений, проверяется гипотеза

$$H_0:\ \alpha_j = 0,\ j = 2,\text{K},p,$$
$$\beta_{jk} = 0,\ j = 2,\text{K},p,\ k = 2,\text{K},p.$$

Статистика критерия равна nR^2, где R^2 - коэффициент детерминации, получаемый при оценивании последней модели.

Если указанная гипотеза верна, то при большом количестве наблюдений n статистика критерия имеет распределение, близкое к распределению хи-квадрат с $\left(p^2+p-2\right)/2$ степенями свободы. Гипотеза H_0 отвергается при заданном уровне значимости α, если вычисленное значение nR^2 превышает

критическое значение, равное квантили уровня $1-\alpha$ указанного распределения, т. е. если

$$\boxed{nR^2 > \left(nR^2\right)_{crit} = \chi^2_{1-\alpha}\left(\left(p^2+p-2\right)/2\right).}$$

Как и в случае критерия Бройша-Годфри, при интерпретации результатов применения обоих вариантов критерия Уайта следует помнить, что этот критерий асимптотический.

Замечание. При описании критериев Уайта мы неявно предполагали, что $x_{i1} \equiv 1$. Если постоянная не включена в исходную модель наблюдений, то в моделях, оцениваемых на втором шаге обоих вариантов критерия Уайта, суммирование следует производить, начиная с $j = 1$.

3.3. НЕАДЕКВАТНОСТЬ ПОДОБРАННОЙ МОДЕЛИ: ПРИМЕРЫ И ПОСЛЕДСТВИЯ

Пример. Рассмотрим статистические данные по США за период с 1959 по 1985 г. г. о следующих макроэкономических показателях:

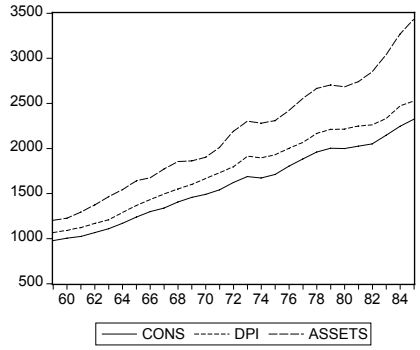
DPI — годовой совокупный располагаемый личный доход;

CONS — годовые совокупные потребительские расходы;

ASSETS — финансовые активы на конец календарного года

(все показатели в млрд. долларов, в ценах 1982 г.).

Представление об изменении этих макроэкономических показателей дает следующий график:



Рассмотрим модель наблюдений

$$CONS_t = \theta_1 + \theta_2 DPI_t + \theta_3 ASSETS_{t-1} + \varepsilon_t, \quad t = 1,\text{K},27,$$

где индексу t соответствует (1958+ t) год. Это модель с 3 объясняющими переменными:

$X_1 \equiv 1$, $X_2 = DPI$, $X_3 = ASSETS(-1)$;

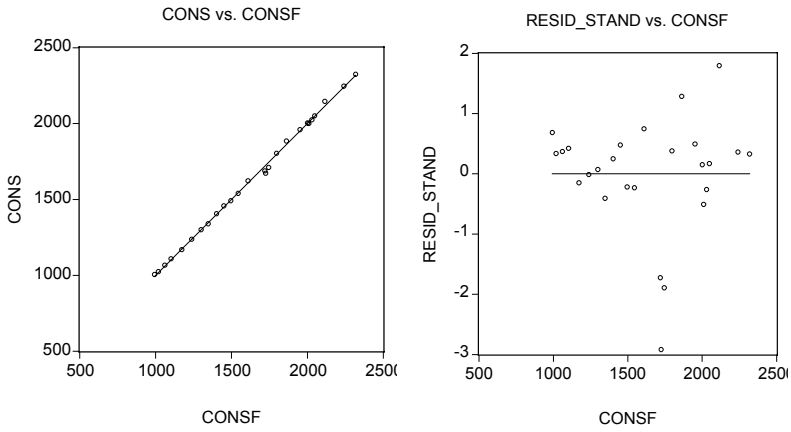
символ $ASSETS(-1)$ обозначает переменную, значения которой ***запаздывают на одну единицу времени*** относительно значений переменной $ASSETS$.

Оценивание этой модели дает следующие результаты: $R^2 = 0.9981$,

$\hat{\theta}_2 = 0.672$, P - value = 0.0000 ;

$\hat{\theta}_3 = 0.174$, P - value = 0.0069 ;

объясняющие переменные $X_2 = DPI$, $X_3 = ASSETS(-1)$ имеют высокую статистическую значимость. Ниже представлены диаграмма рассеяния для предсказанных (CONSF) и наблюдаемых (CONS) значений переменной $CONS$, а также график зависимости стандартизованных остатков $c_i = e_i / S$ (RESID_STAND) от предсказанных (CONSF) значений переменной $CONS$:



Левый график отражает высокое значение коэффициента детерминации. На правом графике заметно возрастание разброса точек относительно нулевого уровня при значениях $\hat{e}_i > 1600$.

Поскольку первый из приведенных в этом примере графиков указывает на возрастание годовых потребительских расходов с течением времени, для реализации процедуры *Goldfeld-Quandt* естественно воспользоваться уже имеющимся упорядочением наблюдений во времени (это и будет направлением ожидаемого возрастания дисперсий случайных ошибок). Заметим теперь, что вследствие использования статистических данных, начиная с 1959 года, мы не имеем в своем распоряжении значения $ASSETS_0$, соответствующего 1958 году. Поэтому реально при оценивании коэффициентов модели наблюдений мы используем только 26 (а не 27) наборов значений (x_{i1}, x_{i2}, x_{i3}), $i = 2,\text{K},27$.

Выделим из этих 26 наблюдений две группы, состоящие из первых 10 и последних 10 наборов значений (x_{i1}, x_{i2}, x_{i3}),

соответствующие периодам с 1960 по 1969 и с 1976 по 1985 годы (так что отброшены $r = 6$ центральных наблюдений). При раздельном подборе линейной модели по этим группам наблюдений получаем остаточные суммы квадратов $RSS_1 = 208.68$ и $RSS_2 = 1299.66$, соответственно, так что наблюдаемое значение F - статистики критерия *Goldfeld-Quandt* равно

$$RSS_2 / RSS_1 = 1299.66/208.68 = 6.228 .$$

Если стандартные предположения о случайных ошибках в модели наблюдений выполнены, то тогда отношение указанных остаточных сумм квадратов как случайных величин имеет F-распределение Фишера $F\left(\frac{26-6}{2}-3, \frac{26-6}{2}-3\right) = F(7,7)$. Если мы, как обычно, задаем уровень значимости равным $\alpha = 0.05$, то соответствующее этому уровню значимости критическое значение F-статистики равно

$$F_{0.95}(7,7) = 3.79 .$$

Наблюдаемое значение этой статистики 6.228 превышает критическое; поэтому гипотеза выполнения стандартных предположений об ошибках отклоняется в пользу гипотезы возрастания дисперсий $D(\varepsilon_i)$ с ростом значений $\theta_1 + \theta_2 DPI + \theta_3 ASSETS(-1)$. Заметим, наконец, что вероятность превышения случайной величиной с распределением $F(7,7)$ значения 6.228 равна

$$P\text{-}value = 0.0138.$$

Сравним результаты применения критерия Голдфелда-Квандта с результатами, получаемыми при использовании двух вариантов критерия Уайта.

При использовании первого варианта наблюдаемое значение статистики критерия равно $nR^2 = 8.884$. Поскольку $p = 3$,

то число степеней свободы соответствующего распределения хи-квадрат равно $2p-2=4$. Вероятность того, что случайная величина, имеющая такое распределение, превысит значение 8.884, равна 0.0641, так что значение $nR^2=8.884$ меньше критического, а значит, гипотеза однородности дисперсий этим вариантом критерия Уайта не отвергается.

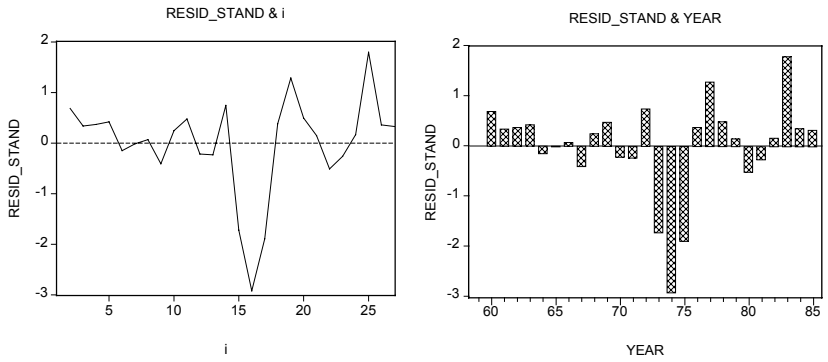
При использовании второго варианта наблюдаемое значение статистики критерия равно $nR^2=9.699$. Число степеней свободы соответствующего распределения хи-квадрат равно $\left(p^2+p-2\right)\big/2=5$. Вероятность того, что случайная величина, имеющая такое распределение, превысит значение 9.699, равна 0.0842, так что значение $nR^2=9.699$ меньше критического, а значит, гипотеза однородности дисперсий не отвергается и этим вариантом критерия Уайта.

Таким образом, статистические выводы относительно однородности дисперсий случайных составляющих в рассматриваемой модели наблюдений оказались противоречивыми: гипотеза однородности отвергается критерием Голфелда-Квандта, но не отвергается обоими вариантами критерия Уайта. Как можно объяснить такое противоречие?

- Оба варианта критерия Уайта асимптотические, тогда как критерий Голдфелда-Квандта учитывает реально имеющееся количество наблюдений.
- Оба варианта критерия Уайта являются ***критериями согласия***, не настроенными на какой-то специфический класс альтернатив гипотезе однородности, тогда как использование критерия Голдфелда-Квандта непосредственно связано с альтернативой, выраженной в форме возрастания дисперсий ошибок для соответствующего упорядочения наблюдений. И здесь проявляется общее положение: критерии, построенные с расчетом на уз-

кий класс альтернатив, оказываются более мощными по сравнению с критериями, рассчитанными на более широкий класс альтернатив, т. е. чаще отвергают нулевую гипотезу, когда она не верна.

