Федеральное агентство железнодорожного транспорта
Федеральное государственное бюджетное
образовательное учреждение высшего
образования
«Московский государственный университет
путей сообщения императора Николая II»

Кафедра “Математика”

А.С.МИЛЕВСКИЙ

ЭКОНОМЕТРИКА

ДЛЯ МАГИСТРОВ

Курс лекций

МОСКВА – 2016

1. Введение

1.1. Предмет эконометрики

Эконометрика – это наука, изучающая конкретные **количественные** и **качественные взаимосвязи** экономических объектов и процессов при помощи **математических** методов и моделей.

Норвежский ученый Р. Фриш определил эконометрику как «…*единство экономической теории, статистики и математики*».

Основные результаты *экономической теории* носят *качественный* характер, например, даются ответы на вопросы вида:

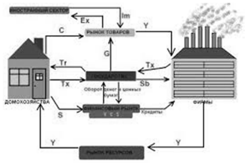
- Какие факторы влияют на совокупный объём инвестиций в частном бизнесе?
- Как зависят темпы экономического роста от размеров дефицита государственного бюджета, объёма инвестиций, внешнего долга, уровня инфляции?
- Каким образом изменит структуру потребления снижение ставок налогов?

Однако для принятия обоснованных решений часто требуется отвечать на вопросы, начинающиеся словами «**на сколько**». Например,

- Какого изменения объёма потребления электроэнергии можно ожидать, если предположить, что цены на неё возрастут на 30%?
- Насколько увеличится спрос на данную продукцию, если удвоить текущие расходы на её рекламу?

Это ***количественное содержание*** и является предметом изучения эконометрики.

Следует отметить, что зачастую не удаётся достичь желаемой степени точности предсказаний. Модели и уравнения, дающие хорошее согласование для одного периода времени, могут оказаться неудовлетворительными для другого. Это связано с большой сложностью экономических процессов.



Но в любом случае анализ в эконометрике производится на основе широкого применения самых разнообразных *математических методов и моделей*.

И, конечно, значимые результаты можно получить лишь путем обработки *реальных статистических данных*.

1.2. Модели

Закономерности в экономике проявляются как взаимосвязи между экономическими показателями. Как правило, их стремятся выразить в сравнительно простой математической форме.

Надо иметь в виду, что математическая модель является лишь упрощённым, концентрированным представлением реальности.

Искусство построения модели состоит в совмещении “желаемого с возможным”.

Пример.

Изучая стоимость Y квадратного метра квартиры в Москве, естественно предполагать, что она зависит от

- количества комнат N,
- жилой площади квартиры P,
- нежилой площади Q,
- расстояния от центра города R,
- расстояния от ближайшей станции метро S,

- вида дома (кирпичный, панельный и т.п.) T,
- этажа U,..

$$Y = f(N, P, Q, R, S, T, U, \ldots)$$

В любом случае влияние неучтённых (или неизвестных) факторов будет представлено как некая случайная добавка

ε к Y.

Наличие ***случайной составляющей*** ε в модели может быть обусловлено причинами двоякой природы.

- Во-первых, она отражает влияние на формирование значения Y *факторов*, не *учтённых в перечне объясняющих переменных* X
- Во-вторых, она может включать в себя *погрешность в измерении* Y.

В реальности возможны также ошибки в выборе вида функциональной зависимости и т.п.

В этом примере видна различная роль рассматриваемых переменных Y, N, P, Q, R, S, T, U (или, как их ещё называют, *факторов*).

В пределах данной модели одни переменные являются *независимыми* (или *объясняющими*, *экзогенными* – N, P, Q, R, S, T, U), а другие *зависимыми* (или *эндогенными* – Y).

В общем случае экономическая модель с одной *эндогенной* *Y* и *m* объясняющими переменными x_1, x_2 ,…, x_m имеет вид:

$$Y = f(x_1, x_2, \ldots, x_m, \varepsilon)$$

Конкретный вид функции может быть различным.

Выбор вида функциональной зависимости называется ***спецификацией модели***, а определение состава объясняющих переменных - ***спецификацией переменных***.

1.3. Статистические данные

Основу эконометрического моделирования составляют статистические данные. Их различают по типам.

Пространственные данные собираются по какому-либо экономическому показателю для разных объектов в один момент времени или в разные моменты в случае, когда время несущественно.

Например, данные по курсам покупки/продажи валюты сегодня в различных обменных пунктах Москвы.

Временные ряды **–** данные для одного объекта в различные моменты времени.

Например, ежеквартальные данные об инфляции, данные о денежной эмиссии за последние годы, ежедневный курс доллара.

Промежуточное положение занимают ***панельные данные***, которые отражают наблюдения по большому числу объектов за относительно небольшой период времени.

Например, данные об объёмах выпуска и капитальных вложениях российских предприятий топливно-энергетического комплекса за 1995–2000 г.

1.4. *Основные этапы эконометрического моделирования*

В процессе эконометрического исследования можно выделить следующие этапы:

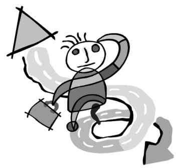
1-й этап (постановочный). Определение конечных целей моделирования, формирование набора участвующих в модели факторов и их ролей.

2-й этап (спецификация модели). Выбор общего вида модели, состава и вида входящих в неё функциональных взаимосвязей.

3-й этап (информационный). Сбор необходимой статистической информации.

4-й этап (идентификация модели). Анализ модели, статистическое оценивание параметров модели.

5-й этап (проверка модели). Сопоставление реальных и оценённых данных, проверка адекватности модели.



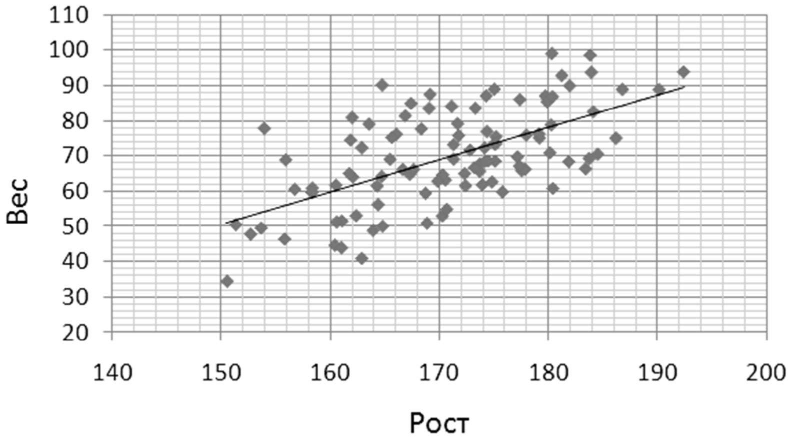
В результате проверки модели нередко приходится неоднократно возвращаться к предыдущим этапам и вносить коррективы.

2. Корреляционный анализ количественных зависимостей

2.1. *Выборочный коэффициент корреляции*

При изучении связи между факторами сразу возникает

Вопрос: в какой степени две переменные СОВМЕСТНО ИЗМЕНЯЮТСЯ? (т.е., влечёт ли за собой увеличение одной переменной увеличение или уменьшение другой, или не влечёт).



Представляет интерес наклон (*направление связи*) и ширина (*сила связи*) воображаемого эллипса

Тесноту линейной связи между двумя количественными факторами обычно измеряют при помощи коэффициента корреляции. Он вычисляется по формуле:

$$\boxed{r^*_{XY} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sqrt{D^*_x} \cdot \sqrt{D^*_y}}}$$

Известно, что

$$\boxed{-1 \le r^*_{XY} \le 1}$$

и линейная зависимость между X и Y тем сильнее, чем ближе *модуль коэффициента корреляции* к 1.

Пример.

X\Y	–1	0	2
0–20	5	3	4
20–40	2	1	5

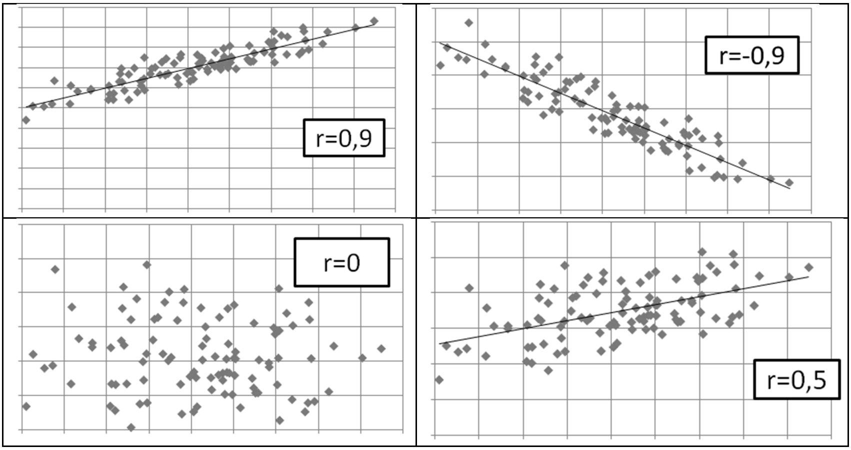
$$\overline{x} = \frac{10 \cdot 12 + 30 \cdot 8}{20} = \ldots,$$

$$\overline{y} = \ldots,$$

$$\overline{x^2} = \frac{10^2 \cdot 12 + \ldots}{20} = \ldots,\ D_x^* = \overline{x^2} - \left(\overline{x}\right)^2 = \ldots,$$

$$\overline{y^2} = \ldots,\ D_y^* = \overline{y^2} - \left(\overline{y}\right)^2 = \ldots$$

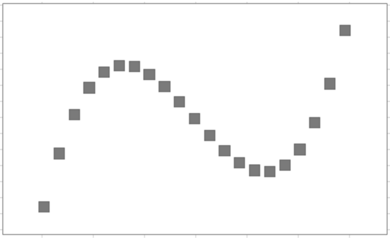
$$\overline{xy} = \frac{10 \cdot (-1) \cdot 5 + \ldots}{20} = \ldots;\ r_{xy}^* = \ldots$$



2.2. Замечания о коэффициенте корреляции

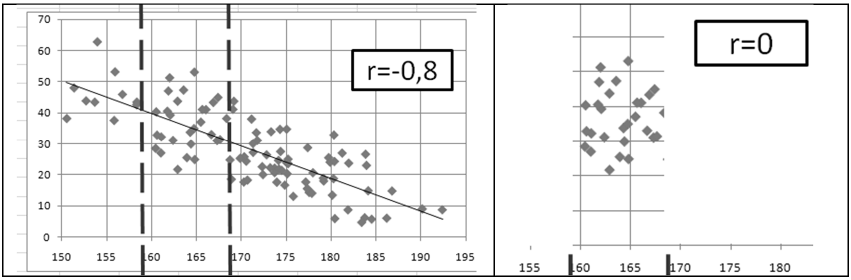
А. Коэффициент корреляции оценивает только линейную связь переменных!

Он не покажет наличие нелинейной связи.



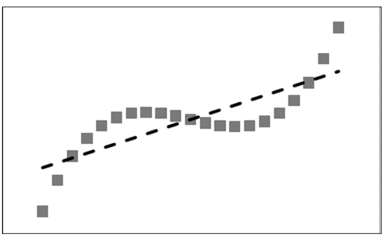
Здесь связь переменных есть, и она очень сильная, но r = 0.

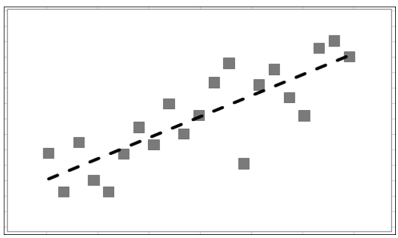
Б. Необходимо, чтобы у переменных была значительная изменчивость! Если сформировать выборку изначально однотипных по одному из рассматриваемых факторов, нечего надеяться выявить там корреляцию фактора с другими.

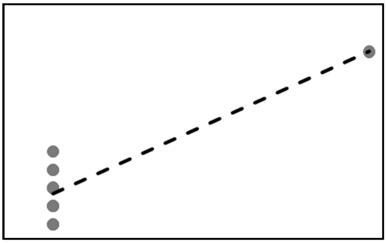


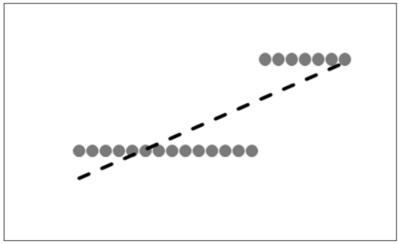
В. Корреляция совершенно не подразумевает наличие причинно-следственной связи!

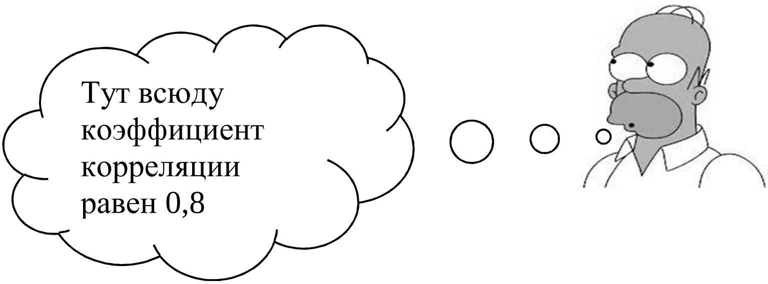
Г. Зная только коэффициент корреляции, нельзя сколько-нибудь детально описать зависимость.











2.3. *Матрица корреляций*

Корреляционная матрица — это квадратная таблица, в которой на пересечении строки и столбца находится коэффициент корреляции между соответствующими параметрами.

Пример. Для следующих табличных данных составить корреляционную матрицу

y	x_1	x_2
7	2	3
14	2	4
11	5	7
12	7	10

Решение. 1) Составляем матрицу

$$M = \begin{pmatrix} 7 & 2 & 3 \\ 14 & 2 & 4 \\ 11 & 5 & 7 \\ 12 & 7 & 10 \end{pmatrix}$$

2) Из каждого столбца вычитаем его среднее значение (например, из второго столбца вычитаем “4”):

$$M_1 = \begin{pmatrix} -4 & -2 & -3 \\ 3 & -2 & -2 \\ 0 & 1 & 1 \\ 1 & 3 & 4 \end{pmatrix}$$

3) Умножаем M_1^T на M_1 и делим на n, получаем ***матрицу ковариаций***:

$$M_1^T \cdot M_1 = \begin{pmatrix} -4 & 3 & 0 & 1 \\ -2 & -2 & 1 & 3 \\ -3 & -2 & 1 & 4 \end{pmatrix} \cdot \begin{pmatrix} -4 & -2 & -3 \\ 3 & -2 & -2 \\ 0 & 1 & 1 \\ 1 & 3 & 4 \end{pmatrix} / 4 = \begin{pmatrix} 6{,}5 & 1{,}25 & 2{,}5 \\ 1{,}25 & 4{,}5 & 5{,}75 \\ 2{,}5 & 5{,}75 & 7{,}5 \end{pmatrix}$$

На диагонали в матрице стоят дисперсии, например Dy=6,5.

4) Делим каждое число на корень из произведения соответствующих дисперсий (например, 1,25 на корень из 6,5·4,5, 2,5 на корень из 6,5·7,5, получаем матрицу корреляций:

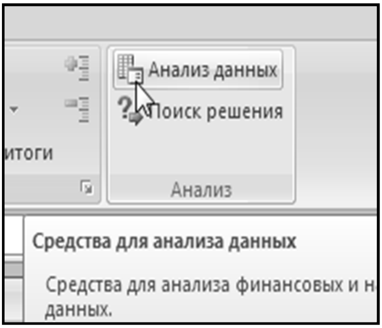
$$r_{YX} = \begin{pmatrix} 1 & 0{,}2311 & 0{,}3581 \\ 0{,}2311 & 1 & 0{,}9898 \\ 0{,}3581 & 0{,}9898 & 1 \end{pmatrix}$$

В MS Excel для вычисления корреляционных матриц используется инструмент **Корреляция** из надстройки **Анализ данных.**

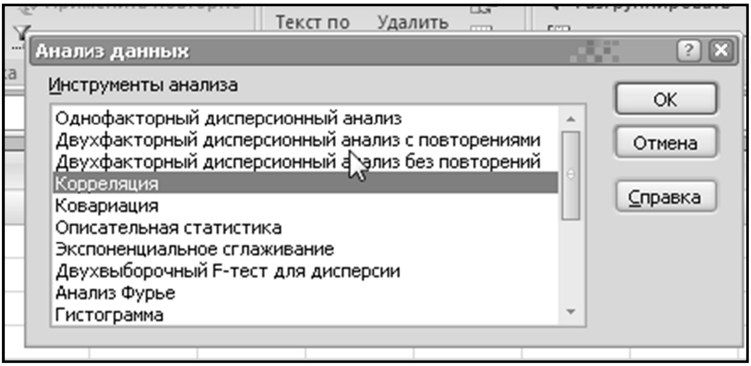
Книга1

	A	B	C
1	Y	X1	X2
2	7	2	3
3	14	2	4
4	11	5	7
5	12	7	10

Вызвать надстройку **Данные → Анализ данных**;

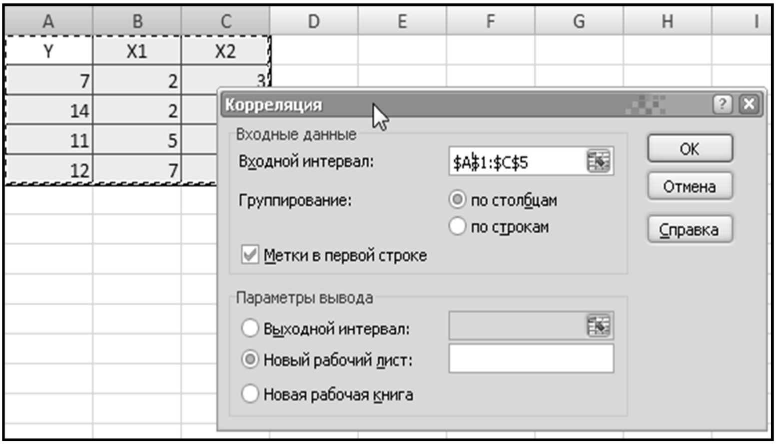


В появившемся списке **Инструменты анализа** выбрать строку **Корреляция** и нажать кнопку ОК;

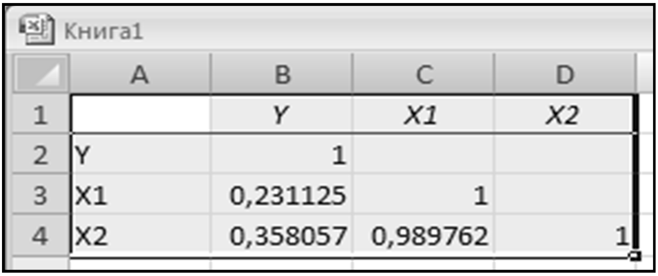


В появившемся окне указать **Входной интервал**.