Рассмотрим теперь график зависимости стандартизованных остатков $c_i = e_i / S$ от номера наблюдений и его вариант в виде зависимости от года наблюдения:



Здесь обращает на себя внимание наличие серий остатков одинакового знака, что сигнализирует о том, что ошибки в модели наблюдений скорее всего имеют положительную автокорреляцию. Для 26 наблюдений и $p = 3$ объясняющих переменных границы для критического значения статистики Дарбина-Уотсона при $\alpha = 0.05$ (односторонний критерий) равны

$d_{L,0.05} = 1.22$, $d_{U,0.05} = 1.55$.

В то же время, вычисленное по остаткам от оцененной модели значение статистики Дарбина-Уотсона равно

$DW = 1.01$,

что меньше нижней границы $d_{L,0.05} = 1.22$. Следовательно, нулевая гипотеза о выполнении стандартных предположений

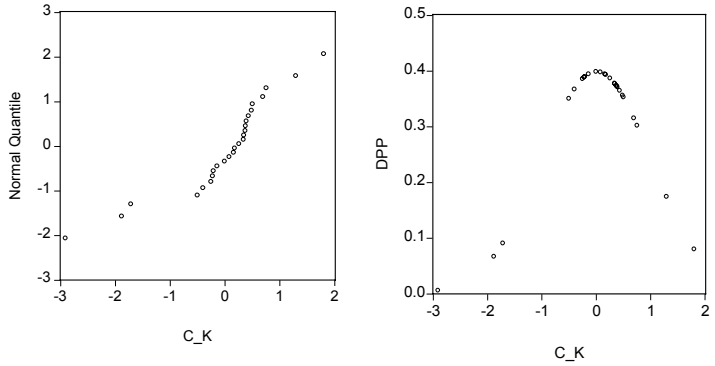
отклоняется в пользу гипотезы о положительной автокоррелированности ошибок.

Сравним результаты применения критерия Дарбина-Уотсона с результатами, получаемые при использовании критерия Бройша-Годфри.

Если исходить из допущения зависимости очищенных случайных ошибок только на один шаг $(K = 1)$, как это делается при использовании критерия Дарбина-Уотсона, то в этом случае вычисленное значение статистики критерия Бройша-Годфри равно $nR^2 = 6.068$, что соответствует P-значению, равному 0.014. Гипотеза независимости ошибок отвергается, что согласуется с результатом, полученным при использовании критерия Дарбина-Уотсона.

В то же время, если взять $K = 5$, то тогда $nR^2 = 10.331$, что соответствует P-значению, равному 0.066. Гипотеза независимости ошибок в этом случае не отвергается при установленном уровне значимости $\alpha = 0.05$, что расходится с результатом, полученным при использовании критерия Дарбина-Уотсона. Эта гипотеза не отвергается также при выборе $K = 6$ $(P\text{-}value = 0.095)$, $K = 7$ $(P\text{-}value = 0.127)$ и т.д., и это вполне объяснимо: выбор $K = 5$, $K = 6$, $K = 7$ соответствует выбору все более широких альтернатив по сравнению с $K = 1$, что приводит к уменьшению вероятности отвергнуть гипотезу независимости ошибок в случае, когда она не верна.

Проверим, наконец, предположение о нормальном распределении ошибок. Сначала рассмотрим диаграмму «квантиль-квантиль»(*Q-Q plot*) и диаграмму плотности (*DPP-plot*):



Первая диаграмма не выглядит удовлетворительной; вторая обнаруживает определенную асимметрию. Выборочный коэффициент асимметрии равен здесь −1.285, а выборочный коэффициент эксцесса равен 5.321. Оба эти значения говорят отнюдь не в пользу нормальности ошибок. Статистика критерия *Jarque-Bera* принимает значение 12.997, что соответствует $P\text{-}value = 0.0015$. Следовательно, имеющиеся данные не подтверждают гипотезу о выполнении стандартных предположений об ошибках и по этому критерию.

В связи со столь неутешительными результатами в отношении проверки гипотезы выполнения стандартных предположений в рассмотренном примере, возникает естественный вопрос о том, как именно влияют нарушения этих предположений на статистические выводы.

Неоднородность дисперсий ошибок (гетероскедастичность, heteroscedasticity). Этот вид нарушений стандартных предположений характерен для статистических данных, относящихся к одному моменту времени, но собранных по различным регионам, различным предприятиям, различным социальным группам (*данные в сечениях, cross-section data*). Неоднородность дисперсий возникает также как результат тех

или иных структурных изменений в экономике, например связанных с мировыми экономическими кризисами. Последний пример как раз и иллюстрирует подобную ситуацию: резкое возрастание абсолютных величин остатков в этом примере относится к периоду глобального нефтяного кризиса.

Последствия неоднородности дисперсий ошибок:

- Оценки дисперсий случайных величин $\hat{\theta}_1, \text{K}, \hat{\theta}_p$ (оценок коэффициентов линейной модели) оказываются смещенными.
- Построенные доверительные интервалы для $\hat{\theta}_1, \text{K}, \hat{\theta}_p$ не соответствуют заявленным уровням значимости.
- Вычисленные значения t - и F - отношений уже нельзя рассматривать как наблюдаемые значения случайных величин, имеющих t - и F -распределения, соответствующие стандартным предположениям. Поэтому сравнение вычисленных значений t - и F - отношений с квантилями указанных t - и F -распределений может приводить к ошибочным статистическим выводам в отношении гипотез о значениях коэффициентов линейной модели.

Автокоррелированность (сериальная корреляция) ошибок (autocorrelation, serial correlation). Этот вид нарушений стандартных предположений характерен для статистических данных, развернутых во времени (*продольные данные, longitudial data*). Автокоррелированность ошибок обычно возникает вследствие направильной спецификации модели, например, при невключении в модель существенной объясняющей переменной с выраженной автокорреляцией.

Последствия автокоррелированности ошибок:

- Оценка $S^2 = RSS/(n-p)$ дисперсии случайных ошибок смещена вниз в случае положительной и смещена вверх в случае отрицательной автокоррелированности ошибок.
- Оценки дисперсий случайных величин $\hat{\theta}_1, \text{K}, \hat{\theta}_p$ (оценок коэффициентов линейной модели) оказываются заниженными в случае положительной и завышенными в случае отрицательной автокоррелированности ошибок.
- Построенные доверительные интервалы для $\hat{\theta}_1, \text{K}, \hat{\theta}_p$ не соответствуют заявленным уровням значимости: в случае положительной автокоррелированности ошибок построенные интервалы неоправданно узки, а в случае отрицательной автокоррелированности ошибок неоправданно широки.
- Вычисленные значения t - и F - отношений нельзя рассматривать как наблюдаемые значения случайных величин, имеющих t - и F -распределения, соответствующие стандартным предположениям. Поэтому сравнение вычисленных значений t - и F - отношений с квантилями указанных t - и F -распределений может приводить к ошибочным статистическим выводам в отношении гипотез о значениях коэффициентов линейной модели. Вычисленные значения t - и F - отношений завышены в случае положительной и занижены в случае отрицательной автокоррелированности ошибок.

При обнаружении нарушений стандартных предположений следует либо улучшить спецификацию модели, привлекая подходящие дополнительные объясняющие переменные, либо использовать для оценивания коэффициентов и оценивания дисперсий коэффициентов модели специальные методы оценивания, принимающие во внимание обнаруженные наруше-

ния (далее мы рассмотрим два таких метода: *взвешенный метод наименьших квадратов* и *авторегрессионное преобразование переменных*).

3.4. КОРРЕКЦИЯ СТАТИСТИЧЕСКИХ ВЫВОДОВ ПРИ НАЛИЧИИ ГЕТЕРОСКЕДАСТИЧНОСТИ (НЕОДНОРОДНОСТИ ДИСПЕРСИЙ ОШИБОК)

Пример. Для исследования вопроса о зависимости количества руководящих работников от размера предприятия были собраны статистические данные по 27 промышленным предприятиям. Далее обозначено:

x_i — численность персонала на *i*-м предприятии,

y_i — количество руководителей на *i*-м предприятии.

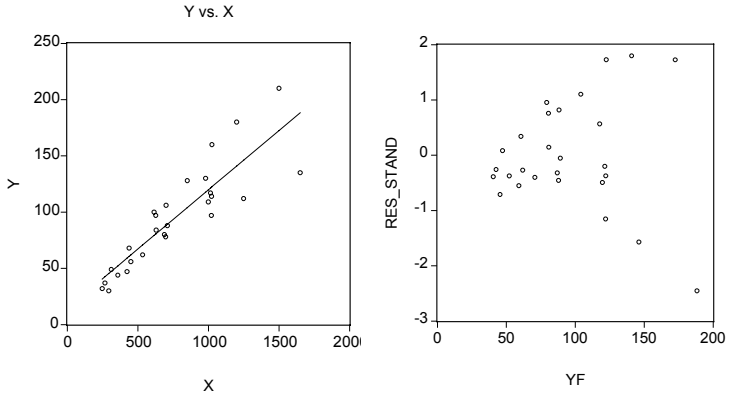
Оцениваем линейную модель наблюдений

$$y_i = \alpha + \beta x_i + \varepsilon_i,\ i = 1,\text{K},27.$$

Регрессионный анализ дает следующие результаты: R^2= 0.776 и

Variable	Coefficient	Std. Error	t-Statistic	P-value.
1	14.448	9.562	1.511	0.1433
X	0.105	0.011	9.303	0.0000

Следующие два графика демонстрируют диаграмму рассеяния с подобранной прямой $y = 14.448 + 0.105x$ (левый график) и зависимость стандартизованных остатков $c_i = e_i / S$ от значений $\hat{y}_i = 14.448 + 0.105x_i$ (правый график).



Похоже, что имеет место тенденция линейного возрастания абсолютных величин остатков с ростом $\hat{y}$, соответствующая наличию приближенной зависимости вида $\boxed{D(\varepsilon_i) = \sigma_i^2 = \sigma^2 \cdot x_i^2}$ для дисперсий ошибок. Чтобы погасить такую неоднородность дисперсий, разделим обе части соотношения $y_i = \alpha + \beta x_i + \varepsilon_i$ на x_i:

$$\frac{y_i}{x_i} = \alpha \frac{1}{x_i} + \beta + \frac{\varepsilon_i}{x_i},$$

т. е. перейдем к модели наблюдений

$$\boxed{y_i^* = \beta + \alpha x_i^* + \varepsilon_i^*,}$$

где

$$\boxed{y_i^* = \frac{y_i}{x_i},\ x_i^* = \frac{1}{x_i},\ \varepsilon_i^* = \frac{\varepsilon_i}{x_i}.}$$

Если действительно выполняется соотношение $D(\varepsilon_i) = \sigma_i^2 = \sigma^2 \cdot x_i^2$, то тогда в преобразованной модели

$$\boxed{E(\varepsilon_i^*) = 0, D(\varepsilon_i^*) = \frac{1}{x_i^2} D(\varepsilon_i) = \sigma^2,}$$

т. е. неоднородность дисперсий ошибок преодолевается.

Результаты оценивания преобразованной модели:

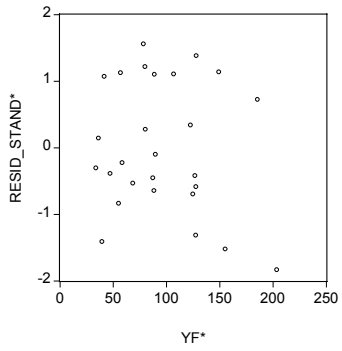
Variable	Coefficient	Std. Error	t-Statistic	P-value.
1	0.121	0.009	13.445	0.0000
1/x	3.803	4.570	0.832	0.4131

В исходных переменных это соответствует модели линейной связи

$$y = 3.803 + 0.121x\,.$$

Отметим уменьшение оцененных стандартных ошибок оценок обоих параметров α и β. Именно на эти значения следует опираться при построении доверительных интервалов для этих параметров. Средними точками этих интервалов будут, соответственно, $\hat{\alpha} = 3.803$ и $\hat{\beta} = 0.121$. Следующий график показывает характер зависимости стандартизованных остатков в преобразованной модели от $\hat{y}^*$.

На сей раз неоднородности дисперсий остатков (по крайней мере явной) не обнаруживается.



Рассмотрим внимательнее наши действия при оценивании преобразованной модели. Оценки коэффициентов, приведенные в последней таблице, получены применением метода наи-

меньших квадратов к модели наблюдений $y_i^* = \beta + \alpha x_i^* + \varepsilon_i^*$, т. е. путем минимизации суммы квадратов

$$\sum_{i=1}^{n}\left(y_i^* - \beta - \alpha x_i^*\right)^2,$$

которую, вспоминая, что обозначают переменные со звездочками, можно записать в виде

$$\boxed{\sum_{i=1}^{n}\left(\frac{y_i}{x_i} - \beta - \alpha\,\frac{1}{x_i}\right)^2 = \sum_{i=1}^{n}\frac{1}{x_i^2}\left(y_i - \alpha - \beta\, x_i\right)^2.}$$

Обозначая теперь

$$\boxed{w_i = \frac{1}{x_i^2},}$$

получаем, что задача минимизации суммы квадратов отклонений в преобразованной модели равносильна задаче минимизации ***взвешенной суммы квадратов*** отклонений в <u>исходной</u> (непреобразованной) модели. Величина w_i интерпретируется в этом контексте как ***вес***, приписываемый квадрату отклонения в i - м наблюдении. Этот вес будет тем меньше, чем больше значение x_i^2, которое в силу наших предположений пропорционально дисперсии случайной ошибки $D\left(\varepsilon_i\right) = \sigma_i^2 = \sigma^2 \cdot x_i^2$ в i -м наблюдении. Следовательно, чем больше дисперсия случайной ошибки ε_i, тем меньше вес, с которым входит квадрат отклонения в i -м наблюдении в минимизируемую сумму.

Имея в виду, что оценивание преобразованной модели наблюдений сводится к минимизации суммы

$$\boxed{\sum_{i=1}^{n} w_i\left(y_i - \alpha - \beta\, x_i\right)^2,}$$

рассмотренный метод оценивания называют ***взвешенным методом наименьших квадратов*** (хотя точнее его следовало бы называть методом наименьших взвешенных квадратов).

Замечание. В некоторых руководствах по эконометрике и в некоторых пакетах статистического анализа данных (например, в пакете EVIEWS) используется несколько иное равносильное представление минимизируемой суммы квадратов в преобразованной модели наблюдений:

$$\sum_{i=1}^{n}\left(w_i\left(y_i-\alpha-\beta\, x_i\right)\right)^2.$$

В этом случае вес приписывается не квадрату отклонения, а самому отклонению $\left(y_i-\alpha-\beta\, x_i\right)$. Разумеется, в рассмотренном примере при таком определении веса последний будет равен

$$w_i=\frac{1}{x_i}\ .$$

На это обстоятельство следует обратить внимание при спецификации весов в процедурах, реализующих взвешенный метод наименьших квадратов.

Обратим теперь внимание на то, в каком виде выдается информация о результатах применения взвешенного метода наименьших квадратов на примере пакета EVIEWS. При этом используем данные из рассмотренного выше примера. Согласно сказанному в Замечании, при обращении к процедуре оценивания взвешенным методом наименьших квадратов в условиях нашего примера мы специфицируем веса как $w=1/x$.

Протокол оценивания имеет следующий вид:

Dependent Variable: Y
Method: Least Squares
Date: Time:
Sample: 1 27
Included observations: 27
Weighting series: 1/X

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3.803296	4.569745	0.832277	0.4131
X	0.120990	0.008999	13.44540	0.0000
Weighted Statistics				
R-squared	0.026960	Mean dependent var		74.04946
Adjusted R-squared	–0.011961	S. D. dependent var		13.08103
S. E. of regression	13.15902	Akaike info criterion		8.063280
Sum squared resid	4328.998	Schwarz criterion		8.159268
Log likelihood	-106.8543	F-statistic		180.7789
Durbin-Watson stat	2.272111	Prob (F-statistic)		0.000000
Unweighted Statistics				
R-squared	0.758034	Mean dependent var		94.44444
Adjusted R-squared	0.748355	S. D. dependent var		45.00712
S. E. of regression	22.57746	Sum squared resid		12743.54
Durbin-Watson stat	2.444541			

В этом протоколе приводятся значения двух видов статистик:

- ***Weighted Statistics (взвешенные статистики)*** — это статистики, основанные на остатках, получаемых по взвешенным данным, т. е. на остатках $\boxed{e_i^* = y_i^* - \beta - \alpha x_i^*}$ в ***преобразованной модели***.
- ***Unweighted Statistics (невзвешенные статистики)*** — это статистики, основанные на «остатках» $\boxed{u_i = y_i - \alpha^{WLS} - \beta^{WLS} x_i}$, т. е. на отклонениях наблюдаемых значений объясняемой переменной y от значе-

ний, предсказываемых линейной моделью связи, в качестве параметров которой берутся их оценки $\alpha^{WLS}, \beta^{WLS}$, полученные в преобразованной модели.

Отметим весьма низкое (0.2696) значение коэффициента детерминации в преобразованной модели. Однако это обстоятельство не должно нас волновать — линейная связь в преобразованной модели значима, о чем говорит весьма высокое значение F-статистики, равное 180.7789, и соответствующее ему P-значение 0.0000 (см. *Weighted Statistics*). В конечном счете нас интересует значение R^2, находящееся в части протокола, соответствующей невзвешенным статистикам, а это значение достаточно велико (0.7580).

Отметим еще, что приведенные в начале таблицы значения оценок параметров, их стандартных ошибок и t-статистик, а также P-значения соответствуют величинам, полученным на стадии оценивания преобразованной модели.

Заметим, наконец, что значение $R^2 = 0.758$, указанное в числе невзвешенных статистик, отличается от значения $R^2 = 0.776$, полученного нами при оценивании исходной (непреобразованной) модели наблюдений. Причина этого, разумеется, в том, что при вычислении значения $R^2 = 0.776$ использовались остатки

$$e_i = y_i - \hat{\alpha} - \hat{\beta}\, x_i ,$$

где $\hat{\alpha}$, $\hat{\beta}$ — оценки наименьших квадратов параметров исходной модели, полученные без использования взвешивания отклонений.

Мы уже отмечали выше, что результатом неоднородности дисперсий случайных ошибок в модели наблюдений является смещение оценок дисперсий случайных величин $\hat{\theta}_1, \ldots, \hat{\theta}_p$. В

то же время, наличие такого нарушения стандартных предположений оставляет оценки $\hat{\theta}_1,\text{K},\hat{\theta}_p$ ***несмещенными***. В связи с этим, один из методов коррекции статистических выводов при неоднородности дисперсий ошибок состоит в использовании обычных оценок наименьших квадратов (***OLS-оценок, Ordinary Least Squares estimates***) $\hat{\theta}_1,\text{K},\hat{\theta}_p$ коэффициентов $\theta_1,\text{K},\theta_p$ вместе со ***скорректированными на гетероскедастичность*** оценками стандартных ошибок $s_{\hat{\theta}_j}$. Один из вариантов получения скорректированных на гетероскедастичность значений $s_{\hat{\theta}_j}$ был предложен Уайтом (*White)* и реализован в ряде пакетов статистического анализа данных, в том числе и в пакете EVIEWS. При этом удовлетворительные свойства оценки Уайта гарантируются только при большом количестве наблюдений. Мы не будем приводить здесь детали получения оценки Уайта, а просто воспользуемся пакетом EVIEWS для анализа данных из только что рассмотренного примера.