Установить переключатель в разделе Группировка (“по столбцам” или “по строкам”);



В выходной диапазон будет выведена корреляционная матрица.



	A	B	C	D
1		*Y*	*X1*	*X2*
2	Y	1		
3	X1	0,231125	1	
4	X2	0,358057	0,989762	1

Хотя в результате будет получена треугольная матрица, корреляционная матрица симметрична. Подразумевается, что в пустых клетках в правой верхней половине таблицы находятся те же коэффициенты корреляции, что и в нижней левой (симметрично расположенные относительно диагонали из единиц).

2.4. Частная корреляция

Корреляция между двумя переменными, вычисленная после устранения влияния всех других переменных, называется ***частной корреляцией***.

Например, ***длина волос*** может коррелировать с ***ростом*** человека (чем выше человек, тем короче волосы), однако эта зависимость становится слабой или совсем исчезает, если устранить влияние ***пола*** людей, поскольку женщины обычно ниже ростом и чаще имеют более длинные волосы, чем мужчины.

Частный коэффициент корреляции между переменными X и Y при исключении влияния переменной Z вычисляется по формуле

$$\boxed{r_{XY|Z} = \frac{r_{XY} - r_{XZ} r_{YZ}}{\sqrt{\left(1 - {r_{XZ}}^2\right)\left(1 - {r_{YZ}}^2\right)}}}$$

Пример. Для предыдущего примера

$$r_{YX} = \begin{pmatrix} 1 & 0{,}2311 & 0{,}3581 \\ 0{,}2311 & 1 & 0{,}9898 \\ 0{,}3581 & 0{,}9898 & 1 \end{pmatrix}$$

$$r_{yx_1|x_2} = \frac{0{,}2311 - 0{,}3581 \cdot 0{,}9898}{\sqrt{\left(1 - 0{,}3581^2\right) \cdot \left(1 - 0{,}9898^2\right)}} = -0{,}925$$

$$r_{yx_2|x_1} = \frac{0{,}3581 - 0{,}2311 \cdot 0{,}9898}{\sqrt{\left(1 - 0{,}2311^2\right) \cdot \left(1 - 0{,}9898^2\right)}} = 0{,}9311$$

И так далее…

$$r_{YX}^{частн} = \begin{pmatrix} 1 & -0{,}925 & 0{,}9311 \\ -0{,}925 & 1 & 0{,}9984 \\ 0{,}9311 & 0{,}9984 & 1 \end{pmatrix}$$

3. Метод наименьших квадратов

3.1. Постановка задачи

Нередко возникает задача: найти кривую заданного вида, наиболее точно приближающую экспериментальные данные.

Математически это формулируется так:

требуется подобрать такие значения параметров β, чтобы график функции

$$\boxed{f(x)=\beta_1\varphi_1(x)+\beta_2\varphi_2(x)+\ldots+\beta_m\varphi_m(x)}$$

проходил как можно ближе к заданным точкам (x_i,y_i), $i=1,2..n$. *Здесь* β_i *– неизвестные коэффициенты,* φ_i *– известные функции.*

Нужно как-то конкретизировать понятие *близости* точек к кривой. Это можно сделать многими способами, наиболее распространён следующий:

Для заданных точек (x_i,y_i), $i=1,2..n$,

найти такие значения β_i,

чтобы минимизировать ***остаточную сумму квадратов***:

$$ESS=\sum_{i=1}^{n}(y_i-f(x_i))^2\rightarrow\min \qquad (*)$$

3.2. Формулы МНК

Приравнивая производные ESS нулю, получаем систему линейных уравнений для нахождения β_i

$$\begin{cases} A_{11}\beta_1 + & A_{12}\beta_2 + \ldots & A_{1m}\beta_m & = B_1 \\ A_{21}\beta_1 + & A_{22}\beta_2 + \ldots & A_{2m}\beta_m & = B_2 \\ \ldots & \ldots & \ldots & \ldots \\ A_{m1}\beta_1 + & A_{m2}\beta_2 + \ldots & A_{mm}\beta_m & = B_m \end{cases}$$

где

$$A_{kl} = \sum_{i=1}^{n} \varphi_k(x_i)\varphi_l(x_i), \quad B_k = \sum_{i=1}^{n} y_i \varphi_k(x_i)$$

Можно сразу и решить эту систему, если использовать *матричные* обозначения

$$\boxed{Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} \varphi_1(x_1) & \varphi_2(x_1) & \cdots & \varphi_m(x_1) \\ \varphi_1(x_2) & \varphi_2(x_2) & \cdots & \varphi_m(x_2) \\ \vdots & \vdots & \vdots & \vdots \\ \varphi_1(x_n) & \varphi_2(x_n) & \cdots & \varphi_m(x_n) \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}}$$

Тогда

$$\boxed{\beta = \left(X^T X\right)^{-1} X^T Y}$$

Пример. Имеются следующие экспериментальные данные

№ п/п	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
X	2	3	2	2	4	3	5	5
Y	1	2	4	3	4	5	7	9

Для этих данных требуется подобрать наилучшую параболу
$y = \beta_1 + \beta_2 x + \beta_3 x^2$ методом наименьших квадратов

$$A_{11} = \sum \varphi_1\varphi_1 = \sum (1 \cdot 1) = 8, \quad A_{12} = A_{21} = \sum \varphi_1\varphi_2 = \sum (1 \cdot x) = 26,$$

$$A_{13} = A_{31} = \sum \varphi_1\varphi_3 = \sum \left(1 \cdot x^2\right) = 96, \; A_{22} = \sum \varphi_2\varphi_2 = \sum x^2 = 96,$$

$$A_{23} = A_{32} = \sum \varphi_2\varphi_3 = .., \quad A_{33} = ...$$

$$B_1 = \sum y\varphi_1 = \sum y = 35, \quad B_2 = \sum (y\varphi_2) = \sum (yx) = ..., \; B_3 = ...$$

Получаем систему уравнений

$$\begin{cases} 8\beta_1 + 26\beta_2 + 96\beta_3 = 35 \\ 26\beta_1 + 96\beta_2 + 392\beta_3 = 133, \\ 96\beta_1 + 392\beta_2 + 96\beta_3 = 559 \end{cases}$$

Откуда $\beta_1 \approx 6{,}15,\ \beta_2 \approx -3{,}06,\ \beta_3 \approx 0{,}68$.

Таким образом, наилучшее МНК-приближение в виде параболы для исходных данных имеет вид
$y = 6{,}15 – 3{,}06x + 0{,}68\ x^2$

Интересно сравнить экспериментальные и расчётные значения для этой формулы:

x_i	2	3	2	2	4	3	5	5
y_i	**1**	**2**	**4**	**3**	**4**	**5**	**7**	**9**
$y_{i,расч}$	**2,76**	**3,1**	**2,76**	**2,76**	**4,8**	**3,1**	**7,87**	**7,87**

4. Классическая модель множественной линейной регрессии

4.1. Описание модели

Модель множественной линейной регрессии имеет вид

$$\boxed{Y = \beta_0 + \beta_1 x_1 + ... + \beta_m x_m + \varepsilon \qquad (4.1)}$$

где y – *зависимая переменная*,
$x_1,\ldots, x_m$ – *объясняющие* переменные (также называемые *регрессорами*),
$\beta_0, \beta_1, \ldots, \beta_m$ – *коэффициенты регрессии*,
ε – *случайный член* (или "*ошибки измерений*").

Коэффициенты $\beta_0,\beta_1 ,\ldots, \beta_m$ оцениваются на основе статистических данных.

Для нахождения оценок проводятся n наблюдений (n>m), в результате чего получается n равенств

$$\boxed{y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_m x_{mi} + \varepsilon_i, \quad i = 1,2,...,n}$$

(индекс i показывает *номер наблюдения*).

Это удобно записать в *матричной форме*

$$\boxed{Y = X\beta + \varepsilon, \quad (4.2)}$$

где

$$\boxed{Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{m1} \\ 1 & x_{12} & \cdots & x_{m2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & \cdots & x_{mn} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad (4.3)}$$

4.2. Оценки параметров регрессии

Будем искать оценки коэффициентов β в задаче (4.1) ***методом наименьших квадратов***. Другими словами, оценки должны удовлетворять условию

$$ESS = \sum (y_i - \hat{y}_i)^2 \to \min; \quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \dots$$

или, в матричной форме,

$$(Y - X\hat{\beta})^T \cdot (Y - X\hat{\beta}) \to \min$$

Если продифференцировать это равенство по β и приравнять производные нулю, получим

$$\boxed{X^T X\hat{\beta} = X^T Y}$$

Отсюда

$$\boxed{\hat{\beta} = \left(X^T X\right)^{-1} X^T Y, \quad (4.4)}$$

Пример. В таблице представлены объём импорта (Y) и ВНП (X_1) и индекс потребительских цен (X_2) в США

Годы	**1964**	**1965**	**1966**	**1967**	**1968**
$\mathbf{X_1}$	636	689	753	796	868
$\mathbf{X_2}$	93	95	97	100	104
Y	28	32	38	41	53

Требуется построить оценки параметров уравнения линейной регрессии Y на X_1, X_2:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon.$$

Решение.

$$Y=\begin{pmatrix}28\\32\\38\\41\\53\end{pmatrix},\ X=\begin{pmatrix}1&636&93\\1&689&95\\1&753&97\\1&796&100\\1&868&104\end{pmatrix},\ \hat{\beta}=\begin{pmatrix}\hat{\beta}_0\\\hat{\beta}_1\\\hat{\beta}_2\end{pmatrix},$$

$$X^TX=\begin{pmatrix}1&1&1&1&1\\636&689&753&796&868\\93&95&97&100&104\end{pmatrix}\cdot\begin{pmatrix}1&636&93\\1&689&95\\1&753&97\\1&796&100\\1&868&104\end{pmatrix}=\begin{pmatrix}5&3742&489\\3742&2833266&367516\\489&367516&47899\end{pmatrix};$$

$$\left(X^TX\right)^{-1}=\begin{pmatrix}2452{,}8&1{,}822&-39{,}02\\1{,}822&0{,}0014&-0{,}03\\-39{,}02&-0{,}03&0{,}625\end{pmatrix}\quad \hat{\beta}=\left(X^TX\right)^{-1}X^TY=\begin{pmatrix}-150{,}99\\0{,}02\\1{,}78\end{pmatrix}$$

$$\boxed{y=-150{,}99+0{,}02x_1+1{,}78x_2}$$

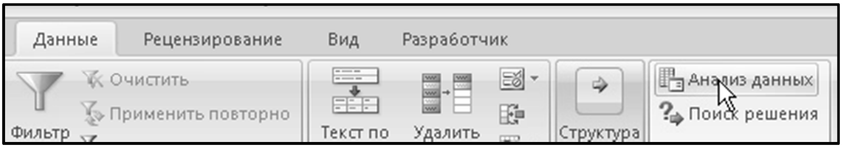
4.3. Нахождение параметров регрессии при помощи MS Excel

Для нахождения параметров регрессии
при помощи MS Excel
необходимо, чтобы была установлена надстройка
«***Анализ данных***».

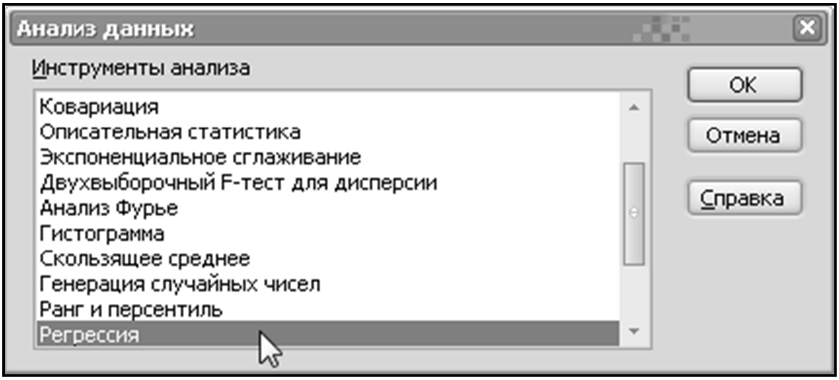
Пример. Возьмём тот же пример и внесём данные в таблицу MS Excel *по столбцам*:

	A	B	C	D
1	Годы	X1	X2	Y
2	1964	636	93	28
3	1965	689	95	32
4	1966	753	97	38
5	1967	796	100	41
6	1968	868	104	53

Выберем пункт меню ***Данные – Анализ данных***

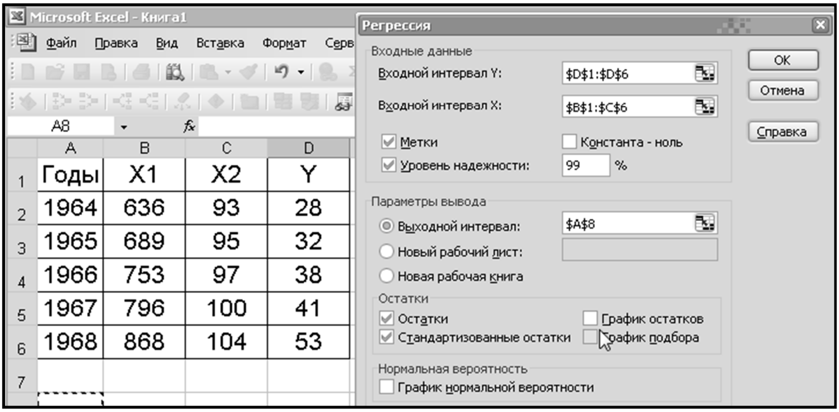


Укажем инструмент анализа *Регрессия*:



Появится диалог “**Регрессия**".

Указываем области, где находятся значения X, Y (см. рис.), указываем, куда выводить результаты и нажимаем кнопку **ОК**.



Программа выдаст множество данных по регрессии, в том числе и оценки параметров регрессии:

Дисперсионный анализ		
	df	*SS*
Регрессия	2,0000	361,6094
Остаток	2,0000	7,5906
Итого	4,0000	369,2000

	Коэффициенты	*Стандартная ошибка*	*t-стат*
Y-пересечение	-150,9880	96,4838	
X1	0,0200	0,0736	
X2	1,7832	1,5403	

В частности, уравнение регрессии:

$$\begin{array}{ccccccc} y & = & -150{,}99 & + & 0{,}02x_1 & + & 1{,}78x_2 \\ & & (96{,}48) & & (0{,}07) & & (1{,}54) \end{array}$$

(под коэффициентами в скобках написаны их стандартные отклонения – см. далее).

4.4. Основные гипотезы

Гипотезы, лежащие в основе *классической модели множественной регрессии*, следующие:

(A) "Истинная" зависимость Y от регрессоров X имеет вид $y=X\beta+\varepsilon$.

(B) Величины X – детерминированные (не случайные).

(C) Столбцы матрицы X линейно независимы (другими словами, *ранг матрицы* X *равен* m+1).

(D) $M\varepsilon_i = 0$, $D\varepsilon_i = \sigma^2$ не зависит от i

(E) $\operatorname{cov}(\varepsilon_i \varepsilon_j)=0$ при $i \neq j$

(*некоррелированность* ошибок для разных наблюдений).

Замечание. Вместо (D) часто добавляют условие

(F) Случайные факторы ε_i имеют *нормальное* распределение $N(0,\sigma^2)$.

В этом случае модель называется *нормальной*.

Замечание. Условия (D) – (F) удобно записать в *матричной форме*

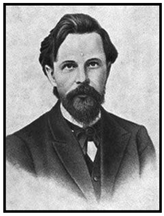
(G) $M\varepsilon = 0$, $M(\varepsilon\varepsilon^T) = \sigma^2 E$ (E – единичная матрица)

($F^/$) $\varepsilon \sim N(0, \sigma^2 E)$

4.5. Теорема Гаусса-Маркова

Теорема. В предположениях (А) – (Е) оценки, полученные методом наименьших квадратов по формуле (4.4), являются *несмещёнными* и обладают *наименьшей дисперсией* среди всех линейных несмещённых оценок параметров β.

$$\boxed{\hat{\beta} = \left(X^T X\right)^{-1} X^T Y, \quad (4.4)}$$



4.6. Коэффициент детерминации

Дисперсия зависимой переменной, помноженная на n, (или ***полная сумма квадратов***) обозначается TSS:

$$\boxed{TSS = \sum \left(y_i - \bar{y}\right)^2 = nD_y^*}$$

Сумма

$$\boxed{RSS = \sum \left(\hat{y}_i - \bar{y}\right)^2 = nD_{\hat{y}}^*}$$

содержит “объяснённую часть дисперсии”.
И, наконец, сумма квадратов “остатков

$$\boxed{ESS = \sum \left(y_i - \hat{y}_i\right)^2 = \sum e_i^2 = nD_e^*}$$

представляет собой *“необъясненную регрессией часть дисперсии”* зависимой переменной.
Нетрудно проверить, что

$$\boxed{RSS + ESS = TSS}$$

Величина

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

называется ***коэффициентом детерминации*** и представляет собой долю дисперсии зависимой переменной y, объясненную при помощи уравнения регрессии.

$$\boxed{0 \le R^2 \le 1}$$

Чем ближе коэффициент детерминации к 1, тем лучше приближаются исходные данные уравнением линейной регрессии.

Замечание. В модели *множественной линейной регрессии*, кроме β, есть ещё один параметр – σ^2. Его оценку можно найти по формуле

$$\boxed{\hat{\sigma}^2 = S^2 = \frac{ESS}{n-m-1}}$$

Пример. В рассмотренном примере

$$Y = \begin{pmatrix} 28 \\ 32 \\ 38 \\ 41 \\ 53 \end{pmatrix}, \hat{Y} = X\hat{\beta} = \begin{pmatrix} 1 & 636 & 93 \\ 1 & 689 & 95 \\ 1 & 753 & 97 \\ 1 & 796 & 100 \\ 1 & 868 & 104 \end{pmatrix} \cdot \begin{pmatrix} 150{,}99 \\ 0{,}03 \\ 1{,}78 \end{pmatrix} = \begin{pmatrix} 27{,}59 \\ 32{,}22 \\ 37{,}07 \\ 43{,}28 \\ 51{,}85 \end{pmatrix},$$

$$\overline{Y} = \frac{28+32+\ldots}{5} = 38{,}4; \quad RSS = (27{,}59 - 38{,}4)^2 + \ldots = 361{,}609;$$

$$TSS = (28 - 38{,}4)^2 + \ldots = 369{,}24; \quad ESS = (28 - 27{,}59)^2 + \ldots = 7{,}5906;$$

$$S^2 = \frac{ESS}{5-2-1} = 3{,}7953; \quad R^2 = \frac{361{,}609}{369{,}24} = 0{,}9794$$

Замечание. На MS Excel эти параметры выглядят следующим образом:

ВЫВОД ОСТАТКА

Наблюдение	Предсказанное Y	Остатки
1	27,58910468	0,4108953
2	32,21717357	-0,2171735
3	37,06559179	0,9344082
4	43,27653597	-2,2765359
5	51,85159399	1,148406

$$\hat{Y} = \begin{pmatrix} 27{,}59 \\ 32{,}22 \\ 37{,}07 \\ 43{,}28 \\ 51{,}85 \end{pmatrix},$$

Дисперсионный анализ

	df	SS	MS
Регрессия	2	361,6094	180,8047
Остаток	2	7,5906	3,7953
Итого	4	369,2000	

$$RSS = 361{,}609;$$
$$ESS = 7{,}5906;$$
$$TSS = 369{,}24;$$

Регрессионная статистика	
Множественный R	0,9897
R-квадрат	0,9794
Нормированный R-кв	0,9589
Стандартная ошибка	1,9481
Наблюдения	5

$$R^2 = 0{,}9794$$
$$S^2 = 3{,}7953;$$
$$S = \sqrt{S} = 1{,}9481$$

Замечание. Важное свойство коэффициента детерминации R^2 состоит в том, что при добавлении в модель новых объясняющих переменных x он <u>не может уменьшиться</u>. Поэтому при сравнении двух моделей с различным числом переменных не всегда ясно, за счёт чего возрос показатель детерминации: за счёт реального влияния дополнительного фактора на результат или просто из-за возрастания числа переменных.