Пример. Используем данные из предыдущего примера, но применим для их анализа последнюю процедуру. Согласно этой процедуре, мы оцениваем коэффициенты α и β обычным методом наименьших квадратов, так что в качестве оценок берутся значения $\hat{\alpha} = 14.448$ и $\hat{\beta} = 0.105$. В качестве же оценок стандартных ошибок $s_{\hat{\alpha}}$ и $s_{\hat{\beta}}$ вместо значений $s_{\hat{\alpha}} = 9.562$ и $s_{\hat{\beta}} = 0.011$, полученных выше при оценивании модели обычным методом наименьших квадратов, берем значения оценок Уайта $s_{\hat{\alpha}} = 10.633$ и $s_{\hat{\beta}} = 0.018$.

Бросающееся в глаза значительное различие оценок для параметра α при применении двух рассмотренных методов

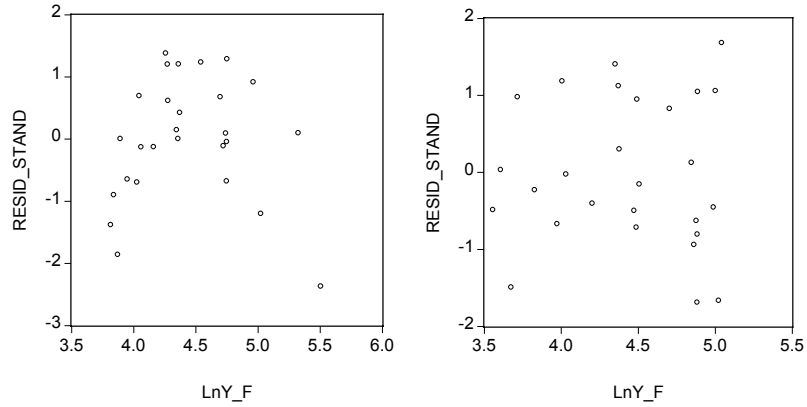
(3.803 и 14.448) в действительности не столь уж удивительно, поскольку оценки стандартной ошибки для $\hat{\alpha}$, полученные каждым из двух методов довольно высоки ($s_{\hat{\alpha}} = 4.570$ и $s_{\hat{\beta}} = 10.633$, соответственно).

Избавиться от неоднородности дисперсий ошибок в ряде случаев позволяет ***переход к логарифмам объясняемой переменной***.

Пример. По данным, использованным в двух предыдущих примерах, оценим модель наблюдений

$$\boxed{\ln y_i = \alpha + \beta x_i + \varepsilon_i,\ i = 1,\text{K},27.}$$

График зависимости стандартизованных остатков, полученных при оценивании этой модели, от предсказанных значений $\ln \hat{y}_i$ (левый график)



указывает на неправильную спецификацию модели, связанную с возможным пропуском квадратичной составляющей x_i^2. Оценивание расширенной модели наблюдений, включающей дополнительную объясняющую переменную x^2, приводит к остаткам, обнаруживающим существенно более удовле-

творительное поведение (см. правый график). Результаты оценивания расширенной модели приведены в следующей таблице.

Variable	Coefficient	Std. Error	t-Statistic	P-value
1	2.851	0.157	18.205	0.0000
x	0.003	0.000399	7.803	0.0000
x^2	-1.10E-06	2.24E-07	-4.925	0.0001

Таким образом, используя преобразования переменных, мы получили две альтернативные оцененные модели связи между переменными x и y:

$\boxed{y = 3.803 + 0.121x}$ и $\boxed{\ln y = 2.851 + 0.003\,x - 1.1 \cdot 10^{-6} x^2}$.

Первую из этих двух моделей можно предпочесть из соображений простоты интерпретации.

3.5. КОРРЕКЦИЯ СТАТИСТИЧЕСКИХ ВЫВОДОВ ПРИ АВТОКОРРЕЛИРОВАННОСТИ ОШИБОК

Пусть мы имеем дело с наблюдениями, производимыми последовательно через равные промежутки времени (ежедневные, еженедельные, ежеквартальные, ежегодные статистические данные) и выявляем по графику зависимости стандартизованных остатков $c_i = e_i / S$ от i тенденцию сохранения знака соседних наблюдений. В таком случае мы можем подозревать нарушение условия независимости случайных ошибок $\varepsilon_1, \text{K}, \varepsilon_n$ в принятой нами модели наблюдений

$$y_i = \theta_1 x_{i1} + \text{K} + \theta_p x_{ip} + \varepsilon_i, \quad i = 1, \text{K}, n,$$

в форме положительной автокоррелированности ряда ошибок.

Простейшей моделью автокоррелированности ошибок является ***модель авторегрессии первого порядка***:

$$\boxed{\varepsilon_i = \rho\, \varepsilon_{i-1} + \delta_i, \quad i = 2, \text{K}, n,}$$

где $|\rho|<1$, а δ_i, $i=2,\text{K},n,$ — независимые в совокупности случайные величины, имеющие одинаковое нормальное распределение $N\left(0,\sigma^2\right)$. Тогда гипотеза

$$\boxed{H_0:\ \rho=0}$$

соответствует (при нашем предположении о нормальности распределения случайных ошибок) ***независимости в совокупности случайных величин*** $\varepsilon_1,\text{K},\varepsilon_n$. В качестве альтернативной используем гипотезу

$$\boxed{H_A:\ \rho>0,}$$

соответствующую ***положительной автокоррелированности случайных величин*** $\varepsilon_1,\text{K},\varepsilon_n$ (т. е. тенденции преимущественного сохранения знака случайной ошибки при переходе от i - го наблюдения к $i+1$-му). Если гипотеза $H_0:\ \rho=0$ отклоняется критерием Дарбина-Уотсона в пользу альтернативной гипотезы $H_A:\ \rho>0$, то для получения правильных статистических выводов относительно коэффициентов модели необходима соответствующая коррекция.

Итерационная процедура Кохрейна-Оркатта (Cochrane-Orcutt).

Умножим обе части выражения для $(i-1)$-го наблюдения на ρ, так что

$$\rho y_{i-1}=\theta_1\rho x_{i-1,1}+\text{K}+\theta_p\rho x_{i-1,p}+\rho\varepsilon_{i-1},\quad i=1,\text{K},n,$$

и вычтем обе части полученного выражения из соответствующих частей выражения для i -го наблюдения:

$$y_i-\rho y_{i-1}=\theta_1\left(x_{i,1}-\rho x_{i-1,1}\right)+\text{K}+\theta_p\left(x_{i,p}-\rho x_{i-1,p}\right)+\left(\varepsilon_i-\rho\varepsilon_{i-1}\right).$$

Тем самым мы приходим к преобразованной модели наблюдений

$$\boxed{y_i'=\theta_1x_{i1}'+\text{K}+\theta_px_{ip}'+\varepsilon_i',\quad i=2,\text{K},n,}$$

где

$$y_i' = y_i - \rho y_{i-1},$$
$$x_{i,1}' = x_{i,1} - \rho x_{i-1,\,1},\ \text{K}\ ,\ x_{i,p}' = x_{i,p} - \rho x_{i-1,\,p},$$
$$\varepsilon_i' = \varepsilon_i - \rho\varepsilon_{i-1}.$$

Поскольку в принятой модели ошибок

$$\varepsilon_i - \rho\,\varepsilon_{i-1} = \delta_i\ ,\ \ i = 2,\text{K}\ ,n,$$

то это означает, что ошибки $\varepsilon_2',\text{K}\ ,\varepsilon_n'$ в преобразованной модели — независимые в совокупности случайные величины, имеющие одинаковое нормальное распределение $N\left(0,\sigma^2\right)$.

Иными словами, ***случайные ошибки в преобразованной модели удовлетворяют стандартным предположениям***. Следовательно, в рамках преобразованной модели никакой дополнительной коррекции обычных статистических выводов о коэффициентах модели не требуется. Проблема только в том, что используемое в процессе преобразования модели значение коэффициента ρ нам ***не известно***. Поэтому реально провести указанное преобразование невозможно. Вместо этого можно пытаться заменить указаное преобразование какой-либо его аппроксимацией с заменой неизвестного значения ρ на его оценку по данным наблюдений. Конечно, при использовании такой аппроксимации мы уже не можем гарантировать, что $\varepsilon_2',\text{K}\ ,\varepsilon_n'$ в преобразованной модели будут независимыми в совокупности случайными величинами, однако есть некоторая надежда на то, что эти ошибки все же будут обнаруживать меньшую автокоррелированность по сравнению с ошибками в исходной модели.

Описываемая здесь процедура Кохрейна-Оркатта использует для получения аппроксимации теоретического преобразования оценку для ρ в виде

$$r = \sum_{i=2}^{n} e_i e_{i-1} \Big/ \sum_{i=2}^{n} e_{i-1}^2 ,$$

где e_1, K, e_n — остатки, получаемые при оценивании исходной модели наблюдений. Аппроксимирующее преобразование определяется соотношениями

$$\begin{aligned} y_i^* &= y_i - r\, y_{i-1}, \\ x_{i,1}^* &= x_{i,1} - r\, x_{i-1,\,1},\ \mathrm{K}\ ,\ x_{i,p}^* = x_{i,p} - r\, x_{i-1,\,p}, \\ \varepsilon\ _{\mathrm{i}}^{\ *} &= \varepsilon_{\,i} - r\varepsilon_{\,i-1}, \end{aligned}$$

которые приводят к преобразованной модели

$$y_i^* = \theta_{\,1} x_{i,1}^* + \mathrm{K} + \theta_{\,p} x_{i,p}^* + \varepsilon\ _i^{\ *}, \quad i = 2, \mathrm{K}\ , n.$$

Если в последней модели автокоррелированность не проявляется, то полученные в рамках этой модели оценки параметров $\vec{\theta}_{\,1}, \mathrm{K}\ , \vec{\theta}_{\,p}$ можно принять в качестве уточненных оценок параметров $\theta_{\,1}, \mathrm{K}\ , \theta_{\,p}$. Если же в преобразованной модели еще остается выраженная автокоррелированность, то процесс преобразования применяют уже к преобразованной модели и еще раз уточняют значения параметров и т.д., пока последовательно уточняемые значения параметров не перестанут изменяться в пределах заданной точности.

Заметим, наконец, что обычно мы предполагаем, что $x_{i\,1} \equiv 1$. Соответственно, для первой объясняющей переменной получаем

$$x_{i,1}^* = x_{i,1} - r\, x_{i-1,1} = 1 - r\ ,$$

так что фактически мы имеем преобразованную модель

$$y_i^* = \alpha^{\,*} + \theta_{\,2} x_{i,2}^* \mathrm{K} + \theta_{\,p} x_{i,p}^* + \varepsilon\ _i^{\ *}, \quad i = 2, \mathrm{K}\ , n,$$

с $\alpha^* = \theta_1(1-r)$. Получив в этой модели оценку $\hat{\alpha}^*$ для α^*, мы можем оценить параметр θ_1 исходной модели, просто полагая

$$\boxed{\hat{\theta}_1 = \hat{\alpha}^* / (1-r).}$$

Пример. Проанализируем статистические данные о совокупных потребительских расходах (CONS) и денежной массе (MONEY) в США за 1952—1956 г. г. (квартальные данные, в млрд. долларов).

obs	MONEY	CONS	obs	MONEY	CONS
1952:1	159.3	214.6	1954:3	173.9	238.7
1952:2	161.2	217.7	1954:4	176.1	243.2
1952:3	162.8	219.6	1955:1	178.0	249.4
1952:4	164.6	227.2	1955:2	179.1	254.3
1953:1	165.9	230.9	1955:3	180.2	260.9
1953:2	167.9	233.3	1955:4	181.2	263.3
1953:3	168.3	234.1	1956:1	181.6	265.6
1953:4	169.7	232.3	1956:2	182.5	268.2
1954:1	170.5	233.7	1956:3	183.3	270.4
1954:2	171.6	236.5	1956:4	184.3	275.6

Результаты оценивания линейной модели наблюдений

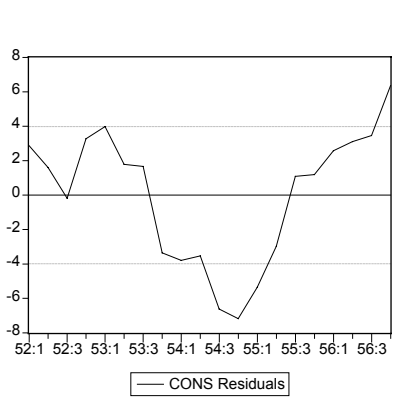
$y_i = \alpha + \beta\, x_i + \varepsilon_i$, $i = 1,\text{K},20,$

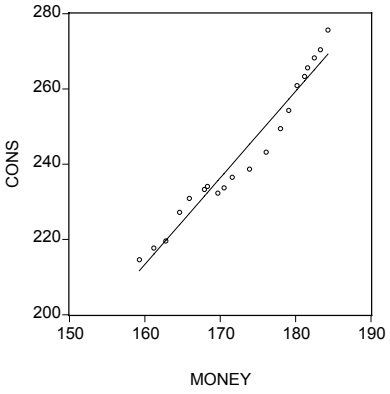
в которой y_i — значения объясняемой переменной CONS, а x_i - значения объясняющей переменной MONEY, приведены в следующей таблице:

Variable	Coefficient	Std. Error	t-Statistic	Prob.
1	–154.719	19.850	-7.794	0.0000
X	2.300	0.114	20.080	0.0000
R-squared	0.957	Durbin-Watson stat		0.328

Хотя коэффициент детерминации весьма близок к единице, значение статистики Дарбина-Уотсона достаточно мало, и это дает возможность подозревать наличие положительной ав-

токоррелированности ошибок в принятой модели наблюдений. Два следующих графика дают представление о рассеянии значений переменных и о поведении остатков.





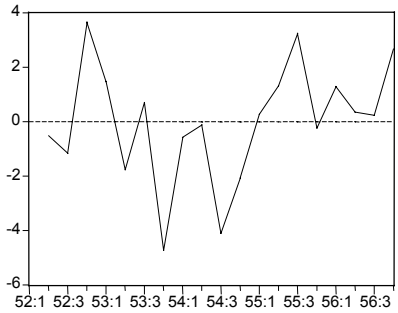
Здесь наблюдаются серии остатков, имеющих одинаковые знаки, что как раз и характерно для моделей, в которых имеется положительная автокоррелированность ошибок.

Для подтверждения положительной автокоррелированности ошибок используем критерий Дарбина-Уотсона. По таблицам находим нижнюю границу для критического значения $d_{0.05}$ при $n = 20$: $d_{L,0.05} = 1.20$. Полученное при оценивании модели значение $DW = 0.328$ существенно меньше этой нижней границы, так что гипотеза H_0: $\rho = 0$ отвергается в пользу альтернативной гипотезы H_A: $\rho > 0$. Для коррекции статистических выводов используем процедуру Кохрейна-Оркатта.

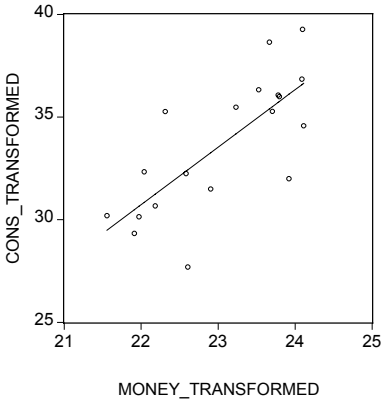
Прежде всего находим оценку для неизвестного значения коэффициента ρ: $r = \sum_{i=2}^{n} e_i e_{i-1} \Big/ \sum_{i=2}^{n} e_{i-1}^2 = 0.874$. Основываясь на этой оценке, переходим к преобразованной модели, оценивание которой дает следующие результаты:

Included observations: 19 after adjusting endpoints				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
1	-30.777	14.043	-2.192	0.0426
X’	2.795	0.609	4.593	0.0003
R-squared	0.554	Durbin-Watson stat		1.667

Хотя в преобразованной модели коэффициент детерминации существенно ниже, чем в непреобразованной модели, значение статистики Дарбина-Уотсона теперь превышает верхнюю границу для критического значения $d_{0.05}$, соответствующего $n = 19$. (В преобразованной модели наблюдений на единицу меньше, чем в исходной, так как при преобразовании используются запаздывающие значения обеих переменных). Поэтому гипотеза о независимости в совокупности ошибок в преобразованной модели не отвергается (в пользу гипотезы об их положительной автокоррелированности). Два следующих графика дают представление о рассеянии значений преобразованных переменных и о поведении остатков в преобразованной модели.



RESID: TRANSFORMED MODEL



Обратим внимание на существенно более нерегулярное поведение остатков по сравнению с исходной моделью.

Обращаясь к результатам оценивания коэффициентов в преобразованной модели, отметим значительное (более, чем в 5 раз!) возрастание оценки стандартной ошибки $s_{\bar{\beta}}$, что подтверждает сделанное ранее замечание о занижении стандартных ошибок при неучете имеющейся в действительности положительной автокорреляции случайных ошибок в модели наблюдений. Столь существенное возрастание значения $s_{\bar{\beta}}$ приводит к возрастанию более, чем в 5 раз, и ширины доверительного интервала для мультипликатора β. Если при оценивании исходной линейной модели 95%-доверительный интервал для этого параметра имел вид $2.058 < \beta < 2.542$, то при оценивании преобразованной модели мы получаем интервал $1.516 < \beta < 4.074$.

Рассмотренный пример ясно демонстрирует опасность пренебрежения возможной неадекватностью построенной модели в отношении стандартных предположений об ошибках и необходимость обязательного проведения в процессе подбора подходящей модели связи между теми или иными экономическими факторами анализа остатков, полученных при оценивании выбранной модели.

Более того, используя преобразованную модель, можно получить улучшенную модель для прогнозирования объемов расходов на потребление при планируемых объемах денежной массы. Поясним это на примере простой линейной модели

$y_i = \alpha + \beta\, x_i + \varepsilon_i$, $i = 1,\text{K}\,,n.$

Предполагая, что $\varepsilon_i - \rho\,\varepsilon_{i-1} = \delta_i$, $i = 2,\text{K}\,,n,$ и используя оценку r для коэффициента ρ, переходим к преобразованной модели

$y_i^* = \alpha^* + \beta\, x_i^* + \varepsilon_i^*,\ i = 2,\text{K},n,$

с $y_i^* = y_i - r\, y_{i-1},\quad x_i^* = \left(x_i - r\, x_{i-1}\right),\ i = 2,\text{K},n,$ и $\alpha^* = \alpha\left(1-r\right),$

и получаем в рамках этой модели оценки $\hat{\alpha}^*$ и $\hat{\beta}$ параметров α^* и β, так что оцененная модель линейной связи между преобразованными переменными имеет вид

$\hat{y}_i^* = \hat{\alpha}^* + \hat{\beta}\, x_i^*,\ i = 2,\text{K},n.$

В исходных переменных последние соотношения принимают вид

$\hat{y}_i - r\, y_{i-1} = \hat{\alpha}\left(1-r\right) + \hat{\beta}\left(x_i - r\, x_{i-1}\right),\ i = 2,\text{K},n,$

где $\hat{\alpha} = \hat{\alpha}^*/\left(1-r\right)$, откуда получаем:

$$\boxed{\hat{y}_i = \hat{\alpha} + \hat{\beta}\, x_i + r\left(y_{i-1} - \hat{\alpha} - \hat{\beta}\, x_{i-1}\right),\ i = 2,\text{K},n.}$$

Если мы собираемся теперь прогнозировать будущее значение y_{n+1}, соответствующее плановому значению x_{n+1} объясняющей переменной, то естественно воспользоваться полученным соотношением и предложить в качестве прогнозного для y_{n+1} значение

$$\boxed{\hat{y}_{n+1} = \hat{\alpha} + \hat{\beta}\, x_{n+1} + r\left(y_n - \hat{\alpha} - \hat{\beta}\, x_n\right).}$$