Для того чтобы можно было *сравнивать различные модели*, вводят так называемый ***скорректированный (***или ***нормированный) коэффициент детерминации***:

$$\boxed{R^2_{корр} = 1 - \frac{ESS/(n-m-1)}{TSS/(n-1)}}$$

Это число также не превосходит 1, но в некоторых случаях может оказаться и отрицательным. В рассматриваемом примере $R^2_{корр} = 0{,}9589$

4.7. Квантили

Формулы, описывающие статистические свойства оценок, содержат табличные значения, называемые *квантилями*.

Квантиль ***нормального закона*** N(0;1): u_p

p	0,9	0,95	0,975	0,99	0,995	0,999	0,9995
u_p	1,282	1,645	1,960	2,325	2,576	3,090	3,291

Свойство:

$$\boxed{u_{1-\alpha} = -u_\alpha}$$

Квантиль ***закона распределения хи-квадрат***: $\chi^2_p(k)$

k \ p	0,01	0,025	0,05	0,1	0,9	0,95	0,975	0,99
1	0,0002	0,001	0,004	0,0158	2,71	3,84	5,02	6,63
2	0,02	0,05	0,103	0,211	4,61	5,99	7,38	9,21

3	0,115	0,216	0,352	0,584	6,25	7,81	9,35	11,3
4	0,297	0,484	0,711	1,06	7,78	9,49	11,1	13,3
10	2,56	3,25	3,94	4,87	16	18,3	20,5	23,2
20	7,63	9,59	10,9	12,4	28,4	31,4	34,2	37,6
30	14,3	16,8	18,5	20,6	40,3	43,8	47,0	50,9
40	22,2	24,4	26,5	29,1	51,8	55,8	59,3	63,7
50	29,7	32,4	34,8	37,7	63,2	67,5	71,4	76,2
75	49,5	53	56,1	59,8	91,1	96,2	100,8	106,4

$$\boxed{\chi_p^2(k) \approx k \cdot \left(1 - \frac{2}{9k} + u_p\sqrt{\frac{2}{9k}}\right)^3}$$ при больших k.

Квантиль ***<u>закона распределения Стьюдента</u>***: t_p (k)

p / k	0,9	0,95	0,975	0,99	0,995
1	3,078	6,314	12,706	31,821	63,657
2	1,886	2,92	4,303	6,965	9,925
3	1,638	2,353	3,182	4,541	5,841
4	1,533	2,132	2,776	3,747	4,604
5	1,476	2,015	2,571	3,365	4,032
10	1,372	1,812	2,228	2,764	3,169
20	1,325	1,725	2,086	2,528	2,845
30	1,312	1,697	2,042	2,457	2,750
40	1,303	1,684	2,021	2,423	2,704
120	1,289	1,658	2,98	2,358	2,617

Свойство:

$$\boxed{t_{1-\alpha}(k) = -t_\alpha(k)}$$

$$\boxed{t_p(k) \approx u_p \cdot \left(1 - \frac{u_p^{\ 2} + 1}{4k}\right)^{-1}}$$ при больших k.

Квантиль ***закона распределения Фишера***: $F_p(k_1,k_2)$

Квантили $\mathbf{F_{0,9}(n_1,n_2)}$ закона распределения Фишера

n_1 \ n_2	1	2	3	4	5	10	15	20	30	120
1	40	8,53	5,54	4,54	4,06	3,29	3,07	2,97	2,88	2,75
2	49,5	9	5,46	4,32	3,78	2,92	2,7	2,59	2,49	2,35
3	53,6	9,16	5,39	4,19	3,62	2,73	2,49	2,38	2,28	2,13
4	55,8	9,24	5,34	4,11	3,52	2,61	2,36	2,25	2,14	1,99
30	62,2	9,46	5,17	3,82	3,17	2,16	1,87	1,74	1,61	1,41

Квантили $\mathbf{F_{0,95}(n_1,n_2)}$ закона распределения Фишера

n_1 \ n_2	1	2	3	5	10	15	20	30	120
1	161	18,5	10,13	6,61	4,96	4,54	4,35	4,24	3,92
2	199	19	9,55	5,79	4,1	3,68	3,49	3,39	3,07
3	216	19,16	9,28	5,41	3,71	3,29	3,1	2,99	2,68
4	225	19,25	9,12	5,19	3,48	3,05	2,87	2,76	2,45
30	250	19,46	8,62	4,5	2,7	2,25	2,04	1,84	1,55

Свойство:

$$\boxed{F_{1-\alpha}(k_1,k_2)=\frac{1}{F_\alpha(k_2,k_1)}}$$

В последних версиях MSExcel квантили можно найти так:

$$u_p = НОРМ.СТ.ОБР(p)$$

$$\chi^2_p(k) = \text{ХИ2.ОБР}(p,k)$$

$$t_p(k) = \text{СТЬЮДЕНТ.ОБР}(p,k)$$

$$F_p(k_1,k_2) = \text{F.ОБР}(p,k_1,k_2)$$

В более старых версиях MS Excel некоторые квантили вычисляются “задом наперёд”

$$\chi^2_{\alpha}(k) = \text{ХИ2ОБР}(1-\alpha, k)$$
$$\chi^2_{1-\alpha}(k) = \text{ХИ2ОБР}(\alpha, k)$$
$$t_{1-\frac{\alpha}{2}}(k) = \text{СТЬЮДРАСПОБР}(\alpha, k)$$
$$F_{1-\alpha}(k_1, k_2) = \text{FРАСПОБР}(\alpha, k_1, k_2)$$

4.8. Статистические свойства оценок параметров регрессии

Теорема. В предположениях (A) – (E) и (F) теоремы Гаусса-Маркова оценки, полученные методом наименьших квадратов по формуле (2.4), обладают следующими свойствами:

1. $\dfrac{(n-m-1)\cdot S^2}{\sigma^2} \sim \chi^2(n-m-1);$
2. S^2 *и* $\hat{\beta}$ *независимы.*

4.9. Доверительные интервалы

Будем считать, что условие нормальности **(F)** выполнено. Тогда можно построить ***доверительные интервалы*** для параметров регрессии, т.е. интервалы, ***содержащие точные значения этих параметров с заданной вероятностью 1–α***.

Вот соответствующие формулы:

$$\boxed{\hat{\beta}_i - t_{1-\alpha/2}(n-m-1)\cdot S_{\beta_i} < \beta_i < \hat{\beta}_i + t_{1-\alpha/2}(n-m-1)\cdot S_{\beta_i}}$$

$$\boxed{\frac{ESS}{\chi^2_{1-\alpha/2}(n-m-1)} < \sigma^2 < \frac{ESS}{\chi^2_{\alpha/2}(n-m-1)}.}$$

Здесь α – заданный “***уровень значимости***”, т.е. ***вероятность того, что построенный интервал не содержит истинного значения параметра***.

Далее, S_β –*стандартные ошибки параметров*:

$$\boxed{S^2_{\beta_i} = S^2 \cdot (X^T \cdot X)^{-1}_{ii}}$$

Как правило, стандартные ошибки параметров вычисляются в эконометрических программах вместе с оценками самих параметров

Пример. В рассмотренном примере

$$\left(X^T X\right)^{-1} = \begin{pmatrix} 2452{,}8 & 1{,}822 & -39{,}02 \\ 1{,}822 & 0{,}0014 & -0{,}03 \\ -39{,}02 & -0{,}03 & 0{,}625 \end{pmatrix}, \quad y = -150{,}99 + 0{,}02x_1 + 1{,}78x_2, \quad S^2 = 3{,}7953$$

$$S^2_{\beta_0} = S^2 \cdot 2452{,}8 = 9309{,}16; \; S_{\beta_0} = 96{,}48$$

$$S^2_{\beta_1} = S^2 \cdot 0{,}0014 = 0{,}0053; \; S_{\beta_1} = 0{,}07$$

$$S^2_{\beta_2} = S^2 \cdot 0{,}625 = 2{,}372; \; S_{\beta_2} = 1{,}54$$

В MS Excel стандартные ошибки коэффициентов находятся здесь:

Дисперсионный анализ			
	df	*SS*	*M*
Регрессия	2,0000	361,6094	
Остаток	2,0000	7,5906	
Итого	4,0000	369,2000	
	Коэффициенты	***Стандартная ошибка***	*t-стат*
Y-пересечение	**-150,9880**	**96,4838**	
X1	**0,0200**	**0,0736**	
X2	**1,7832**	**1,5403**	

Далее, на уровне значимости α=0,1:

$$t_{1-\alpha/2}(n-m-1) = t_{0,95}(2) = 2,92;$$

$$-150,99 - 2,92\cdot 96,48 = -432,72 < \beta_0 < 130,74 = -150,99 + 2,92\cdot 96,48$$

Аналогично,

$$-0,19 < \beta_1 < 0,23;$$
$$-2,71 < \beta_2 < 6,28;$$

Наконец,

$$\chi^2_{1-\alpha/2}(n-m-1) = \chi^2_{0,95}(2) = 5,99;\ \chi^2_{\alpha/2}(n-m-1) = \chi^2_{0,05}(2) = 0,1;$$

$$\frac{ESS}{5,99} = \frac{7,5906}{5,99} = 1,27 < \sigma^2 < 75,9 = \frac{7,5906}{0,1}$$

4.10. Проверка гипотез о параметрах регрессии

В этом параграфе также будем считать, что условие нормальности **(F)** выполнено:

(F) Случайные факторы ε_i имеют *нормальное* распределение $N(0,\sigma^2)$.

При исследовании модели приходится проверять гипотезы о равенстве истинного значения параметра регрессии заданному числу. Для этого **достаточно построить доверительный интервал и посмотреть, попадает ли в него это число**.

Особенно часто проверяют гипотезу $\mathbf{H_0}$**:** $\beta_i=0$.

Дело в том, что если принимается альтернативная гипотеза

$$\boxed{H_1 : \beta_i \neq 0}$$

то говорят, что коэффициент β_i ***значим***.

Таким образом, получаем следующий ***критерий значимости параметра*** $\boldsymbol{\beta_i}$**:**

$$\frac{\left|\hat{\beta}_i\right|}{S_{\beta_i}} \geq t_{1-\alpha/2}(n-m-1) \qquad (4.5)$$

Пример. В рассмотренном примере было

$$-0{,}19 < \beta_1 < 0{,}23;$$
$$-2{,}71 < \beta_2 < 6{,}28;$$

Доверительные интервалы для коэффициентов содержат 0, так что оба коэффициента *не значимы*.

Замечание. Дробь в левой части неравенства 4.5 называют ***t-статистикой***. Эти значения в MS Excel автоматически:

	Коэффициен	артная	t-статистика	P-Значение
Y-пересечени	1665,367	1442,75	1,154294475	0,264348886
Total investm	-68,6798	42,0316	-1,634003237	0,120639348
Inflation, aver	-1,3E-11	4,43E-1	-0,290871293	0,774668562
Unemployme	-34,9101	30,5155	-1,144010272	0,268472918
Population	7,296819	1,18972	6,133186538	1,10426E-05
General gover	20,79136	21,9242	0,948326344	0,356250483
Current accou	-0,51724	3,00883	-0,171907903	0,865539759

Замечание. Значимость коэффициентов удобнее всего оценивать, используя ***P-значение***. Это минимальное значение уровня значимости, на котором коэффициент ещё значим. Таким образом, для значимости коэффициента необходимо и достаточно, чтобы P-значение было <u>меньше α.</u>

	Коэффициен	артная о	t-статистика	P-Значение
Y-пересечени	1665,367	1442,757	1,15429447	0,264348886
Total investm	-68,6798	42,03165	-1,63400323	0,120639348
Inflation, aver	-1,3E-11	4,43E-11	-0,29087129	0,774668562
Unemployme	-34,9101	30,51556	-1,14401027	0,268472918
Population	7,296819	1,189727	6,13318653	1,10426E-05
General gover	20,79136	21,92427	0,94832634	0,356250483
Current accou	-0,51724	3,008832	-0,17190790	0,865539759

На рисунке значимым на уровне 0,1 является только коэффициент при Population.

Задача. По выборке объёма 8 получено следующее уравнение линейной регрессии:

$$\begin{array}{ccccccccc} y & = & 5{,}27 & - & 0{,}89x_1 & + & 1{,}69x_2 & - & 0{,}41x_3 \\ & & (2{,}97) & & (0{,}53) & & (0{,}58) & & (0{,}32) \end{array}$$

Какие из его коэффициентов значимы на уровне значимости α =0,1?

Решение.

$$n = 8,\ m = 3,\ t_{1-\alpha/2}(n-m-1) = t_{0,95}(4) = 2{,}13.$$

$$\frac{0{,}89}{0{,}53} \approx 1{,}68 < 2{,}13 \Rightarrow \ldots$$

$$\frac{1{,}69}{0{,}59} \approx \ldots \Rightarrow \ldots$$

4.11. Проверка гипотезы о значимости уравнения регрессии

При построении уравнения возникает вопрос: в какой мере можно ему доверять, насколько оно *значимо*?

Уравнение считается значимым, если отвергается гипотеза о том, что все коэффициенты при переменных равны нулю.

Для проверки *значимости уравнения* используется ***F-статистика Фишера***:

$$\boxed{F = \frac{RSS/m}{ESS/(n-m-1)}}$$

Разделив числитель и знаменатель на TSS, получим ещё одну формулу для *F:*

$$\boxed{F = \frac{R^2/m}{\left(1-R^2\right)/\left(n-m-1\right)}}$$

> Если вычисленное значение F-статистики оказывается больше $F_{1-\alpha}(m, n - m - 1)$,
> то оцененное уравнение регрессии является
> значимым при выбранном α

Задача. По выборке получено уравнение линейной регрессии: $y = 4{,}8 - 2{,}15x_1 + 2{,}18x_2$

y	x_1	x_2
6	2	3
10	2	4
10	5	7
11	7	10
3	7	6

Значимо ли это уравнение на уровне 0,1?

Решение. 1) Вычислим расчётные значения Y:

$\hat{y}$
7,04
9,22
9,31
11,55
2,83

2) Вычислим RSS и ESS

$$3) F = \frac{RSS/m}{ESS/\left(n-m-1\right)} = \frac{43{,}5/2}{2{,}5/\left(5-2-1\right)} = 17{,}4$$

Квантиль $F_{0,9}(2;5\text{-}2\text{-}1)=F_{0,9}(2;2)$ по таблице равен *9*.

Так как 17,4 >9, то уравнение значимо.

Замечание. В MS Excel ***F-значение*** вычисляется автоматически:

Дисперсионный анализ					
	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	6	63713945	10618990,9	14,2521	8,42827E-06
Остаток	17	12666421	745083,611		
Итого	23	76380367			

Замечание. Значимость уравнения в MS Excel удобнее всего оценивать, используя величину “***значимость F”***. Это – *минимальное значение уровня α, на котором уравнение ещё значимо*. Таким образом, для значимости уравнения необходимо и достаточно, чтобы число “***Значимость F”*** было меньше α.

Дисперсионный анализ					
	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	6	63713945	10618990,9	14,2521	8,42827E-06
Остаток	17	12666421	745083,611		
Итого	23	76380367			

5. Спецификация модели

5.1. Основные этапы

Спецификация модели включает следующие этапы:

1. Определение списка объясняющих и зависимых переменных.

2. Выбор функциональной формы модели.

Рассмотрим подробнее первый пункт

5.2. Отбор объясняющих переменных

При определении состава факторов, включаемых в уравнение регрессии, руководствуются **теоретическими представлениями** о взаимосвязях этих факторов. Однако часто встречается ситуация, когда имеется большое число факторов, **но нет априорной модели изучаемого явления**, так что неясно, какие именно переменные включать в модель. В этом случае используются различные **эвристические процедуры**. Приведём простой пример такой процедуры

Процедура пошагового отбора независимых переменных.

1. Из исходного набора переменных отбирается (включается в модель) переменная, имеющая **наибольший по модулю коэффициент корреляции с зависимой переменной Y**.
2. На каждом последующем шаге в модель **добавляется та** из переменных, добавление которой **максимально увеличивает скорректированный коэффициент детерминации**.

Пример. Произвести пошаговый отбор переменных регрессии для выборки**.**

Y	**X_1**	**X_2**	**X_3**
12,4	1,1	1,0	2,2
23,5	2,0	1,4	3,0
43,5	3,5	5,0	3,0
38,6	4,0	3,4	4,6
25,1	2,7	2,1	3,0
38,0	4,1	2,0	6,0
59,8	5,0	8,1	8,0
54,3	7,0	4,0	9,0
85,7	8,0	11,0	3,0
52,2	6,4	5,0	8,0

Решение. 1) Найдём матрицу корреляций.

Наибольший по модулю коэффициент корреляции с Y имеет переменная X_2.