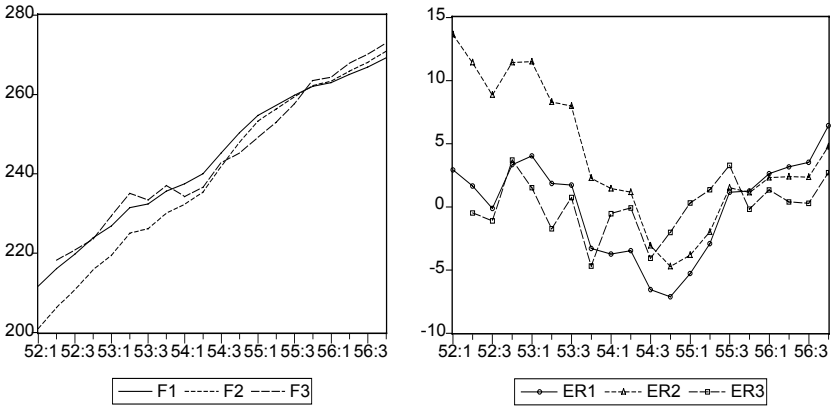
При таком способе вычисления прогнозного значения для y_{n+1} учитывается тенденция сохранения знака остатков: если в последнем наблюдении наблюдавшееся значение y_n <u>превышало</u> значение $\hat{\alpha} + \hat{\beta}\, x_n$, предсказываемое линейной моделью связи $y = \hat{\alpha} + \hat{\beta}\, x$, то и последующее значение y_{n+1} <u>прогнозируется с превышением</u> значения $\hat{\alpha} + \hat{\beta}\, x_{i+1}$, предсказываемого этой линейной моделью связи при $r > 0$. Если же

значение y_n меньше, чем $\hat{\alpha} + \hat{\beta}\, x_n$, то тогда будущее значение y_{n+1} прогнозируется меньшим значения $\hat{\alpha} + \hat{\beta}\, x_{i+1}$.

Пример. Продолжим рассмотрение предыдущего примера. В этом примере $r = 0.874,\ \hat{\alpha} = \hat{\alpha}^* / (1 - r) = -30.777 / (1 - 0.874) = -244.262$, $\hat{\beta} = 2.795$. Наблюдавшимся значениям x_2, K, x_{20} можно сопоставить:

- наблюдавшиеся значения y_2, K, y_{20};
- значения
  $\hat{y}_i = -154.700 + 2.300 x_i$,
  получаемые по модели, построенной без учета автокоррелированности ошибок;
- значения
  $\hat{y}_i = -244.262 + 2.795 x_i$,
  получаемые по модели, параметры которой скорректированы с учетом автокоррелированности ошибок;
- значения
  $\hat{y}_i = -244.262 + 2.795\, x_i + 0.874 \left(y_{i-1} + 244.262 - 2.795\, x_{i-1} \right)$,
  отличающиеся от значений, указанных в предыдущем пункте, учетом значения остатка в предшествующем наблюдении.

Ниже приведены графики значений $\hat{y}_i$, получаемых указанными тремя методами, и графики соответствующих им расхождений $\hat{y}_i - y_i$. Индексы 1, 2, 3 указывают на один из трех способов получения значений $\hat{y}_i$, в том порядке, в котором они были перечислены выше).



Сравним средние квадраты расхождений $(1/19)\sum_{i=2}^{20}\left(\hat{y}_i - y_i\right)^2$ при использовании указанных трех методов вычисления значений $\hat{y}_i$. Эти средние квадраты равны, соответственно,

$MSE_1 = 14.583,\ MSE_2 = 37.025,\ MSE_3 = 4.533,$

что говорит о большей гибкости прогноза, построенного по последнему (третьему) методу.

Рассмотрим еще одно важное следствие автокоррелированности ошибок в линейной модели

$y_i = \alpha + \beta\ x_i + \varepsilon_i\ ,\ i = 1,\text{K}\ ,n,$

с $\varepsilon_i - \rho\ \varepsilon_{i-1} = \delta_i\ ,\ \ i = 2,\text{K}\ ,n.$ Преобразование

$y_i' = y_i - \rho\, y_{i-1},\ x_i' = x_i - \rho\, x_{i-1}$

приводит к модели наблюдений

$y_i' = \alpha' + \beta\ x_i' + \delta_i\ , i = 1,\text{K}\ ,n,$

на основании которой получаем соотношение

$y_i = \alpha\left(1-\rho\right) + \rho\, y_{i-1} + \beta\ \left(x_i - \rho\, x_{i-1}\right)\ + \delta_i\ ,\quad i = 2,\text{K}\ ,n.$

Вспомним теперь о нашем предположении, что $0 < \rho < 1$, и преобразуем последнее соотношение следующим образом:

$$y_i = \alpha\,(1-\rho) + y_{i-1} - (1-\rho)y_{i-1} + \beta\,(x_i - x_{i-1} + (1-\rho)x_{i-1}) + \delta_i$$
$$= y_{i-1} + (1-\rho)(\alpha + \beta\,x_{i-1} - y_{i-1}) + \beta\,(x_i - x_{i-1}) + \delta_i\,,$$

или

$$\boxed{\Delta y_i = \beta\,\Delta x_i + (\rho-1)(y_{i-1} - \alpha - \beta\,x_{i-1}) + \delta_i\,.}$$

Здесь $\Delta y_i = y_i - y_{i-1}$, $\Delta x_i = x_i - x_{i-1}$ и $-1 < (\rho-1) < 0$. Второе слагаемое в правой части по-существу поддерживает «долговременную» линейную связь (тенденцию)

$$\boxed{y = \alpha + \beta\,x.}$$

Если в момент $i-1$ отклонение y_{i-1} от $(\alpha + \beta\,x_{i-1})$ положительно $(y_{i-1} > \alpha + \beta\,x_{i-1})$, то второе слагаемое будет отрицательным, действуя в сторону уменьшения приращения $\Delta y_i = y_i - y_{i-1}$. Если же отклонение y_{i-1} от $(\alpha + \beta\,x_{i-1})$ отрицательно $(y_{i-1} < \alpha + \beta\,x_{i-1})$, то второе слагаемое будет положительным, действуя в сторону увеличения приращения $\Delta y_i = y_i - y_{i-1}$.

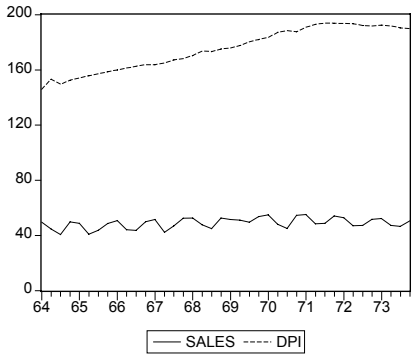
Указанная модель коррекции приращений переменной y использует «истинные» значения параметров α, β, ρ. Поскольку эти значения нам не известны, мы в состоянии построить только аппроксимацию такой модели, использующую оценки параметров. При этом естественно воспользоваться оценкой r и уточненными оценками $\hat{\alpha}, \hat{\beta}$, полученными на базе преобразованной модели.

В рассмотренном примере аппроксимирующая модель коррекции приращений принимает вид

$$\boxed{\Delta y_i = 2.795\,\Delta x_i - 0.126\,(y_{i-1} + 244.262 - 2.795\,x_{i-1}).}$$

3.6. КОРРЕКЦИЯ СТАТИСТИЧЕСКИХ ВЫВОДОВ ПРИ НАЛИЧИИ СЕЗОННОСТИ. ФИКТИВНЫЕ ПЕРЕМЕННЫЕ

Приведенный ниже график показывает динамику изменения совокупного располагаемого дохода ***DPI*** и объемов продаж ***SALES*** лыжного инвентаря в США (квартальные данные; *DPI* — в *млрд* долларов, *SALES* — в *млн* долларов, в ценах 1972 г.).



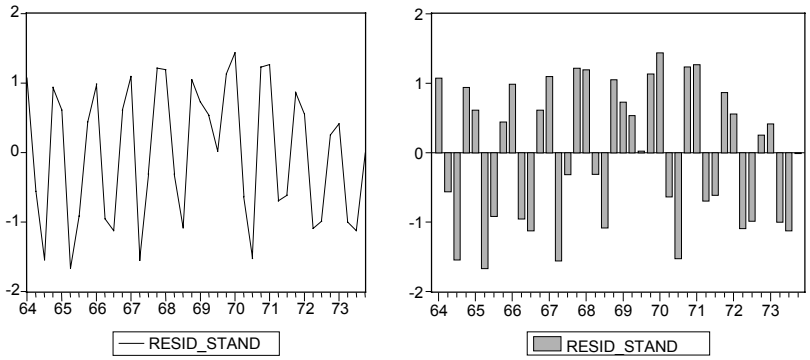
Оценивание линейной модели связи указанных переменных дает следующие результаты.

Dependent Variable: SALES
Method: Least Squares
Sample: 1964:1 1973:4
Included observations: 40

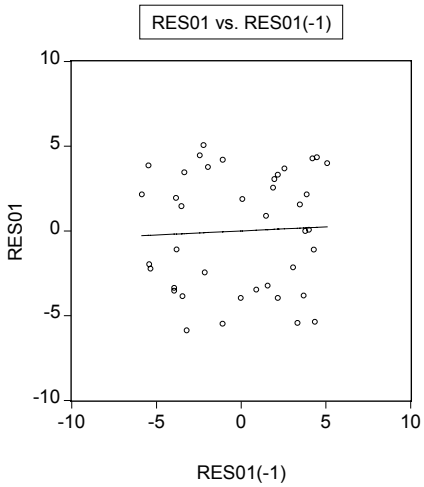
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	29.97613	6.463626	4.637665	0.0000
DPI	0.108402	0.036799	2.945768	0.0055

R-squared	0.185904	Mean dependent var	48.94571
Adjusted R-squared	0.164481	S. D. dependent var	3.852032
S. E. of regression	3.521017	Akaike info criterion	5.404084
Sum squared resid	471.1074	Schwarz criterion	5.488528
Log likelihood	–106.0817	F-statistic	8.677546
Durbin-Watson stat	1.874403	Prob (F-statistic)	0.005475

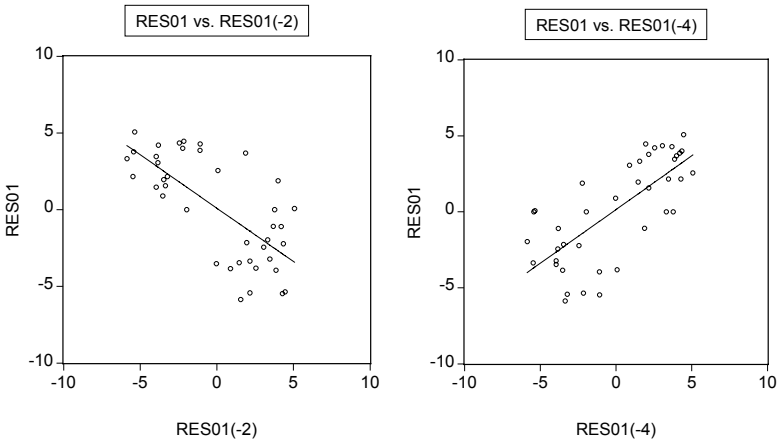
Коэффициент при переменной *DPI* статистически значим. Однако график стандартизованных остатков (приведенный для удобства в двух формах)



обнаруживает явную неадекватность построенной модели имеющимся наблюдениям. Однако характер этой неадекватности таков, что он не улавливается критерием Дарбина-Уотсона: значение 1.874 статистики Дарбина-Уотсона близко к 2. И это не удивительно: за положительными остатками с равным успехом следуют как положительные, так и отрицательные остатки, что соответствует практическому отсутствию корреляции между соседними ошибками и подтверждается диаграммой рассеяния



(Здесь *RES*01 — переменная, образованная остатками от подобранной модели линейной связи, а *RES*01(−1) — переменная, образованная запаздывающими на один квартал значениями переменной *RES*01.)



В то же время, налицо отрицательная коррелированность остатков для наблюдений, отстоящих на два квартала, и положительная — для наблюдений, отстоящих на четыре квартала:

В отличие от критерия Дарбина-Уотсона, критерий Бройша-Годфри «замечает» такую коррелированность: допуская коррелированность ошибок для наблюдений, разделенных двумя кварталами, получаем $P-value = 0.000037$, что ведет к безусловному отклонению гипотезы о независимости ошибок.

Обратим теперь внимание на весьма специфическое поведение остатков. Все остатки, соответствуюшие первому и четвертому кварталам, положительны, а все (за исключением двух) остатки, соответствующие второму и третьему кварталам, отрицательны. Такое положение, конечно, просто отражает тот факт, что спрос на зимний спортивный инвентарь возрастает в осенне-зимний период и снижается в весенне-летний период года, т. е. имеет ***сезонный характер***.

Построенная нами модель не учитывает фактор сезонности спроса и потому оказывается неадекватной. Вследствие этого, такая модель не может, в частности, использоваться для прогнозирования объема спроса в зависимости от величины совокупного располагаемого дохода.

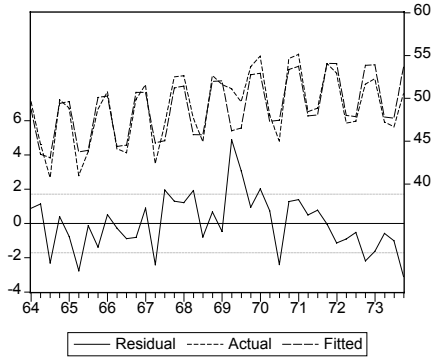
Для коррекции моделей связи в подобных ситуациях часто привлекают искусственно построенные переменные — «***фиктивные переменные*** *(«dummy» variables)*. В нашем случае в качестве такой дополнительной переменной можно взять, например, переменную $DUMMY$, значение которой равно 1 для первого и четвертого кварталов и равно 0 для второго и третьего кварталов. Добавление такой переменной в качестве объясняющей позволяет учесть сезонные колебания спроса. Оценивание расширенной модели дает следующие результаты.

Dependent Variable: SALES

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	26.21787	3.152042	8.317742	0.0000
DPI	0.112653	0.017847	6.312227	0.0000
DUMMY	6.028524	0.539997	11.16399	0.0000

R-squared	0.813644	Mean dependent var	48.94571
Adjusted R-squared	0.803571	S. D. dependent var	3.852032
S. E. of regression	1.707233	Akaike info criterion	3.979663
Sum squared resid	107.8419	Schwarz criterion	4.106329
Log likelihood	-76.59327	F-statistic	80.77244
Durbin-Watson stat	1.452616	Prob (F-statistic)	0.000000

Оцененное значение 6.029 коэффициента при переменной *DUMMY* фактически означает, что спрос на лыжный инвентарь в течение первого и четвертого кварталов возрастает по сравнению со спросом в течение второго и четвертого кварталов в среднем примерно на 6 *млн* долларов (в ценах 1972 г.). Следующий график иллюстрирует качество подобранной расширенной модели.



На сей раз значение $P-value$ для статистики критерия Бройша-Годфри равно 0.157197 против прежнего значения 0.000037, так что этот критерий теперь не отвергает гипотезу независимости случайных ошибок $\varepsilon_1, \text{K}, \varepsilon_n$.

По-существу, мы подобрали две различные модели линейной связи между DPI и $SALES$:

модель

$SALES = 26.21787 + 0.112653 DPI$

для весенне-летнего периода;

модель

$SALES = (26.21787 + 6.028524) + 0.112653 DPI$

для осенне-зимнего периода.

При этом, предельная склонность к закупке лыжного инвентаря в обеих моделях остается одинаковой и оценивается величиной 0.112653.

Замечание. Вместо подбора отдельных моделей для осенне-зимнего и весенне-летнего периодов можно было бы заняться подбором отдельных моделей для каждого из четырех кварталов года. С этой целью в качестве дополнительных объясняющих переменных можно взять, например, переменные $DUMMY4$, $DUMMY1$, $DUMMY2$, принимающие значение 1, соответственно, в четвертом, первом и втором кварталах, и равные нулю в остальных кварталах. При оценивании такой расширенной модели для наших данных оказывается незначимым коэффициент при $DUMMY2$, что означает близость в среднем уровней продаж во втором и в третьем кварталах. Более того, оказываются близкими оценки коэффициентов при переменных $DUMMY4$ и $DUMMY1$. Гипотеза о совпадении двух последних коэффициентов не отвергается, и в итоге мы возвращаемся к модели с одной фиктивной переменной $DUMMY$, которую мы уже оценили ранее.

Использование фиктивных переменных полезно при анализе ***агрегированных (объединенных) данных***, полученных при объединении наблюдений, относящихся к различным полам (мужчины и женщины), к различным возрастным, языковым и социальным группам, к различным периодам времени. В таких ситуациях модели, построенные по отдельным группам, могут существенно различаться, и тогда модель, построенная по объединенным данным, не учитывает этого различия. Привлечение фиктивных переменных позволяет оценить значимость такого различия и по результатам этой оценки остановиться на модели с агрегированными данными или на модели, в которой учитывается различие параметров связи для различных групп (периодов времени).

В качестве примера, попробуем построить модель связи между переменными Z и X, которые в 15 наблюдениях имели следующие значения:

X	Z	X	Z	X	Z
1	1.257	6	0.865	11	1.804
2	1.812	7	1.930	12	1.956
3	3.641	8	2.944	13	3.134
4	4.401	9	4.316	14	4.649
5	5.561	10	5.323	15	4.559

Этим данным соответствует приведенная ниже диаграмма рассеяния;

Прямая на диаграмме соответствует подобранной модели связи

$Z = 2.414 + 0.099X$;

t - статистика для коэффициента при X принимает значение 1.087, что дает $P - value = 0.297$ и ведет к неотвержению гипотезы о равенстве этого коэффициента нулю. Регрессия переменной Z на переменную X признается незначимой.

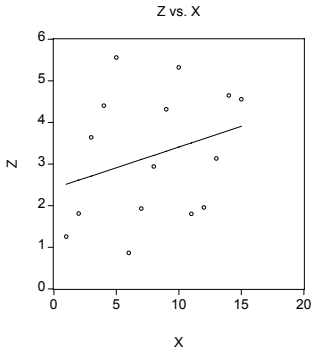
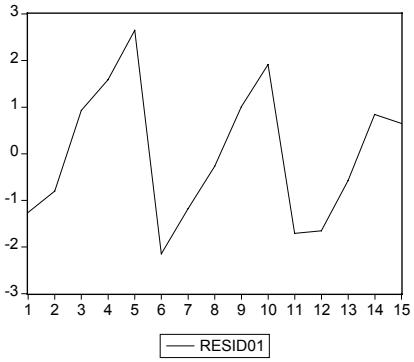


График указывает на наличие трех режимов линейной связи между переменными Z и X, соответствующим 5 первым, 5 центральным и 5 последним наблюдениям. Коэффициент при X кажется одинаковым для всех трех режимов, тогда как постоянные различаются.

В то же время, график остатков от подобранной модели связи явно указывает на неправильную спецификацию модели:



Чтобы учесть обнаруженное по графику остатков наличие трех режимов, привлечем в качестве дополнительных объясняющих переменных две фиктивные переменные: переменную $D2$, равную 1 в пяти центральных наблюдениях и равную 0 в остальных наблюдениях, а также переменную $D3$, равную 1 в

пяти последних наблюдениях и равную 0 в остальных наблюдениях. Оценивание расширенной модели с участием этих дополнительных объясняющих переменных дает следующий результат:

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.264368	0.274073	0.964591	0.3555
X	1.023398	0.070765	14.46185	0.0000
D2	-5.375960	0.430449	-12.48920	0.0000
D3	-10.34806	0.748910	-13.81749	0.0000
R-squared	0.950286	Mean dependent var		3.210213
Durbin-Watson stat	2.205754	Prob (F-statistic)		0.000000

На этот раз регрессия оказывается не только статистически значимой, но и имеет очень высокую значимость; то же относится и к коэффициентам при переменных X, $D2$ и $D3$. Высокая значимость двух последних коэффициентов подтверждает значимое отличие констант в моделях линейной связи между переменными Z и X.