	Y	*X_1*	*X_2*	*X_3*
Y	1,00			
X_1	0,93	1,00		
X_2	0,94	0,77	1,00	
X_3	0,40	0,59	0,21	1,00

2) Строим регрессию Y на X_2.

$$y = 16{,}48 + 6{,}24 x_2, \quad R^2_{корр} = 0{,}868$$

3.1) Добавим к X_2 переменную X_1.

$$y = 6{,}65 + 7{,}79 x_1 + 3{,}65 x_2, \quad R^2_{корр} = 0{,}986$$

3.2) Добавим к X_2 переменную X_3.

$$y = 8{,}86 + 5{,}944 x_2 + 1{,}786 x_3, \quad R^2_{корр} = 0{,}9$$

$R^2_{корр}$ сильнее возрос в варианте 3.1. Оставляем этот вариант.
4.1) Добавим к оставленному варианту (X_1 , X_2) переменную X_3.

$$y = 6{,}9 + 4{,}96x_1 + 3{,}58x_2 - 0{,}145x_3, \quad R^2_{корр} = 0{,}980$$

Больше добавлять к (X_1 , X_2) нечего – других вариантов нет.
Ответ. Результат пошагового отбора переменных:

$$y = 6{,}65 + 7{,}79x_1 + 3{,}65x_2, \quad R^2_{корр} = 0{,}986$$

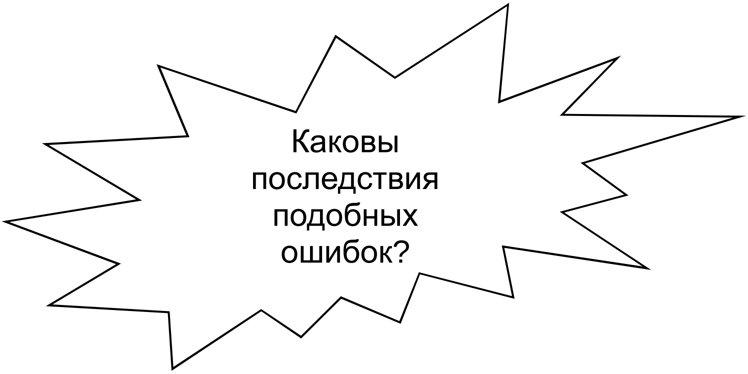
5.3. Влияние ошибок спецификации переменных

Неправильная спецификация переменных может произойти из-за

- отбрасывания переменной, которая должна быть в составе объясняющих переменных

  или

- включения “лишней” переменной, которой не должно быть в составе объясняющих переменных.



5.3.1. Пропуск существенных переменных

Пусть истинная зависимость имеет вид

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m + \gamma_1 z_1 + \ldots + \gamma_l z_l + \varepsilon, \quad (A)$$

или, в матричной форме,

$$Y = X\beta + Z\gamma + \varepsilon, \quad (A)$$

а мы, не зная об этом, ищем регрессию в виде:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m + \varepsilon, \quad (B)$$

то есть,

$$Y = X\beta + \varepsilon \quad (B)$$

В случае такого пропуска существенных переменных:

- ❖ МНК-оценка параметров β в общем случае оказывается смещённой;
- ❖ Стандартные ошибки коэффициентов β имеют неотрицательное смещение;
- ❖ Оценка дисперсии σ^2 имеет неотрицательное смещение

Таким образом, **последствия** достаточно **серьёзны**.

5.3.2. Включение несущественных переменных

Пусть истинная зависимость имеет вид

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m + \varepsilon, \quad (B)$$

или, в матричной форме,

$$Y = X\beta + \varepsilon \quad (B)$$

а мы, не зная об этом, ищем регрессию в виде:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m + \gamma_1 z_1 + \ldots + \gamma_l z_l + \varepsilon, \quad (A)$$

то есть,

$$Y = X\beta + Z\gamma + \varepsilon, \quad (A)$$

Тогда

- МНК-оценки параметров β несмещённые;
- Стандартные ошибки коэффициентов β имеют неотрицательное смещение;
- Оценка дисперсии σ^2 имеет неотрицательное смещение

Последствия такой ошибки спецификации являются не столь серьёзными, как в предыдущем случае.

Однако увеличение стандартных ошибок может привести к неверным выводам при проверке гипотез и ухудшению интервальных оценок.
В частности, может возникнуть незначимость коэффициентов.

5.4. Фиктивные переменные

Как правило, независимые переменные принимают непрерывный ряд значений (цена товара, уровень инфляции и т.д.).

Однако нередко приходится учитывать *качественные* признаки.

Например, можно изучать влияние пола работника на уровень его зарплаты (переменная “пол” принимает всего два значения), или спросом на прохладительные напитки в зависимости от времени года (сколько различных значений принимает эта переменная?).

В подобных ситуациях “качественному” значению условно присваивается числовая метка, как правило, это 0 или 1.

Если же качественный признак принимает k значений (k>2), то вводят <u>k–1</u> переменную, каждая из которых принимает
значения 0 и 1.

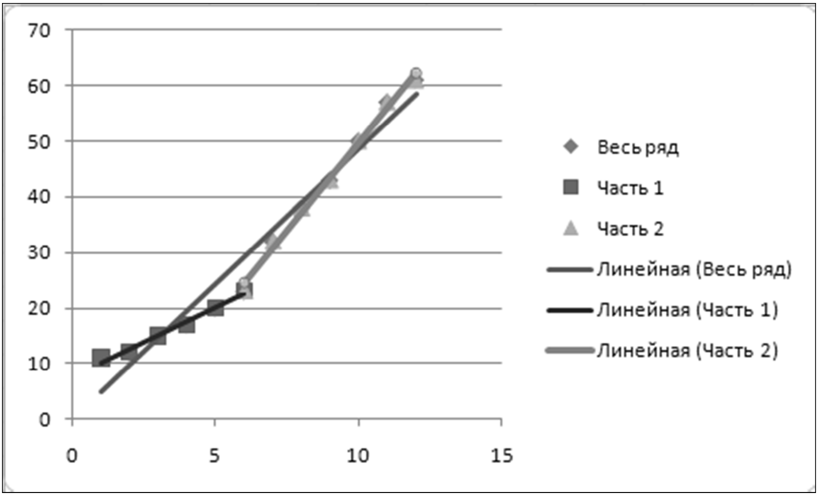
Пример. Качественный признак – **время года**,
значения: зима, весна, лето, осень.
Фиктивные переменные:
X_1=1, если зима, иначе X_1=0,
X_2=1, если весна, иначе X_2=0,
X_3=1, если лето, иначе X_3=0.

Переменную $\mathbf{X_4}$
аналогичным образом
вводить бессмысленно
(и нельзя!), так как

5.5. Тест Чоу (Chow).

В некоторых случаях можно наблюдать изменения тренда, вызванные различными факторами.

В таком случае уравнения тренда на разных участках будут различаться



В ситуации на рисунке возникает альтернатива – описывать тренд одной синей линией или совокупностью из двух линий.

Другими словами, выборка состоит из нескольких частей и требуется выяснить, следует ли их объединить в одну, или рассматривать каждую подвыборку отдельно.

Обозначим регрессию по первой выборке (объёма n_1) через A, по второй (объёма n_2) – через B, а объединённую (объёма $n=n_1+n_2$)– через R. Ясно, что

$$\mathbf{ESS_A + ESS_B \le ESS_R.}$$

Если левая часть неравенства близка к правой, то это говорит о том, что регрессия по объединённой выборке почти так же хороша, как “составная” их двух регрессий по двум выборкам. В этом случае объединение выборок в одну допустимо. Проверка этого составляет суть ***критерия Чоу***.

1) Составляется статистика

$$F = \frac{\left(ESS_R - ESS_A - ESS_B\right)/(m+1)}{\left(ESS_A + ESS_B\right)/(n-2m-2)}$$

2) Если вычисленное значение оказывается меньше $F_{1-\alpha}(m+1, n-2m-2)$, то объединение выборок возможно при данном α.

Замечание. Если одна из выборок слишком мала, то применяется другая формула. Пусть, например, $n_2 \le m+1$, тогда

$$\boxed{F = \frac{\left(ESS_R - ESS_A\right)/(n-n_1)}{ESS_A/(n_1-m-1)}}$$

Если вычисленное значение оказывается меньше $F_{1-\alpha}(n-n_1, n_1-m-1)$, то объединение выборок возможно при данном α.

Замечание. Другой способ проверки состоит в использовании *фиктивных переменных*.

Добавим в модель фиктивную переменную Z, равную 0 для первой выборки и 1 для второй, а также переменные ZX_1, ZX_2,…. Если <u>все</u> эти переменные окажутся *незначимыми*, то объединение выборок возможно.

Пример. Проверить возможность объединения выборок при α=0,1.

y	x_1	x_2
3	2	3
5	3	4
5	3	6
6	4	1
7	5	7

y	x_1	x_2
9	4	7
9	6	8
10	6	9
12	7	9

Решение. 1 способ (критерий Чоу).

Для первой выборки:

$$y = 0{,}86 + 1{,}26x_1 + 0{,}01x_2,\ ESS = 0{,}42$$

Для второй выборки:

$$y = 3 + 0{,}5x_1 + 0{,}5x_2,\ ESS = 2{,}5$$

Для объединённой выборки:

$$y = 0{,}11 + 1{,}32x_1 + 0{,}23x_2,\ ESS = 7{,}38$$

$$F = \frac{(7{,}38 - 0{,}42 - 2{,}5)/(2+1)}{(0{,}42 + 2{,}5)/(9 - 2\cdot 2 - 2)} = 1{,}527$$

$$F_{0,9}(2+1; 9 - 2\cdot 2 - 2) = F_{0,9}(3;3) = 5{,}4$$

Так как 1,527<5,4, то объединение возможно.

2 способ (фиктивная переменная). Добавляем новые переменные z, zx_1, zx_2.

Проверяем их значимость

$$
\begin{aligned}
y = \underset{(1,59)}{0,86} + \underset{(0,46)}{1,26x_1} + \underset{(0,22)}{0,01x_2} \\
+ \underset{(6,63)}{2,14z} + \underset{(1,13)}{0,76zx_1} + \underset{(0,38)}{0,49zx_2}
\end{aligned}
$$

y	x_1	x_2	z	zx_1	zx_2
3	2	3	0	0	0
5	3	4	0	0	0
5	3	6	0	0	0
6	4	1	0	0	0
7	5	7	0	0	0
9	4	7	1	4	7
9	6	8	1	6	8
10	6	9	1	6	9
12	7	9	1	7	9

Все добавленные переменные незначимы, следовательно, объединение выборок возможно.

5.6. Проверка гипотезы о линейной связи коэффициентов

Пусть рассматривается линейная модель:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m + \varepsilon, \quad (U)$$

Мы хотим проверить гипотезу о том, не удовлетворяют ли коэффициенты модели q линейным ограничениям:

$$H_0 : H\beta = r, \quad H : q \times (m+1)$$

Например, три ограничения (q=3):

$$H_0 : \begin{cases} \beta_1 = 2\beta_2 \\ \beta_3 = 0 \\ \beta_4 = 5 \end{cases}$$

Чтобы проверить такую гипотезу, нужно

1) Вычислить $ESS=ESS_U$ для исходной регрессии.
2) Вычислить $ESS=ESS_R$ для регрессии, в которой выполнены условия гипотезы H_0.
3) Вычислить значение F-статистики

$$\boxed{F = \frac{\left(ESS_R - ESS_U\right)/q}{ESS_U/(n-m-1)}}$$

Гипотеза H_0 принимается, если **вычисленное значение F-статистики оказывается** ***меньше*** $\mathbf{F_{1-\alpha}(q\ ;\ n-m-1)}$

5.7. RESET-тест Рамсея (Ramsey)

RESET-тест Рамсея (сокр. от англ. *Regression Equation Specification Error Test*) отвечает на вопрос, надо ли включать

в регрессию дополнительные (в частности, нелинейные) члены.

Пусть рассматривается линейная модель:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m + \varepsilon, \quad (*)$$

Мы хотим выяснить, не следует ли вместо (*) рассмотреть более сложную функциональную модель, включающую степени x.

Алгоритм теста Рамсея

1) Найти оценки параметров регрессии (*).

2) Вычислить расчётные значения y^ зависимой переменной y.

3) Оценить регрессию

$$y=\beta_0+\beta_1 x_1+\ldots+\beta_m x_m+\alpha_2\hat{y}^2+\ldots+\alpha_k\hat{y}^k+\varepsilon, \quad (**)$$

4) Проверить гипотезу H_0: *все* $\alpha_j=0$
(то есть что дополнительные слагаемые в (**) не нужны).
Для этого вычислить

$$\boxed{F=\frac{\left(ESS_{*}-ESS_{**}\right)/(k-1)}{ESS_{**}/(n-m-k)}}$$

Гипотеза H_0 принимается, если **вычисленное значение F-статистики оказывается *<u>меньше</u>* $F_{1-\alpha}$(k–1 ; n–m–1)**

Пример. Заработная плата в Нидерландах.

Имеется 150 наблюдений, 75 мужчин и 75 женщин, работавших на полную ставку.
W - заработная плата (гульденов в час),
SEX - пол (1– мужчины, 2 – женщины),
EDU – уровень образования (1– нач. школа,… 5 – университет) ,
AGE возраст .

	W	SEX	EDU	AGE
1	10,44	1	1	19
2	13,52	1	1	20
3	19,12	1	1	21
4	20,28	1	1	25
5	14,63	1	1	26
...	...	...	...	...
22	20,32	1	2	29
23	15,96	1	2	32
24	17,07	1	2	37
25	13,04	1	2	39
...	...	...	...	...
148	26,09	2	5	27
149	24,95	2	5	37
150	41,05	2	5	52

Оценим линейную регрессию

$$W = 3{,}51 - 3{,}55 \cdot SEX + 3{,}24 \cdot EDU + 0{,}44 \cdot AGE,$$

$$ESS = 7631$$

Проверим при помощи **теста Рамсея** правильность функциональной формы. Для этого вычислим столбец W^ расчётных значений и добавим к исходным регрессорам его вторую степень (ограничимся второй степенью).

Получим уравнение регрессии

$$W = 12{,}09 + 0{,}29 \cdot SEX - 0{,}49 \cdot EDU - 0{,}09 \cdot AGE + 0{,}02 \cdot \hat{W}^2,$$

$$ESS = 7154$$

Вычислим F-статистику

$$F = \frac{(ESS_{*} - ESS_{**})/(k-1)}{ESS_{**}/(n-m-k)} = \frac{(7631-7154)/(2-1)}{7154/(150-3-2)} = 9{,}67$$

Далее, на уровне значимости α=0,1

$$F_{1-\alpha}(k-1;\, n-m-k) = F_{0,9}(1;\,145) = 2{,}74 < 9{,}67$$

В уравнение

$$W = \beta_0 + \beta_1 \cdot SEX + \beta_2 \cdot EDU + \beta_3 \cdot AGE + \varepsilon$$

надо попробовать включить нелинейные члены, например, AGE^2 или другие переменные.

6. Некоторые дополнительные вопросы

6.1. Коэффициент эластичности

Зависимость изменения значения переменной y=f(x) от небольшого изменения значения переменной x выражается производной по x:

$$y' = \frac{dy}{dx} \approx \frac{\Delta y}{\Delta x}$$

Если, например, мы выражаем объём спроса Q через цену P товара (и, возможно, какие-то другие параметры), то (частная) производная

$$Q_p^{/} \approx \frac{\Delta Q}{\Delta P},$$

где
ΔP - изменение цены,
ΔQ - вызванное им изменение объема спроса,

покажет, ***на сколько*** единиц изменится спрос в расчете ***на единичное изменение цены*** при данных значениях прочих параметров.

Пусть повышение цены за 1 кг картофеля на 1 руб. снижает годовой объем спроса на 20 кг, т. е. ΔP = 0,1 руб./кг, ΔQ = −20 кг/год (знак "минус" соответствует уменьшению).
Считая эти изменения малыми, можно приближенно оценить производную:

$$Q_p' \approx -\frac{20}{1} = -20\,\frac{\text{кг}^2}{\text{руб.}\cdot\text{год}}$$

Допустим, аналогичным образом мы установили, что для билетов в кино

$$Q_p' \approx -0{,}5\,\frac{\textit{шт}^2}{\text{руб.}\cdot\text{год}}$$

Какая из этих величин больше? Вопрос бессмыслен, и не только из-за того, что величины, измеренные в различных единицах, несопоставимы.

Эти трудности можно преодолеть, если в качестве основного показателя реакции спроса на изменение цены использовать не **производную**, а
эластичность спроса по цене − предел отношения ***относительного приращения*** объема $\delta Q = \Delta Q/Q$ ***к относительному приращению цены*** $\delta P = \Delta P / P$ при условии, что последнее стремится к нулю

В общем случае эластичность вычисляется по формуле

$$\boxed{E_x(y) = \frac{dy/y}{dx/x} = \frac{x}{y}\cdot y'}$$

Эластичность - безразмерная величина;
ее использование снимает сложности, связанные с единицами и масштабами рассматриваемых величин.

Пример (степенная зависимость). Вычислим эластичность для степенного вида зависимости.

$$y = a \cdot x^b$$

$$E_x(y) = y' \cdot \frac{x}{y} = a \cdot b \cdot x^{b-1} \cdot \frac{x}{a \cdot x^b} = b$$

То есть для степенного вида зависимости эластичность постоянна. При изменении x на 1 % величина y изменяется на b %.

Пример (линейная зависимость). Вычислим эластичность для линейного вида зависимости.

$$y = a + bx$$

$$E_x(y) = y' \cdot \frac{x}{y} = b \cdot \frac{x}{a+bx} = \frac{bx}{a+bx}$$

6.2. *Парная линейная регрессия*

В частном случае, когда <u>m=1</u> (***одна*** независимая переменная x) и ищется прямая линия $y=\beta_0+\beta_1 x$, получаются легко запоминающиеся формулы:

$$\boxed{\begin{array}{l} y = \beta_0 + \beta_1 x \\ \hat{\beta}_1 = \dfrac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \left(\overline{x}\right)^2}, \quad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \end{array}}$$

6.3. *Нелинейные формы зависимости*

Использование лишь линейных зависимостей для описания экономических взаимосвязей часто оказывается недостаточным. Необходимо использовать и нелинейные соотношения.

Например, *производственная функция Кобба-Дугласа*:

$$Y = A \cdot K^{\alpha} \cdot L^{\beta}$$

где Y– выпуск, K– затраты капитала, L– затраты труда, A, α, β, – параметры.

В некоторых случаях ***нелинейные*** зависимости путем замены переменных можно преобразовать к ***линейному*** виду.

Пример (гиперболическая зависимость).
По исходным данным оценить коэффициенты в формуле

$$y = a + \frac{b}{x} + \varepsilon$$

y	x
5	2
3	3
3	5
2	5
2	7
1	8

Решение. Заменой z =1/x зависимость приводится к линейному виду y=a+bz+ε.

$$y = 0{,}43 + 8{,}94z + \varepsilon$$

$$Ответ: y = 0{,}43 + \frac{8{,}94}{x} + \varepsilon$$

y	$z = 1/x$
5	0,5
3	0,3333
3	0,2

Пример (полиномиальная зависимость).
По исходным данным оценить коэффициенты в формуле

$$y = a + bx + cx^2 + \varepsilon$$

Решение. Заменой z =x^2 зависимость приводится к линейному виду y=a+bx+cz+ε.

y	x
5	2
3	3
3	5
2	5
2	7
1	8

$$y = 6{,}54 - 1{,}11x + 0{,}06z + \varepsilon$$

$$Ответ: y = 6{,}54 - 1{,}11x + 0{,}06x^2 + \varepsilon$$

y	x	$z = x^2$
5	2	4
3	3	9
3	5	25
2	5	25

Пример (логистическая кривая).
По исходным данным оценить коэффициенты в формуле

$$y = \frac{1}{a + be^{-x} + \varepsilon}$$

Решение. Заменой z =1/y, t=e^{-x} зависимость приводится к линейному виду z=a+bt+ε.

y	x
1,61	1
1,78	2
1,92	2
1,90	3
2,06	4

$z = 0{,}49 + 0{,}35t + \varepsilon$

$$\text{Ответ}: y = \frac{1}{0{,}49 + 0{,}35e^{-x} + \varepsilon}$$

$z = 1/y$	$t = e^{-x}$
0,62	0,37
0,56	0,14
0,52	0,14
0,52	0,05
0,49	0,02

Пример (степенная зависимость). К линейному виду можно привести *степенную* зависимость:

$$y = a \cdot x_1^{b_1} x_2^{b_2} \cdot \ldots \cdot x_m^{b_m} \cdot (1+\varepsilon)$$

где a, b_1, …, b_m – параметры, а ε – случайный множитель. Прологарифмируем это соотношение

$$\ln y = \ln a + b_1 \ln x_1 + \ldots + b_m \ln x_m + \ln(1+\varepsilon)$$

Если теперь обозначить

$$\tilde{y} = \ln y; \beta_0 = \ln a; \tilde{x}_i = \ln x_i; \tilde{\varepsilon} = \ln(1+\varepsilon)$$

то оно примет вид ***линейной*** регрессионной модели

$$\tilde{y} = \beta_0 + \beta_1 \tilde{x}_1 + \ldots + \tilde{\varepsilon}$$

Пример. По исходным данным оценить коэффициенты А,α,β в формуле Кобба-Дугласа

$$Y = A \cdot K^{\alpha} \cdot L^{\beta} \cdot (1+\varepsilon)$$

Y	K	L
34	2	4
53	3	6
81	5	7
81	5	7
113	7	9
97	9	4

Решение. Логарифмированием зависимость приводится к линейному виду

lnY=lnA+αlnK+βlnL+ln(1+ε).

Ответ.

$$Y = 12 \cdot K^{0,69} \cdot L^{0,41} \cdot (1+\varepsilon)$$

$\ln Y$	$\ln K$	$\ln L$
3,53	0,69	1,39
3,96	1,1	1,79
4,39	1,61	1,95
4,39	1,61	1,95

6.4. Нелинейные модели, не приводимые к линейному виду

Всё же далеко не всякую зависимость заменой переменных можно привести к линейному виду.

Например, если искать логистическую зависимость в виде

$$y = \frac{1}{a + be^{-x}} + \varepsilon$$

вместо

$$y = \frac{1}{a + be^{-x} + \varepsilon}$$

то никакими переобозначениями **к линейному виду её не привести**.

Другие примеры:

$$y = ax^b + \varepsilon, \quad y = a + bx^c + \varepsilon, \quad y = ae^{bx} + \varepsilon, ...$$

В общем случае уравнение нелинейной регрессии с аддитивным случайным членом ε имеет вид

$$y = f(x_1, ..., x_m, \beta_0, ..., \beta_k) + \varepsilon,$$

Для нахождения оценок этих параметров можно использовать, как и в линейном случае, **метод наименьших квадратов**

$$ESS = \sum_1^n e_i^{\,2} = \sum_1^n \left(y_i - \hat{y}_i\right)^2 \to \min$$

ESS является функцией *k+1* переменных β_0, …, β_k . Задача минимизации решается при помощи итеративных методов оптимизации, например, градиентным методом.

Пример. По исходным данным оценить коэффициенты a, b в формуле

$$y = a \cdot e^{bx} + \varepsilon$$

y	x
2	1
9	2
47	3

Решение. Задача не сводится к линейной. Выпишем формулу для ESS:

$$ESS = \sum_1^n \left(y_i - \hat{y}_i\right)^2 = \left(2 - ae^b\right)^2 + \left(9 - ae^{2b}\right)^2 + \left(47 - ae^{3b}\right)^2.$$

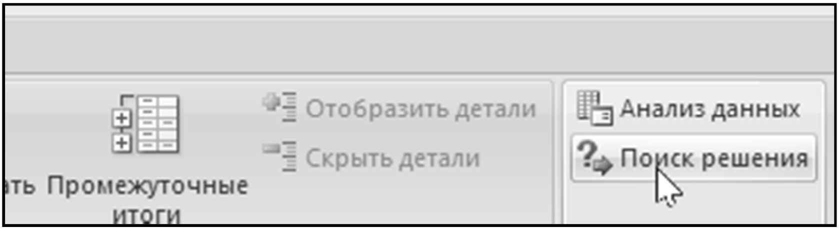
Требуется тем или иным способом найти **минимум** этой функции. Используем надстройку **Поиск решения** из **MS Excel.**

Внесём данные и формулы в таблицу:

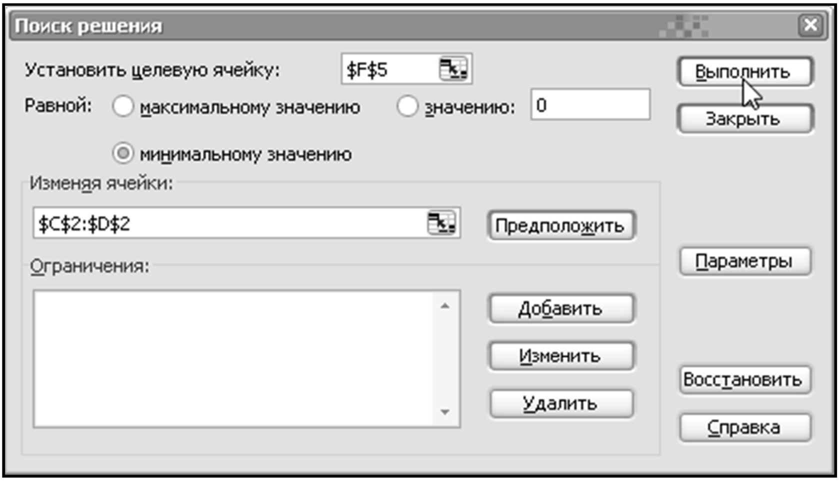
	A	B	C	D	E	F
1	Y	X	a	b	F(x)	
2	2	1	0	0	=$C$2*EXP($D$2*B2)	=(E2-A2)^2
3	9	2			=$C$2*EXP($D$2*B3)	=(E3-A3)^2
4	47	3			=$C$2*EXP($D$2*B4)	=(E4-A4)^2
5					**ESS**	**=СУММ(F2:F4)**

Начальные значения a и b положены равными нулю.

Вызовем надстройку **Поиск решения**



и введём данные в диалоге



Получим результат:

	A	B	C	D	E	F
1	Y	X	a	b	F(x)	
2	2	1	0,341	1,642	1,76	0,06
3	9	2			9,09	0,01
4	47	3			46,99	0,00
5					**ESS**	**0,07**

Таким образом, оценки параметров модели: a=0,341, b=1,642,

$$y = 0{,}341 \cdot e^{1{,}642x} + \varepsilon$$

7. Нарушения допущений классической линейной модели

7.1. Мультиколлинеарность

7.1.1. Определение

Одним из условий **классической регрессионной модели** является **предположение (C) о линейной независимости столбцов матрицы X**.

Нарушение этого условия – **линейная зависимость двух или более объясняющих переменных** – называется (точной) ***мультиколлинеарностью***.

На практике более реальна ситуация, когда между объясняющими переменными существует не точная линейная, а сильная корреляционная зависимость.

Эту ситуацию мы и будем называть ***мультиколлинеарностью***.

Например, если в состав объясняющих переменных X входят и доход, и потребление, то обе эти переменные будут сильно коррелированными.

7.1.2. Последствия мультиколлинеарности

- ❖ Оценки коэффициентов регрессии по-прежнему остаются наилучшими линейными несмещенными оценками.
- ❖ Стандартные ошибки коэффициентов при коррелированных регрессорах увеличиваются, так что коэффициенты могут стать незначимыми

7.1.3. Признаки мультиколлинеарности

- ❖ Небольшое изменение исходных данных приводит к существенному изменению оценок параметров модели.
- ❖ Оценки коэффициентов имеют малую значимость, в то время как модель в целом является значимой.
- ❖ Оценки коэффициентов имеют неправильные с точки зрения теории знаки.

Пример. Рассмотрим следующую выборку.

Y	X_1	X_2
2,0	1,1	6,9
3,0	2	6,8
3,0	3,5	5,4
4,0	4	4,3
3,0	2,7	5,5
5,0	4,1	4,6
2,0	5	3,2
8,0	7	1,8
6,0	8	0,6
5,0	6,4	2,5

Матрица корреляций имеет вид

	Y	X_1	X_2
Y	1		
X_1	0,78	1	
X_2	-0,7	-0,99	1

Видна сильная линейная зависимость между X_1 и X_2.
(На самом деле $X_2=8–0,5X_1+\varepsilon$).

Оценим регрессию

$$Y=-12,54+2,22X_1+1,65X_2$$

F	*Значимость F*
6,961	0,022

	Коэффициенты	*P-Значение*
Y-пересечен	-12,54	0,33
X_1	2,22	0,14
X_2	1,65	0,28

Регрессия значима при $\alpha>0,022$, т.е., скажем, при $\alpha=0,1$ или 0,05

Коэффициенты незначимы даже при $\alpha=0,1$

Изменим только одно значение X_2 с 6,9 на 7,1:

Y	X_1	X_2
2,0	1,1	7,1
3,0	2	6,8
3,0	3,5	5,4

Получим регрессию

$$Y=-13,61+2,33X_1+1,77X_2$$

Видим, что совсем небольшое изменение исходных данных привело к заметному изменению коэффициентов.

7.1.4. Способы обнаружения

На практике о наличии МК первоначально судят по матрице парных корреляций между регрессорами. Большое по модулю значение коэффициента корреляции между двумя x говорит о наличии МК.
Близкое к нулю значение определителя этой матрицы также говорит о возможной МК.
Ещё один часто используемый для обнаружения МК показатель называется VIF ("фактор инфляции вариации").
Пусть исследуется регрессия

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m + \varepsilon$$

Построим регрессию регрессора X_i <u>на остальные X</u> и вычислим соответствующий R^2:
Тогда

$$\boxed{VIF(X_i) = \frac{1}{1 - R_i^2}}$$

Если какие-то VIF оказываются больше 10, то это служит признаком МК**.**

В предыдущем примере $VIF(X_1)=VIF(X_2)=49,3$.

7.1.5. Методы устранения

Все кардинальные методы борьбы с МК имеют свои недостатки.

- Если задачей исследования является ***прогноз*** будущих значений зависимой переменной, то МК не является серьёзной проблемой, так как она не ухудшает значимость уравнения**.**
- Естественным методом устранения мультиколлинеарности является ***исключение отдельных переменных*** из модели.

Но при этом могут возникнуть новые трудности. Например, не всегда ясно, какие переменные являются “лишними”.
Удаление переменных может изменить содержательный смысл модели.
Кроме того, коэффициент при оставшейся переменной получает смещение, т.к. теперь измеряет не только свое «чистое» влияние на Y, но и влияние на Y всех отброшенных переменных

- Иногда можно **увеличить объём выборки или качество исходных данных**.

7.2. Гетероскедастичность

7.2.1. Определение

Условие **(D)** теоремы Гаусса-Маркова требует постоянства дисперсии случайного члена ε_i во всех наблюдениях:

$$D\varepsilon_i = \sigma^2 \text{ не зависит от } i$$

Это называется ***гомоскедастичность.***

Если же дисперсия случайного члена *меняется* от наблюдения к наблюдению, то мы имеем дело с ***гетероскедастичностью.***

7.2.2. Последствия гетероскедастичности

- ❖ Оценки коэффициентов регрессии несмещённые.
- ❖ Стандартные ошибки коэффициентов являются заниженными.

Это приводит к завышению t-статистик и даёт неправильное (завышенное) представление о точности оценок.

7.2.3. Способы обнаружения

Существуют различные способы выявления гетероскедастичности.

7.2.3.1. Тест Уайта

- Для исходной модели (объём выборки = n) строится линейная регрессия и
  находятся остатки e_i.
- Строится регрессия квадратов этих остатков на все исходные переменные, их квадраты, попарные произведения и константу (пусть всего получится N переменных, не считая константы).
- Если для второй регрессии $n \cdot R^2$ больше $\chi^2_{1-\alpha}(N)$, то гетероскедастичность есть.

Пример. Проверить на гетероскедастичность при α=0,1, используя тест Уайта

y	x_1	x_2
3	2	3
5	3	4
5	3	6
6	4	1
7	5	7
9	4	7
9	6	8

Решение. 1) Построим регрессию и найдём расчётные значения и остатки e_i.

$$y = 0{,}77 + 1{,}15x_1 + 0{,}22x_2$$

$\hat{y}$	$e = y - \hat{y}$
3,72	$-0{,}72$
4,87	0,13
5,52	$-0{,}52$
5,59	0,41
8,03	$-1{,}03$
6,99	2,12
9,4	$-0{,}4$

2) Построим регрессию по выборке, описанной в тесте Уайта:

e_i	e_i^2	x_1	x_2	x_1^2	x_2^2	x_1x_2
$-0{,}72$	0,52	2	3	4	9	6
0,13	0,017	3	3	9	9	9
$-0{,}52$	0,27	3	6	9	36	18
0,41	0,17	4	1	16	1	4
$-1{,}03$	1,06	5	7	25	49	35
2,12	4,5	4	7	16	49	28
$-0{,}4$	0,16	6	8	36	64	48

3) Для этой регрессии вычислим:

$$nR^2 = 7 \cdot 0{,}4835 = 3{,}3845,$$

$$\chi^2_{1-\alpha}(N) = \chi^2_{0,9}(5) = 9{,}236$$

3,3845<9,236, так что гетероскедастичности нет.

7.2.3.2. Тест Голдфелда-Квандта

В этом тесте предполагается, что стандартное отклонение σ ***пропорционально значению некоторой независимой переменной*** x_i.

1) Наблюдения упорядочиваются по возрастанию значений этой переменной x_i.

2) Берутся первые n' и последние n' наблюдений, для них строятся регрессии, и оцениваются ESS_1 и ESS_2 соответственно.
Число n' можно взять примерно равным $n/3$.

3) Составляется статистика

$$F = \frac{ESS_2}{ESS_1}$$

4) Значение этой статистики, большее

$$F_{1-\alpha}(n'-m-1,\ n'-m-1)$$

где m – число объясняющих переменных в исходном уравнении регрессии,
означает **наличие гетероскедастичности**

Пример. Проверить на гетероскедастичность по переменной x_1 при α=0,1, используя тест Голдфелда-Квандта.

y	x_1	x_2
3	2	3
5	3	4
5	3	6
6	6	1
7	5	7
9	4	7
9	6	8
11	6	7
18	8	4
32	9	5
68	12	7
78	15	9

Решение. 1) Объём выборки=12, так что выберем $n'=12/3=4$. Отсортируем выборку по возрастанию x_1 и построим две регрессии (по первым и последним четырём строкам для x_1).

$$y = -3{,}37 + 2{,}79x_1 + 0{,}11x_2, \quad ESS_1 = 0{,}95$$

$$y = -0{,}67 - 13{,}67x_1 + 32x_2, \quad ESS_2 = 112{,}7$$

2)

$$F = \frac{ESS_2}{ESS_1} = \frac{112{,}7}{0{,}95} \approx 118$$

$$F_{1-\alpha}(n'-m-1; n'-m-1) = F_{0{,}9}(4-2-1;4-2-1) = 40,$$

118>40, так что гетероскедастичность в виде пропорциональности σ и x_1 есть.

7.2.3.3. Тест Глейзера

- Оценивается регрессионная зависимость

$$|e_i| = a + bx_i^k + u_i$$

- Параметр k изменяется с некоторым шагом, и для каждого его значения строится регрессия.
- *Статистическая значимость* уравнения при каком-либо значении означает *наличие гетероскедастичности*.

7.3. Автокорреляция

7.3.1. Определение

Условие (E) теоремы Гаусса-Маркова требует отсутствия корреляции между значениями случайного фактора ε_i во всех наблюдениях:

$$\text{cov}(\varepsilon_i \varepsilon_j)=0 \text{ при } i \neq j$$

При нарушении этого условия, т.е. при наличии связи между случайными членами для разных наблюдений, возникает явление *автокорреляции*.

Обычно автокорреляция рассматривается при изучении ***временных рядов*** (т.е. когда номер наблюдения соответствует ***моменту времени***).

Пример. *Авторегрессия первого порядка AR(1).*

$$\boxed{\varepsilon_i = \rho \cdot \varepsilon_{i-1} + u_i, \quad -1 < \rho < 1}$$

ε – случайный член уравнения регрессии,
ρ – коэффициент авторегрессии,
u – случайный член, не подверженный автокорреляции
В этом случае ε в данном наблюдении прямо связан лишь с ε в предыдущем наблюдении.

Если ρ>**0,** то автокорреляция ***положительная,***
если ρ<**0,** то ***отрицательная.***

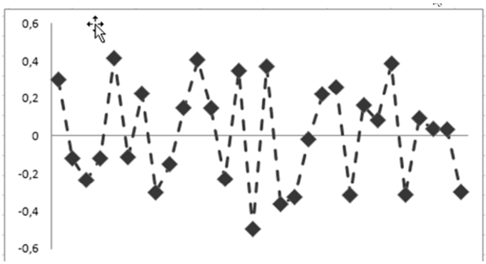
Замечание. Коэффициент ρ в AR(1) равен *коэффициенту корреляции соседних значений* ε. Действительно:

$$r(\varepsilon_i, \varepsilon_{i-1}) = r(\rho \cdot \varepsilon_{i-1} + u_i, \varepsilon_{i-1}) = \rho \cdot r(\varepsilon_{i-1}, \varepsilon_{i-1}) = \rho$$

Пример. *Авторегрессия второго порядка AR(2).*

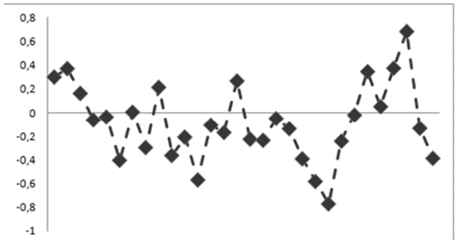
$$\boxed{\varepsilon_i = \rho_1 \cdot \varepsilon_{i-1} + \rho_2 \cdot \varepsilon_{i-2} + u_i, \quad -1 < \rho_{1,2} < 1}$$

Пример. Нулевая автокорреляция.



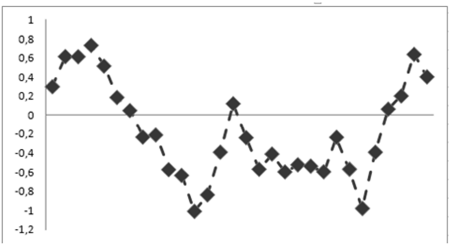
$$\boxed{\varepsilon_i = 0 \cdot \varepsilon_{i-1} + u_i}$$

Пример. Положительная автокорреляция.