В заключение обратимся опять к примеру, рассмотренному в параграфе 3.3. Мы обнаружили там, что модель линейной связи

$$CONS_t = \theta_1 + \theta_2 DPI_t + \theta_3 ASSETS_{t-1} + \varepsilon_t, \quad t = 2,\text{K},27,$$

оказалась неудовлетворительной, поскольку анализ остатков от оцененной модели выявил гетероскедастичность и автокоррелированность ошибок и отличие распределения ошибок от нормального. Приведенные там график зависимости стандартизованных остатков $c_i = e_i / S$ от номера наблюдений и его вариант в виде зависимости от года наблюдения указывают на явную разницу в поведении остатков в первой части перио-

да наблюдений (до 1972 года) и во второй его части (1973-1985 годы). Такое различие в поведении остатков свидетельствует о том, что в 1973 году произошел структурный сдвиг в экономической ситуации, связанный с мировым топливо-энергетическим кризисом, который изменил характер связи между рассматриваемыми макроэкономическими факторами. Последнее могло, например, выразиться в изменении значений параметров $\vartheta_1, \vartheta_2, \vartheta_3$ при переходе ко второй части периода наблюдений. Возможность такого изменения учитывает расширенная модель

$$CONS_t = \gamma_1 (D1)_t + \gamma_2 (D2)_t + \gamma_3 (DPI1)_t + \gamma_4 (DPI2)_t + \gamma_5 (ASSLAG1)_t + \gamma_6 (ASSLAG2)_t + \varepsilon_t, \quad t = 2,\text{K},27.$$

Здесь

$(D1)_t$ - фиктивная переменная, равная 1 для $t = 1,\text{K},14$ (что соответствует периоду с 1959 по 1972 год) и равная 0 для $t = 15,\text{K},27$ (что соответствует периоду с 1973 по 1985 год),

$(D2)_t = 1 - (D1)_t$ - фиктивная переменная, равная 0 для $t = 1,\text{K},14$ и равная 1 для $t = 15,\text{K},27$,

$(DPI1)_t = DPI_t \cdot (D1)_t$ - переменная, равная $(DPI)_t$ для $t = 1,\text{K},14$ и равная 0 для $t = 15,\text{K},27$,

$(DPI2)_t = DPI_t \cdot (D2)_t$ - переменная, равная 0 для $t = 1,\text{K},14$ и равная $(DPI)_t$ для $t = 15,\text{K},27$,

$(ASSLAG1)_t = ASSETS_{t-1} \cdot (D1)_t$ - переменная, равная $ASSETS_{t-1}$ для $t = 2,\text{K},14$ и равная 0 для $t = 15,\text{K},27$,

$(ASSLAG2)_t = ASSETS_{t-1} \cdot (D2)_t$ - переменная, равная 0 для $t = 2,\text{K},14$ и равная $ASSETS_{t-1}$ для $t = 15,\text{K},27$.

Заметим, что при этом

$(DPI1)_t + (DPI2)_t = DPI_t, \; t = 1,\text{K},27,$

$(ASSLAG1)_t + (ASSLAG2)_t = ASSETS_{t-1}, \; t = 2,\text{K},27.$

В рамках расширенной модели проверим гипотезу

$$H_0: \gamma_1 = \gamma_2,\ \gamma_3 = \gamma_4,\ \gamma_5 = \gamma_6,$$

используя F-критерий. Значению F-статистики 10.490 соответствует P-значение 0.0002, так что гипотеза H_0 отвергается, и это говорит об изменении хотя бы одного из параметров $\vartheta_1, \vartheta_2, \vartheta_3$ при переходе ко второй части периода наблюдений. Поскольку оценки параметров γ_5 и γ_6 статистически незначимы (им соответствуют P-значения 0.1157 и 0.5599), проверим гипотезу о равенстве нулю обоих этих параметров. Получаемое P-значение 0.2412 означает, что последняя гипотеза не отвергается, так что допуская изменение параметров модели при переходе ко второй части периода наблюдений, можно вообще отказаться от включения в модель переменной *ASSETS* и ограничиться моделью

$$CONS_t = \gamma_1 (D1)_t + \gamma_2 (D2)_t + \gamma_3 (DPI1)_t + \gamma_4 (DPI2)_t + \varepsilon_t,$$
$$t = 1, \text{K}, 27.$$

Оценивание этой модели дает следующие результаты: $R^2 = 0.9992$,

$\gamma_1 = 57.834$, P - value = 0.0059;

$\gamma_2 = -234.836$, P - value = 0.0000;

$\gamma_3 = 0.865$, P - value = 0.0000;

$\gamma_4 = 1.012$, P - value = 0.0000;

Гипотеза $H_0: \gamma_3 = \gamma_4$ здесь отвергается (P - value = 0.0000), как и гипотеза $H_0: \gamma_1 = \gamma_2$, так что структурный сдвиг затрагивает и постоянную и коэффициент при *DPI*.

Значение статистики Дарбина-Уотсона равно $DW = 2.06$ и не выявляет автокоррелированности ошибок. К тому же ре-

зультату приводит и применение критерия Бройша-Годфри с $K = 1,\ K = 2,\ K = 3$. Критерий Уайта дает P - value $= 0.433$, не выявляя гетероскедастичности, а критерий Жарка-Бера дает P - value $= 0.445$, не выявляя существенных отклонений распределения ошибок от нормального.

Вспомним, однако, про критерий Голдфелда-Квандта. Опять выделяя периоды с 1960 по 1969 год и с 1976 по 1985 год, получаем значение F -статистики 3.354, соответствующее P - value $= 0.0832$, так что на сей раз и этот критерий не обнаруживает существенной гетероскедастичности.

Тем самым, мы имеем основания принять в качестве возможной модели наблюдений, объясняющей изменения объема совокупного потребления на периоде с 1959 по 1985 год, оцененную модель

$$CONS_t = 57.834(D1)_t - 234.836(D2)_t + 0.865(DPI1)_t + 1.012(DPI2)_t + \varepsilon_t,\ \ t = 1,\text{K},27.$$

Эту модель можно также записать в виде

$$CONS_t = \begin{cases} 57.834 + 0.865\ DPI_t + \varepsilon_t, & t = 1,\text{K},14, \\ -234.836 + 1.012\ DPI_t + \varepsilon_t, & t = 15,\text{K},27. \end{cases}$$

Соответственно последней форме записи такая модель называется ***двухфазной линейной регрессией (***или ***линейной моделью с переключением)***. Заметим, наконец, что допустив возможность изменения постоянной и коэффициента при DPI при переходе ко второй части периода наблюдений, мы можем допустить при этом и изменение дисперсии ошибок, т.е. полагать, что $D(\varepsilon_t) = \sigma_1^2$ для $t = 1,\text{K},14$ и $D(\varepsilon_t) = \sigma_2^2$ для $t = 15,\text{K},27$. Оценки для σ_1 и σ_2 в этом случае равны, соответственно, 8.517 и 14.886.

ЗАКЛЮЧЕНИЕ

В рамках короткого вводного курса мы успели рассмотреть только основы построения и статистического анализа моделей связи между экономическими факторами. Базовым являлось предположение о том, что объясняющие переменные являются неслучайными величинами, на которые накладываются случайные ошибки, имеющие нормальное распределение.

Отказ от предположения нормальности распределения ошибок в модели наблюдений во многих ситуациях компенсируется возможностью использовать изложенные методы при "больших выборках", т.е. при большом количестве наблюдений. Отказ от предположения о неслучайном характере объясняющих переменных чреват более серьезными последствиями и требует применения более тонких и сложных методов статистического анализа, изучение которых, в свою очередь, требует существенных знаний в области теории вероятностей и математической статистики. Особенно это относится к исследованию связей между переменными, эволюционирующими во времени (временными рядами).

Как уже отмечалось в Предисловии, заинтересованный читатель может обратиться далее к цитировавшейся там книге К.Доугерти, где в доступной форме изложены некоторые вопросы, связанные с неслучайностью объясняющих переменных, моделированием динамических процессов и оцениванием систем одновременных уравнений. Полезно также обратиться к книге Я.Р.Магнуса, П.К.Катышева и А.А.Пересецкого (1997), в которой те же вопросы изложены в более компактном, но и более формальном виде. Затем можно ознакомиться с основами статистического анализа временных рядов, обратившись к

книге С.А.Айвазяна и В.С.Мхитаряна (1998). Разнообразные эконометрические модели и методы анализа этих моделей обсуждаются в книге W. H. Green (1993). Подробный обзор современных методов статистического анализа связей между временными рядами, имеющими выраженный тренд, имеется в книге Maddala G.,S., Kim In-Moo (1999), однако чтение этой книги требует существенной математической подготовки. В приводимом ниже списке литературы перечислены и некоторые другие руководства различной степени сложности, изданные в последнее десятилетие.

СПИСОК ЛИТЕРАТУРЫ

Айвазян С.А., Мхитарян В.С. (1998), *Прикладная статистика и основы эконометрики.* М., ЮНИТИ.-1022 с.

Магнус Я.Р., Катышев П.К., Пересецкий А.А. (1997), *Эконометрика. Начальный курс*. 3-е изд. М., Дело.-400 с.

Доугерти Кристофер (1997), *Введение в эконометрику.* Пер. с англ.- М., ИНФРА-М.- XIV, 402 с.

Maddala G.L., Kim In-Moo (1999), *Unit Roots, Cointegration, and Structural Change*. Cambridge Univ. Press.

Davidson R., MacKinnon J.G. (1993), *Estimation and Inference in Econometrics*. Oxford Univ. Press.

Hatanaka M. (1996), *Time-Series Based Econometrics. Unit Root and Cointegration*. Oxford Univ. Press.

Green W.H. (1993), *Econometric Analysis* (second edition). Macmillan Publishing Company.

Johnston, J., DiNardo J. (1997), *Econometric Methods*. McGraw-Hill, Inc.