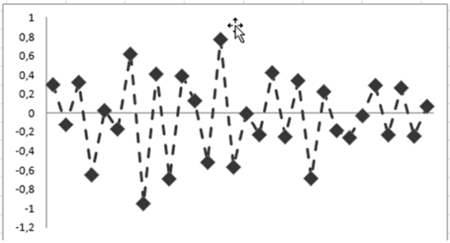
$$\boxed{\varepsilon_i = 0{,}5 \cdot \varepsilon_{i-1} + u_i}$$

Пример. Очень сильная положительная автокорреляция.



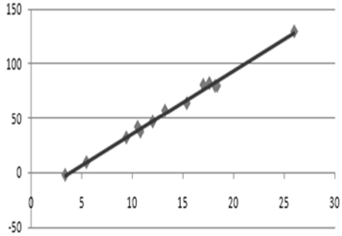
$$\boxed{\varepsilon_i = 0{,}9 \cdot \varepsilon_{i-1} + u_i}$$

Пример. Отрицательная автокорреляция.

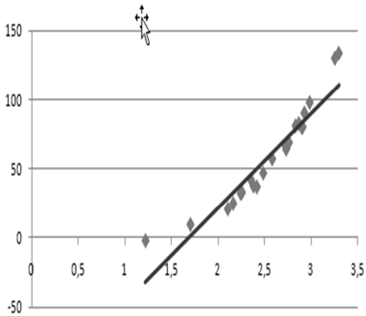


$$\boxed{\varepsilon_i = -0{,}8 \cdot \varepsilon_{i-1} + u_i}$$

Пример. Ложная автокорреляция, вызванная неправильной функциональной формой.



$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$



Те же данные, но другая формула

$$Y = \beta_0 + \beta_1 \ln X_1 + \varepsilon$$

7.3.2. Способы обнаружения

Существуют различные способы выявления автокорреляции.

7.3.2.1. Графический метод.

1) При помощи метода наименьших квадратов оценивается регрессия.

2) Вычисляются остатки e_i.

3) Строится график зависимости остатков e_i от номера наблюдения или от значений какой-то из переменных.

4) Далее анализируется внешний вид этого графика

7.3.2.2. Метод рядов

В этом методе рассматриваются знаки остатков e_i

"Ряд" это непрерывная последовательность одинаковых знаков.

При большом числе наблюдений применяется следующая процедура.

1) Рассчитываются статистики.

$$M = \frac{2n_+ n_-}{n_+ + n_-} + 1, \quad D = \frac{2n_+ n_-\left(2n_+ n_- - n_+ - n_-\right)}{\left(n_+ + n_-\right)^2\left(n_+ + n_- - 1\right)}$$

где n_+ - количество положительных остатков, n_- – количество отрицательных.

2) Если

$$M - u_{1-\alpha/2} \cdot \sqrt{D} < k < M + u_{1-\alpha/2} \cdot \sqrt{D}$$

где k – количество рядов, то принимается гипотеза об отсутствии автокорреляции.

7.3.2.3. Критерий Дарбина-Уотсона.

Этот метод применяется для обнаружения авторегрессии AR(1)

$$\boxed{\varepsilon_i = \rho \cdot \varepsilon_{i-1} + u_i}$$

В ***критерии Дарбина-Уотсона*** проверяется гипотеза ρ=**0** при альтернативной ρ≠**0**.

При этом используется статистика Дарбина-Уотсона**:**

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

Для больших выборок **DW ≈ 2·(1–ρ)**.
При значительной положительной автокорреляции

ρ ≈ 1 и DW ≈ 0.
При отрицательной автокорреляции

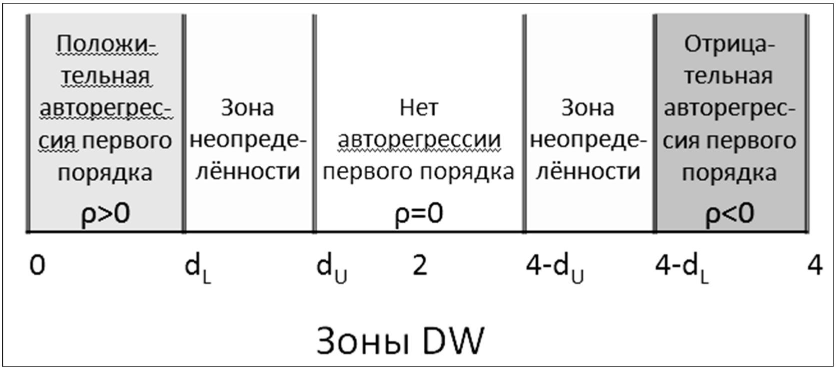
ρ ≈ –1 и DW ≈ 4.

При отсутствии автокорреляции

ρ ≈ 0 и DW ≈ 2.

Критические значения статистики Дарбина-Уотсона зависят не только от числа переменных, но и от значений, которые они принимают в выборке. Поэтому **невозможно составить таблицы критических значений DW**.

Но можно указать верхнюю d_U и нижнюю d_L границы для DW. Они определяются в зависимости от n и числа оцениваемых параметров m.



Границы d_L, d_U распределения Дарбина-Уотсона при α=0,05

m ↓	$n \rightarrow$	7	8	9	10	20	30
1	d_L	0,7	0,76	0,82	0,88	1,2	1,35
1	d_U	1,356	1,33	1,32	1,32	1,41	1,49
2	d_L	0,47	0,36	0,63	0,7	1,1	1,29
2	d_U	1,9	1,78	1,7	1,64	1,54	1,57
3	d_L		0,37	0,44	0,53	1	1,21
3	d_U		2,29	2,13	2,02	1,68	1,65
4	d_L			0,3	0,38	0,9	1,14
4	d_U			2,39	2,41	1,83	1,74

Пример. Построена регрессия с n=8, m=2. Используя тест Дарбина-Уотсона, проверить гипотезу о наличии авторегрессии первого порядка при α=0,05.

y	$\hat{y}$
3	3,09
5	4,68
5	5,79
6	6,13
7	8,43
9	7,39
9	10,02
11	9,46

Решение. Заполним таблицу

$e_i = y_i - \hat{y}_i$	e_{i-1}	$(e_i - e_{i-1})^2$	e_i^2
−0,09	–	–	0,0081
−0,32	−0,09	0,17	0,1024
−0,79	−0,32	1,23	0,6241
−0,13	−0,79	0,44	0,0169
−1,43	−0,13	1,69	2,0449
1,61	−1,43	9,24	2,5921
−1,02	1,61	6,92	1,0404
1,54	−1,02	6,55	2,3716
Σ		= 26,2	= 8,8

$$DW = \frac{26,2}{8,8} \approx 3$$

По таблице d_L=0,36, d_U=1,78. Число DW=3 попадает в зону $(4-d_U;4-d_L)$=(2,22;3,64). Поэтому критерий Дарбина-Уотсона в этом примере ответа не даёт.

7.3.3. Методы устранения автокорреляции

7.3.3.1. Исправление спецификации модели

Как уже отмечалось, важную роль играет правильный выбор функциональной формы зависимости. Например, при выборе линейной формы зависимости в ситуации, когда имеет место экспоненциальная, возникает положительная автокорреляция.

7.3.3.2. Авторегрессионное преобразование

Рассмотрим случай AR(1). Если

$$\boxed{\varepsilon_i = \rho \cdot \varepsilon_{i-1} + u_i,}$$

и коэффициент ρ известен, то для исходного уравнения регрессии

$$\boxed{y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,}$$

можно выполнить ***авторегрессионное преобразование***. Запишем это уравнение для предыдущего номера

$$y_{i-1} = \beta_0 + \beta_1 x_{i-1} + \varepsilon_{i-1},$$

умножим обе его части на ρ и вычтем из исходного уравнения

$$y_i - \rho y_{i-1} = \beta_0 \cdot (1-\rho) + \beta_1 \cdot (x_i - \rho \cdot x_{i-1}) + u_i$$

Если ввести новые переменные

$$\boxed{y_i^* = y_i - \rho \cdot y_{i-1}, \quad x_i^* = x_i - \rho \cdot x_{i-1}, \quad \beta_0^* = \beta_0 \cdot (1-\rho),}$$

то получим **уравнение без автокорреляции**

$$\boxed{y_i^* = \beta_0^* + \beta_1 x_i^* + u_i.}$$

Найдя оценки коэффициентов этого уравнения, мы тем самым сможем вычислить и оценки коэффициентов исходного.

Однако *проблема* заключается в том, что **величина ρ неизвестна.**

Поэтому используют разные методы получения **оценки для ρ,** такие как ***метод Хилдрета-Лу, метод Кохрейна-Оркатта,*** или просто $\rho \approx 1 - \dfrac{DW}{2}$.

7.4. *Стохастические объясняющие переменные*

До сих пор мы предполагали, что переменные x_1, x_2 ,…, x_m являются неслучайными, детерминированными (в отличие от y).

Но иногда их приходится считать случайными. Выясним, к чему это может привести.

7.4.1. Случай некоррелированности X и ε.

Пусть, как и раньше,

$$\boxed{Y = X \cdot \beta + \varepsilon}$$

Будем считать элементы матрицы X <u>случайными величинами</u>. Потребуем, чтобы были выполнены условия

(C_1) Ранг матрицы X равен m+1 с вероятностью единица
(D_1) $M(\varepsilon \mid X)=0$, $M(\varepsilon\varepsilon^T \mid X)=\sigma^2 E$.

Здесь M(…|X) – условные математические ожидания.

В этих предположениях

- ✓ МНК-оценки параметров регрессии являются несмещёнными.
- ✓ Среди всех линейных условно несмещённых оценок
- ✓ МНК-оценка обладает наименьшей условной дисперсией.

Если в каждом наблюдении значения объясняющих переменных (регрессоров) выбираются из одной и той же генеральной совокупности, а случайные факторы независимы, одинаково распределены и не зависят от X, то МНК-оценки состоятельны.

7.4.2. Коррелированность объясняющих переменных и случайного фактора.

Если же объясняющие переменные **X** и ошибки ε коррелированны (т.е. $M(\varepsilon|X)\neq 0$), то в общем случае

- ✓ МНК-оценки могут быть смещёнными и несостоятельными
- ✓ Содержательная интерпретация зависимостей ошибочна
- ✓ Рекомендации, сделанные на основе модели неверными.

8. Анализ временных рядов.

8.1. Определение

Временной ряд – это набор наблюдений какой-либо случайной величины, произведённых в последовательные моменты времени.

Это могут быть, например, цены на батон хлеба в соседнем магазине, курс обмена доллара на рубли в ближайшем обменном пункте или годовые объёмы добычи нефти странами ОПЕК.

8.2. Основные факторы, формирующие временной ряд

Обычно выделяют четыре типа факторов, под влиянием которых формируются значения временного ряда.

1. Долговременные − формируют общую тенденцию изменений в длительной перспективе. Функция, описывающая эту тенденцию, называется трендом.
2. Сезонные − формируют периодические, повторяющиеся в определённое время года колебания изменяемого показателя.
3. Циклические − формируют изменения показателя под влиянием действия циклов экономической, демографической и другой природы.
4. Случайные − эти факторы не поддаются учёту.

Не обязательно, чтобы присутствовали все четыре составляющие временного ряда.

Но будем считать, что случайные факторы присутствуют всегда (иначе неинтересно).

Выводы о том, участвуют или нет факторы данного типа в формировании значений y(t), могут базироваться как на анализе содержательной сущности задачи, так и на специальном статистическом анализе исследуемого временного ряда.

8.3. Аддитивная и мультипликативная модели временного ряда

Модель, в которой временной ряд представлен в виде суммы тренда, сезонной (или циклической) и случайной компонент, называется аддитивной.

$$\boxed{y_t = f(t) + \varphi(t) + \varepsilon_t}$$

Модель, в которой временной ряд представлен в виде произведения тренда, сезонной (или циклической) и случайной компонент, называется мультипликативной.

$$\boxed{y_t = f(t)\cdot\varphi(t)\cdot\varepsilon_t}$$

Мультипликативную модель часто можно свести к аддитивной логарифмированием.

8.4. Выделение основных компонентов временного ряда

8.4.1. Порядок анализа модели

Построение модели включает следующие шаги

1. Сглаживание исходного ряда.
2. Расчёт значений периодической компоненты.
3. Устранение периодической компоненты из исходного ряда и получение аналитического выражения для тренда.
4. Анализ остатков.

8.4.2. Сглаживание временного ряда методом скользящей средней

Чтобы уменьшить влияние случайных и циклических факторов, упростив нахождение тренда, используется сглаживание временного ряда *методом скользящей средней*. Сглаживание представляет собой вычисление взвешенного среднего значений, наблюдаемых в окрестности рассматриваемой точки. Оно определяется для каждого момента времени, *за исключение нескольких первых и нескольких последних точек.*

В простейшем случае вычисление производится следующим образом: берется k последовательных значений временного ряда (k нечётно), со средним, соответствующим текущему моменту времени t:

$$y_t,\ y_{t-1},\ y_{t+1},\ y_{t-2},\ y_{t+2},$$

и находится их среднее арифметическое:

$$\boxed{\tilde{y}_t = \frac{y_t + y_{t-1} + y_{t+1} + \ldots}{k}}$$

Затем точка *t* сдвигается вправо на один шаг, опять производится усреднение значений временного ряда и т.д.

Число k (т.е. временной интервал, по которому производится усреднение), называется ***окном***.

Замечание. Иногда требуется усреднять за период, равный чётному количеству шагов. Например, за год=12 месяцев. В этом случае применяют взвешенные средние. Например, в случае года:

$$\boxed{\tilde{y}_t = \frac{y_{t-6} + 2y_{t-5} + 2y_{t-4} + \ldots + 2y_t + 2y_{t+1} + \ldots + 2y_{t+5} + y_{t+6}}{24}}$$

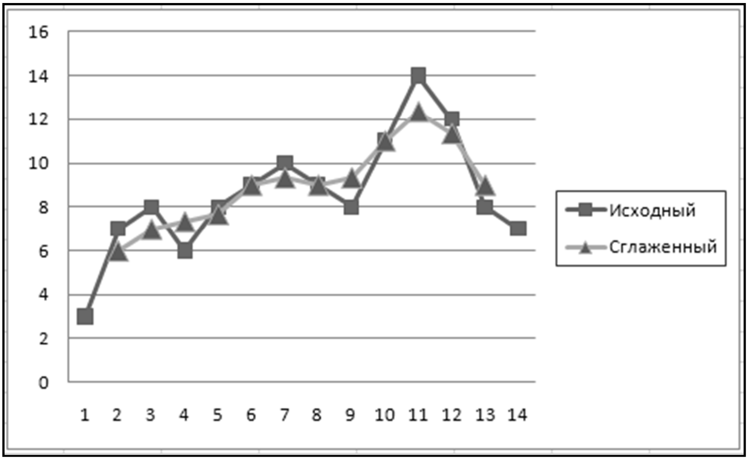
Пример. Сгладить временной ряд с окном 3.

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14
y_t	3	7	8	6	8	9	10	9	8	11	14	12	8	7

Решение. Вычислим поочерёдно

$$\tilde{y}_2 = \frac{y_1 + y_2 + y_3}{3} = \frac{3+7+8}{3} = 6, \quad \tilde{y}_3 = \frac{y_2 + y_3 + y_4}{3} = \frac{7+8+6}{3} = 7,\ldots$$

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14
y_t	3	7	8	6	8	9	10	9	8	11	14	12	8	7
$\tilde{y}_t$	–	6	7	7,3	7,7	9	9,3	9	9,3	11	12,3	11,3	9	–



8.4.3. Экспоненциальное сглаживание

Экспоненциально сглаженный ряд S_t определяется формулами

$$\boxed{s_1 = y_1, \quad s_t = s_{t-1} + \alpha(y_t - s_{t-1})}$$

где α − параметр сглаживания, $0 \le \alpha \le 1$. Чем меньше α, тем сильнее сглаживание

Можно дать следующую интерпретацию экспоненциальной средней:

если S_{t-1} — прогноз значения ряда y_t, то разность $y_t - S_{t-1}$ есть *погрешность* прогноза;

таким образом прогноз s_t для следующего момента времени t+1 учитывает ставшую известной в момент t ошибку прогноза.

Метод экспоненциального сглаживания часто применяется для краткосрочного прогнозирования. При этом основной задачей является выбор параметра сглаживания α.

Пример. Экспоненциально сгладить временной ряд с параметрами α=0,5 и α=0,3.

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14
y_t	3	7	8	6	8	9	10	9	8	11	14	12	8	7

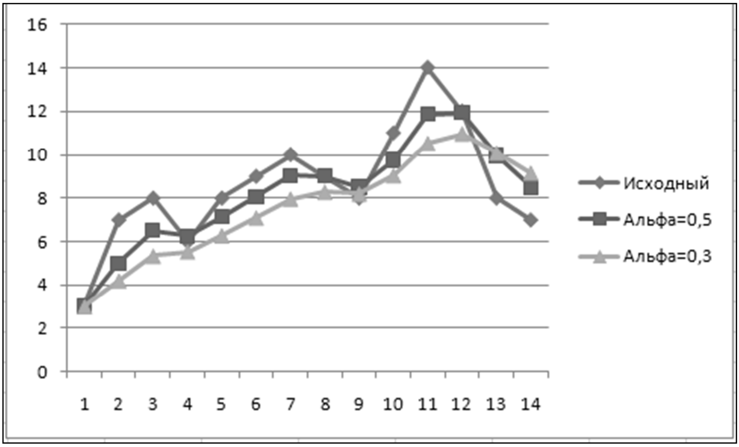
Решение. Пусть α=0,5.

$$s_1 = y_1 = 3, \quad s_2 = s_1 + 0{,}5 \cdot (y_2 - s_1) = 3 + 0{,}5 \cdot (7-3) = 5, \ldots$$

Пусть теперь α=0,3.

$$s_1 = y_1 = 3, \quad s_2 = s_1 + 0{,}3 \cdot (y_2 - s_1) = 3 + 0{,}3 \cdot (7-3) = 4{,}2, \ldots$$

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14
y_t	3	7	8	6	8	9	10	9	8	11	14	12	8	7
$\alpha = 0{,}5$	3	5	6,5	6,3	7,1	8,1	9	9	8,5	9,8	11,9	11,9	10	8,5
$\alpha = 0{,}3$	3	4,2	5,3	5,5	6,3	7,1	8	8,3	8,2	9	10,5	11	10,1	9,2



8.4.4. Выделение тренда временного ряда в аналитической форме

Трендом (или тенденцией) называют неслучайную медленно меняющуюся составляющую временного ряда, на которую могут накладываться циклические и случайные составляющие.

Замечание. Слова «*медленно меняющаяся составляющая» в* определении тренда носят относительный характер.

Например, медленное увеличение количества осадков в течение периода в сотню лет может быть понято как тренд.

Однако на самом деле рост осадков, характерный для этого столетия, может оказаться частью некоторого медленного колебательного процесса, происходящего в пределах нескольких тысячелетий.

При различении тренда и циклической компоненты невозможно полностью исключить из рассуждений элемент субъективности.

В качестве функции тренда чаще всего берут

- Гиперболический тренд: f(t)=a+b/t;
- Линейный тренд: f(t)=a+bt;
- Экспоненциальный тренд: $f(t)=a\cdot b^t$;
- Степенной тренд: $f(t)=a\cdot t^b$;
- Полиномиальный тренд: $f(t)=a+bt+ct^2+\ldots$;
- Логистическая кривая: $f(t)=1/(a+be^{-x})$

Выбор вида функции обычно производится с учётом выводов экономической теории и визуального анализа графика ряда.

Вычисление коэффициентов тренда часто производится по МНК.

Задача. Вычислить коэффициенты линейного тренда $f(t) = a + bt$ для временного ряда

t	1	2	3	4	5	6	7
y_t	82	87	99	104	107	121	118

Решение. При помощи обычного метода наименьших квадратов получаем. $\boxed{f(t) = 76{,}29 + 6{,}57 \cdot t}$

Задача. Для того же временного ряда вычислить коэффициенты степенного тренда $f(t)=a\cdot t^b$

Решение. Линеаризуем, взяв логарифм от обеих частей:

$$\ln f = \ln a + b \ln t$$

$\ln t$	0	0,7	1,1	1,4	1,6	1,8	1,9
$\ln y_t$	4,4	4,5	4,6	4,6	4,7	4,8	4,8

$$\begin{pmatrix} \ln a \\ b \end{pmatrix} = \begin{pmatrix} 4{,}37 \\ 0{,}21 \end{pmatrix} \qquad a = e^{4{,}37} = 79{,}13 \qquad \boxed{f(t) = 79{,}13 \cdot t^{0{,}21}}$$

Использование MS Excel для выделения тренда

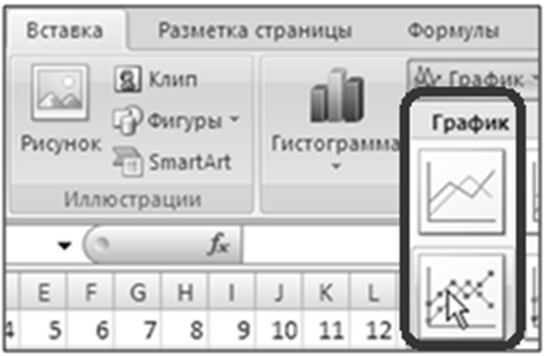
Пример. Найдём аналитическое выражение тренда в различных формах для временного ряда

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
y_t	3	5	4	4	5	4	7	6	11	9	11	14	18	21	22	29

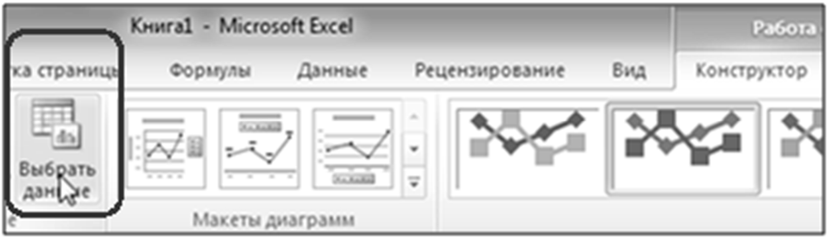
Решение. Занесём данные в таблицу MS Excel.

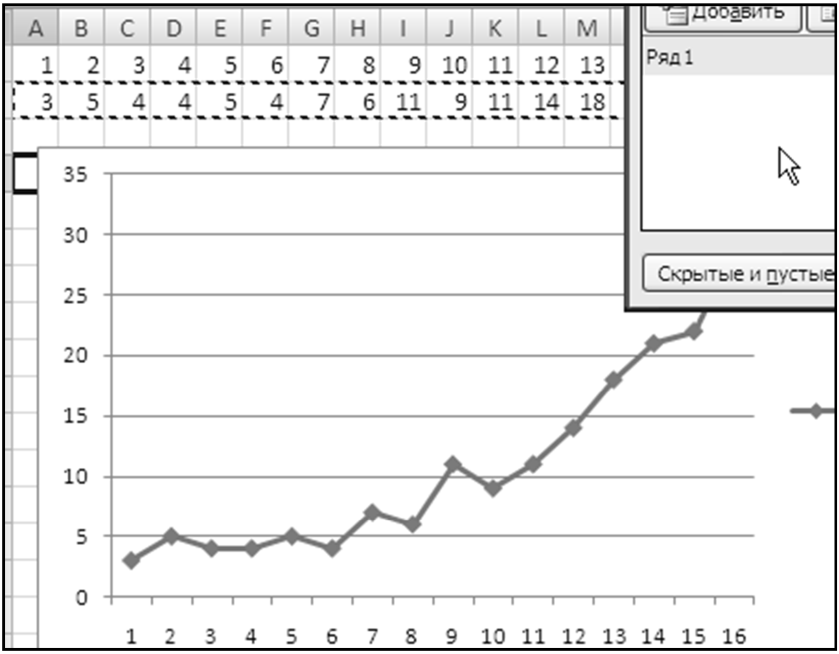
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
2	3	5	4	4	5	4	7	6	11	9	11	14	18	21	22	29

Вставка→ Диаграммы→График:

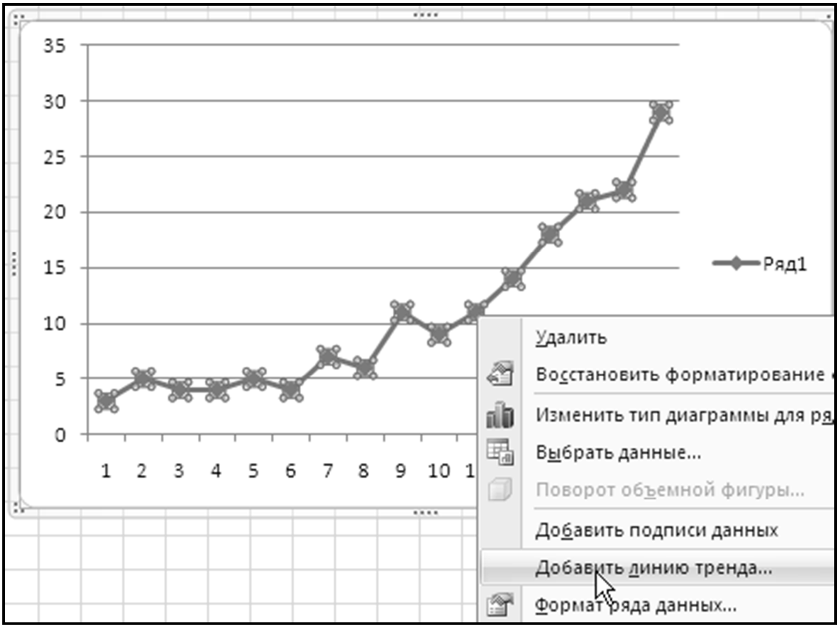


Работа с диаграммами → Конструктор→Выбрать данные

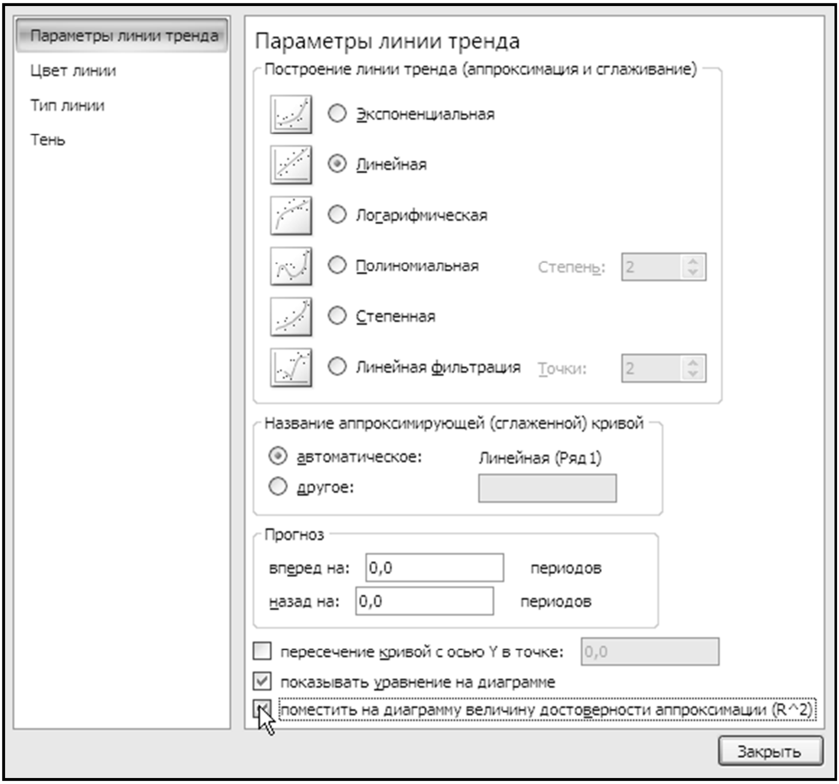




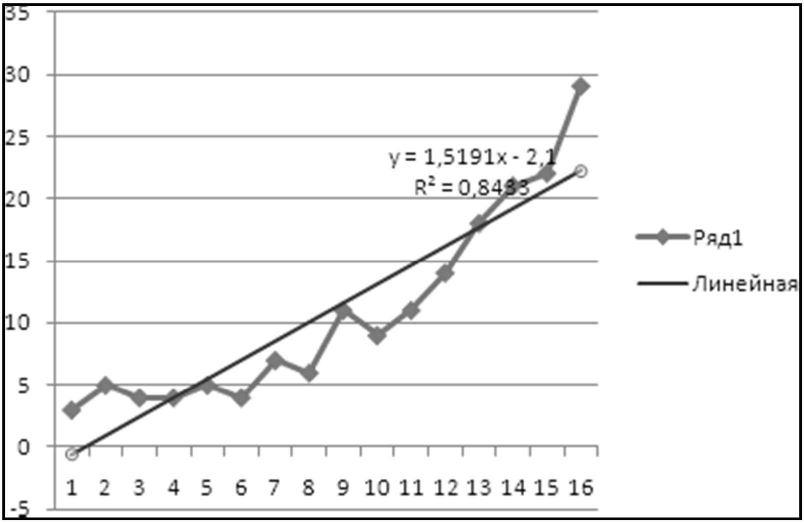
Выделим мышкой график и выберем пункт **меню Добавить линию тренда**



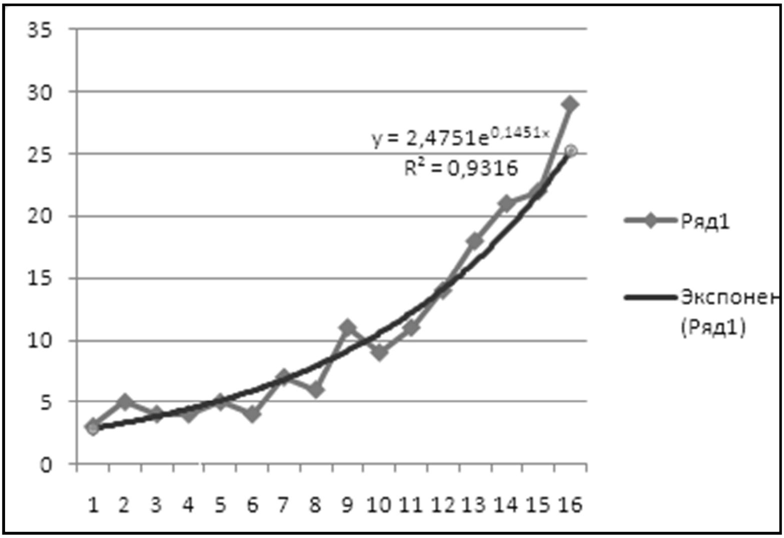
Появится **диалог**, в котором можно выбрать **тип тренда**



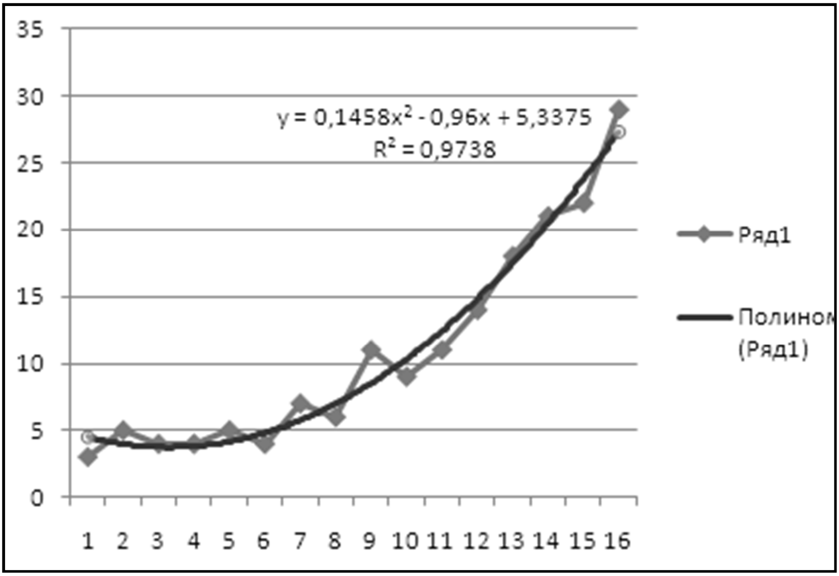
Выберем, например, линейный



Экспоненциальный:



Полиномиальный второй степени:



По критерию R^2 наилучшее значение у последнего вида тренда.

8.4.5. Выделение компонентов ряда при помощи фиктивных переменных в аддитивной модели

Компоненты временного ряда можно выделять и другим способом, используя фиктивные переменные.

Пример. Выделить линейный тренд и циклическую компоненту временного ряда, считая, что циклическая имеет период 3.

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
y_t	8	2	4	10	5	6	12	8	7	14	11	8	16	14	10

Решение. Для выделения циклической компоненты периода 3 заводим 2=3-1 фиктивные переменные D_1 и D_2.

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
y_t	8	2	4	10	5	6	12	8	7	14	11	8	16	14	10
D_1	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0
D_2	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0

Строим регрессию:

$$y_t = 0{,}6 + 0{,}711 \cdot \mathrm{t} + 6{,}422 \cdot \mathrm{D}_1 + 1{,}711 \cdot \mathrm{D}_2$$

Среднее арифметическое коэффициентов при D (считая ещё один коэффициент нулём):

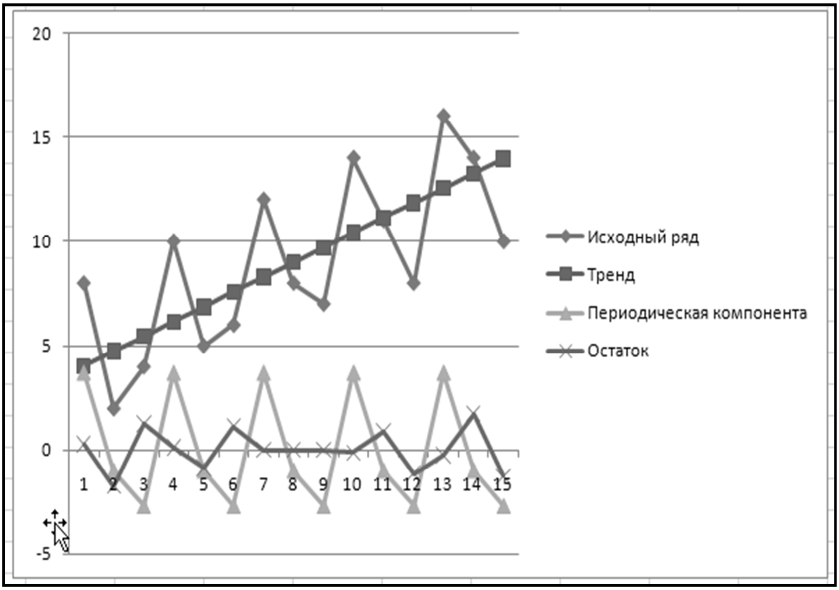
$$\frac{6{,}422 + 1{,}711 + 0}{3} = 2{,}711$$

Получаем циклическую компоненту

$\varphi(1)$	$\varphi(2)$	$\varphi(3)$
$6{,}422 - 2{,}711 = 3{,}711$	$1{,}711 - 2{,}711 = -1$	$0 - 2{,}711 = -2{,}711$

и линейный тренд:

$$f(t) = (0{,}6 + 2{,}711) + 0{,}711 \cdot \mathrm{t} = 3{,}311 + 0{,}711 \cdot \mathrm{t}$$



Замечание. ***Сезонная*** компонента имеет период 4, так что в этом случае заводятся 3=4–1 фиктивные переменные, примерно так:

t	1	2	3	4	5	6	7	8	9	10	11	12	...
y_t	...	...	...	...	...	...	...	...	...	...	...	...	...
D_1	1	0	0	0	1	0	0	0	1	0	0	0	...
D_2	0	1	0	0	0	1	0	0	0	1	0	0	...
D_3	0	0	1	0	0	0	1	0	0	0	1	0	...

И далее аналогично предыдущему примеру (только в среднем арифметическом деление на 4)

$$y_t = \beta_0 + \beta_1 \cdot \mathrm{t} + \beta_2 \cdot \mathrm{D}_1 + \beta_3 \cdot \mathrm{D}_2 + \beta_4 \cdot \mathrm{D}_3$$

8.4.6. Выделение компонентов временного ряда в случае мультипликативной модели

Пример. Выделить основные компоненты временного ряда, считая, что циклическая имеет период 4

t	1	2	3	4	5	6	7	8	9	...	16
y_t	72	100	90	64	70	92	80	58	62	...	30

Взятие логарифма превращает мультипликативный ряд в аддитивный:

$$y_t = f(t)\cdot\varphi(t)\cdot\varepsilon_t$$

$$\ln y_t = \ln f(t) + \ln\varphi(t) + \ln\varepsilon_t$$

Поэтому проще всего перейти от временного ряда к его логарифму, выделить компоненты, как в аддитивном, а затем вернуться к исходному.

t	1	2	3	4	5	6	...	16
$\ln y_t$	4,277	4,605	4,5	4,159	4,25	4,52	...	3,4

Выделяем периодическую компоненту и тренд, получаем:

$\ln\varphi(1)$	$\ln\varphi(2)$	$\ln\varphi(3)$	$\ln\varphi(4)$
$-0{,}045$	$0{,}058$	$0{,}1$	$-0{,}113$

$$\ln f(t) = 4{,}659 - 0{,}075\cdot t$$

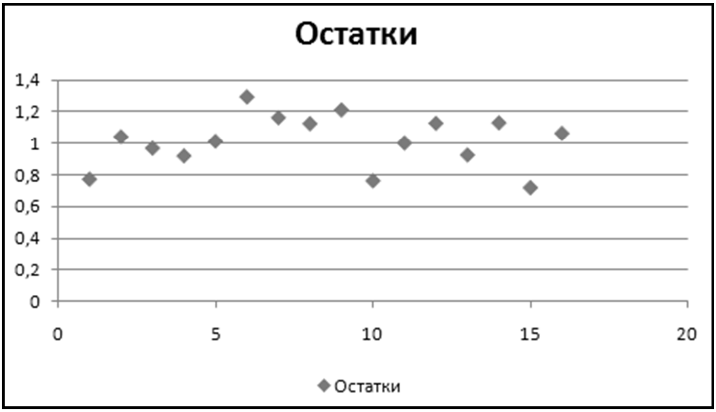
Отсюда

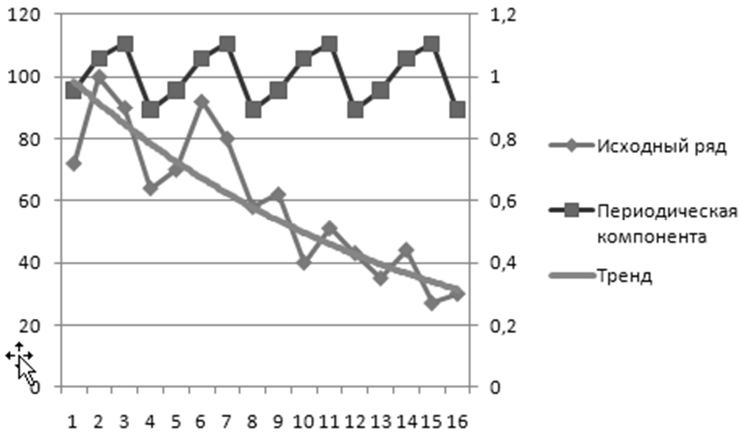
$\varphi(1)$	$\varphi(2)$	$\varphi(3)$	$\varphi(4)$
$e^{-0{,}045} = 0{,}956$	$e^{0{,}058} = 1{,}06$	$e^{0{,}1} = 1{,}106$	$e^{-0{,}113} = 0{,}893$

$$f(t) = e^{4,659-0,075\cdot t} = 105,497e^{-0,075\cdot t}$$

Остаток находится делением исходного ряда на φ и f

t	1	2	3	4	5	6	7	8	...	16
y_t	72	100	90	64	70	92	80	58	...	30
ε	0,77	1,04	0,967	0,918	1,012	1,293	1,162	1,124	...	1,062





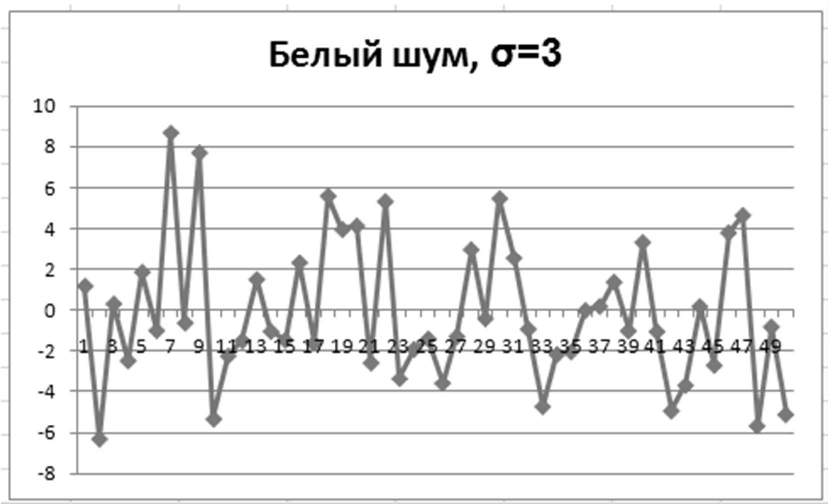
8.5. Стационарность

Остаток временного ряда, освобождённый от периодической составляющей и тренда, должен быть в каком-то смысле стационарным.

Ряд $\mathbf{y}_t$ называется ***стационарным в широком смысле***, если его среднее значение и cov($\mathbf{y}_{t}$, $\mathbf{y}_{t+k}$) (в частности, дисперсия) не зависят от t.

В частности, для стационарного ряда недопустимы тренд и циклическая компонента!

8.6. Пример: Белый шум



Белый шум (БШ), очевидно, является стационарным временным рядом.

8.7. Пример: Случайное блуждание

$$\boxed{y_t = y_{t-1} + \varepsilon_t, \quad \varepsilon_t - БШ}$$

Случайное блуждание не является стационарным временным рядом!

$$\boxed{My_t = 0}$$

Это не зависит от t

Дисперсия зависит от **t**!

$$\boxed{Dy_t = Dy_{t-1} + \sigma^2 = \ldots = t \cdot \sigma^2}$$

8.8. Проверка гипотез о стационарности

8.8.1. Графический метод

Первым этапом при проверке стационарности является графический анализ. Наличие *видимого тренда или систематического изменения амплитуды* колебаний говорит о *нестационарности*.

8.8.2. Критерий Фостера-Стюарта

Ещё один критерий наличия тренда среднего и дисперсии.

Вычисляются

$$S_> = \sum_{i=2}^{n} u_i, \quad u_i = \begin{cases} 1, & \text{если } x_i > \text{всех} \quad \text{предыдущих} \\ 0, & \text{иначе} \end{cases}$$

$$S_< = \sum_{i=2}^{n} l_i, \quad l_i = \begin{cases} 1, & \text{если } x_i < \text{всех} \quad \text{предыдущих} \\ 0, & \text{иначе} \end{cases}$$

$$t_1 = \frac{S_> + S_< - f^2}{l}, \quad t_2 = \frac{S_> - S_<}{f},$$

$$l \approx \sqrt{2\ln n - 3{,}4253}, \quad f \approx \sqrt{2\ln n - 0{,}8456},$$

Если $|t_1| < t_{1-\alpha/2}(n)$, то гипотеза о наличии тренда среднего отклоняется

Если $|t_2| < t_{1-\alpha/2}(n)$, то гипотеза о наличии тренда дисперсии отклоняется

Задача. Проверить гипотезу о стационарности при α=0,01.

Решение. n=15

t	y_t
1	11
2	6
3	4
4	10
5	5
6	6
7	12
8	8
9	7
10	14
11	11
12	8
13	16
14	14
15	10

$$S_> = \sum_{i=2}^{n} u_i = 3, \quad S_< = \sum_{i=2}^{n} l_i = 2$$

t	y_t	u	l
1	11	—	—
2	6	0	1
3	4	0	1
4	10	0	0
5	5	0	0
6	6	0	0
7	12	1	0
8	8	0	0
9	7	0	0
10	14	1	0
11	11	0	0
12	8	0	0
13	16	1	0
14	14	0	0
15	10	0	0

$$l \approx \sqrt{2\ln 15 - 3{,}4253} \approx 2{,}14,$$

$$f \approx \sqrt{2\ln 15 - 0{,}8456} \approx 1{,}5,$$

$$t_1 = \frac{3+2-1{,}5^2}{2{,}14} \approx 1{,}28, \quad t_2 = \frac{3-2}{1{,}5} \approx 0{,}67$$

$$t_{1-\alpha/2}(n) = t_{0,995}(15) \approx 2{,}94$$

$$|t_1| < t_{0,95}(15)$$

$$|t_2| < t_{0,95}(15)$$

Гипотеза о наличии трендов отвергается.

Стационарность возможна.

8.9. Модель авторегрессии AR(p)

8.9.1. Определение

Моделью авторегрессии AR(p) называется временной ряд, удовлетворяющий соотношению

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t,$$

где случайная составляющая ε_t – “белый шум”.

В частности, **AR(1)** – это временной ряд, удовлетворяющий соотношению

$$y_t = c + \phi \cdot y_{t-1} + \varepsilon_t$$

Пример. Случайное блуждание является процессом вида AR(1) с α=1 и с=0:

$$\boxed{y_t = y_{t-1} + \varepsilon_t, \quad \varepsilon_t - БШ}$$

8.9.2. Условия стационарности ряда AR(p)

Чтобы ряд AR(p) ,был стационарен, необходимо, чтобы все корни уравнения

$$Z^p = \phi_1 Z^{p-1} + \ldots + \phi_p Z$$

по модулю были меньше 1

1) В частности, для AR(1)

$$\boxed{y_t = c + \phi\, y_{t-1} + \varepsilon_t}$$

$$\boxed{Z^1 = \phi\, Z^0 \Leftrightarrow Z = \phi}$$

так что,

Ряд AR(1) стационарен, если и только если |φ|<1

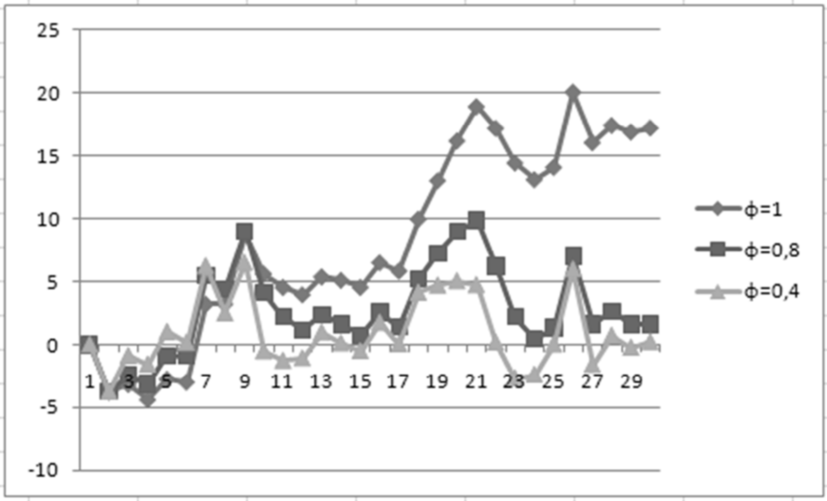
2) Для AR(2)

$$\boxed{y_t = c + \phi_1\, y_{t-1} + \phi_2\, y_{t-2} + \varepsilon_t}$$

$$\boxed{Z^2 = \phi_1\, Z + \phi_2}$$

Для стационарности ряда AR(2) корни этого уравнения должны по модулю быть меньше 1

Примеры реализации временных рядов AR(1) при различных φ:



8.9.3. Тест Дики-Фуллера (Dickey-Fuller)

Перепишем уравнение AR(1) в виде

$$\Delta y_t = c + \rho y_{t-1} + \varepsilon_t,\ \rho = \phi - 1$$

Если ρ=0, что соответствует φ = 1, то исходный ряд не стационарен.

Если ρ < 0, то φ < 1 и исходный ряд стационарен.

Тест Дики-Фуллера проверяет гипотезу H_0: ρ=0 при альтернативе H_1: ρ<0.

Если $\tau = \dfrac{\hat{\rho}}{s_\rho} > \tau_{крит}$,

где $\tau_{крит}$ – табличное значение, то гипотеза H_0 принимается (ряд нестационарный), в противном случае отвергается (ряд считается стационарным).

Для использования теста Дики-Фуллера требуются соответствующие табличные значения.

$\alpha \setminus n$	25	50	100	∞
0,01	–3,75	–3,58	–3,51	–3,43
0,05	–3,33	–3,22	–3,17	–3,12
0,1	–3	–2,93	–2,89	–2,86

Пример. Проверить гипотезу о стационарности при помощи теста DF. $\alpha=0,1$

t	y_t
1	-7,1
2	6,0
3	-7,2
4	7,7
5	-7,1
6	14,4
7	-12,1
8	17,4
9	-19,2
10	13,2
11	-12,0
12	11,1
13	-9,9
14	6,5
15	-2,8

Решение. Заполним таблицу.

t	y_t	Δ_t	y_{t-1}
1	-7,1	—	—
2	6,0	13,1	-7,1
3	-7,2	-13,2	6,0
4	7,7	14,9	-7,2
5	-7,1	-14,8	7,7
6	14,4	21,5	-7,1
7	-12,1	-26,4	14,4
8	17,4	29,5	-12,1
9	-19,2	-36,6	17,4
10	13,2	32,4	-19,2
11	-12,0	-25,2	13,2
12	11,1	23,1	-12,0
13	-9,9	-21,0	11,1
14	6,5	16,4	-9,9
15	-2,8	-9,3	6,5

Построим регрессию обычным методом МНК

$$\Delta y_t = c + \rho y_{t-1} + \varepsilon_t$$

	Коэффициенты	*Стандартная ошибка*	*татисти*
Y-пересечен	0,517	0,939	0,551002
Переменная	-1,947	0,082	-23,8356

$$\Delta y_t = 0,517 - 1,947\ y_{t-1}$$

$$\frac{\hat{\rho}}{S_\rho} = \frac{-1,947}{0,082} = -23,8356$$

Так как -23,836 гораздо меньше критического значения из таблицы, тест DF стационарность не отвергает.

Модифицированный тест Дики-Фуллера рассматривает авторегрессионные процессы более высокого порядка A(p).

9. Системы эконометрических уравнений

9.1. Введение

При моделировании часто приходится вводить не одно, а несколько связанных между собой эконометрических уравнений, т.е. описывать модель **системой уравнений**.

Пример.

$$\begin{cases} Y_1 = a_0 + a_1 Y_2 + a_2 X_1 + \varepsilon_1 \\ Y_2 = b_0 + b_1 Y_1 + b_2 X_2 + \varepsilon_2 \end{cases}$$

Классическим примером является формирование спроса Q^D и предложения Q^S товара в зависимости от его цены P и доходов I

$$\begin{cases} Q^S{}_t = a_0 + a_1 P_t + \varepsilon_1 \\ Q^D{}_t = b_0 + b_1 P_t + b_2 I_t + \varepsilon_2 \end{cases}$$

В ситуации <u>равновесия</u> на рынке $Q^D = Q^S$ и получается система уравнений

$$\begin{cases} Q_t^D = a_0 + a_1 P_t + \varepsilon_1 \\ Q_t^S = b_0 + b_1 P_t + b_2 I_t + \varepsilon_2 \\ Q_t^D = Q_t^S \end{cases}$$

Уравнения, не содержащие неизвестных коэффициентов (как третье в этом примере), называются *тождествами*.

9.2. Структурная и приведённая форма записи системы уравнений

Роль переменных в системе уравнений различна.

Эндогенные переменные − это **зависимые** переменные, значения которых которые определяются **внутри** модели. Как правило, их количество равно числу уравнений.

Экзогенные переменные − это **независимые** переменные, значения которых формируются вне модели.

С математической точки зрения их особенность в том, что они не коррелируют со случайным фактором ε.

Разделение переменных на эндогенные и экзогенные зависит от теоретической концепции.

Внеэкономические переменные (например, климатические условия) являются экзогенными.

Близкую к экзогенным роль играют значения эндогенных переменных в предыдущие моменты времени (так как они уже сформировались), т.е. лаговые переменные.

Экзогенные и лаговые эндогенные переменные называются ***предопределёнными***

Так, потребление текущего года y_t зависит не только от одновременных экономических факторов, но и от

потребления предыдущего года y_{t-1} (и это уже ***предопределённая*** переменная).

Исходная форма записи системы, когда в правой присутствуют и предопределённые и эндогенные переменные, называется ***структурной***.

Как правило, каждое уравнение в левой части содержит свою эндогенную переменную, а в правой − какие-то другие переменные, эндогенные и предопределённые.

$$\begin{cases} y_1 = a_{12}y_2 + \ldots + b_{11}x_1 + b_{12}x_2 + \ldots, \\ y_2 = a_{21}y_1 + \ldots + b_{21}x_1 + b_{22}x_2 + \ldots, \\ \quad \ldots, \\ y_m = a_{m1}y_1 + \ldots + b_{m1}x_1 + b_{m2}x_2 + \ldots \end{cases}$$

Из-за наличия в правой части эндогенных переменных для нахождения коэффициентов нельзя применять МНК непосредственно к структурной форме.

С этой целью её преобразуют в приведенную форму.

В ***приведённой*** форме записи эндогенные переменные выражены через предопределённые, в каждом уравнении в левой части стоит своя эндогенная переменная.

$$\begin{cases} y_1 = \pi_{11}x_1 + \pi_{12}x_2 + \ldots, \\ y_2 = \pi_{21}x_1 + \pi_{22}x_2 + \ldots, \\ \quad \ldots, \\ y_m = \pi_{m1}x_1 + \pi_{m2}x_2 + \ldots \end{cases}$$

Пример. Рассмотрим модель

$$\begin{cases} Q_t = b_0 + a_1 P_t + \varepsilon_1 \\ Q_t = b_1 + a_2 P_t + b_2 I_t + \varepsilon_2 \end{cases}$$

Здесь Q − равновесный спрос-предложение (эндогенная переменная), P − цена (эндогенная переменная), I − доход (экзогенная переменная).

Это ***структурная*** форма записи, числа a_j, b_j – ***структурные коэффициенты***.

Чтобы получить приведённую форму записи, выразим P,Q через I.

Перенесём все эндогенные переменные в левую часть

$$\begin{cases} Q_t - a_1 P_t = b_0 + \varepsilon_1 \\ Q_t - a_2 P_t = b_1 b_2 I_t + \varepsilon_2 \end{cases}$$

и запишем в матричной форме

$$AY = BX + \varepsilon,$$

$$A = \begin{pmatrix} 1 & -a_1 \\ 1 & -a_2 \end{pmatrix}, Y = \begin{pmatrix} Q_t \\ P_t \end{pmatrix}, B = \begin{pmatrix} b_0 & 0 \\ b_1 & b_2 \end{pmatrix}, X = \begin{pmatrix} 1 \\ I_t \end{pmatrix},$$

Если существует обратная матрица

$$A^{-1} = \begin{pmatrix} 1 & -a_1 \\ 1 & -a_2 \end{pmatrix}^{-1}$$

то, помножив на неё равенство AX+BY=ε, получим

$$Y = A^{-1}BX + A^{-1}\varepsilon \quad \Rightarrow$$

$$Y = \Pi X + \xi, \quad \xi = B^{-1}\varepsilon,$$

где

$$\Pi = \begin{pmatrix} \pi_1 & \pi_2 \\ \pi_3 & \pi_4 \end{pmatrix} = A^{-1}B$$

т.е. получаем приведённую форму модели

$$\begin{cases} Q_t = \pi_1 + \pi_2 I_t + \xi_{1t} \\ P_t = \pi_3 + \pi_4 I_t + \xi_{2t} \end{cases}$$

В правой части только предопределённые переменные, так что оценку коэффициентов здесь уже можно проводить методом наименьших квадратов.

9.3. Косвенный метод наименьших квадратов

9.3.1. Идея метода

Попробуем применить подход из предыдущего примера в общем случае.

1) Пусть исходная система дана в структурной форме

$$AY = BX + \varepsilon.$$

2) Перепишем её (если возможно) в приведённой форме

$$Y = \varPi X + \xi, \qquad \varPi = A^{-1}B.$$

3) Рассчитаем оценки коэффициентов матрицы П методом наименьших квадратов, после чего

4) Выразим через них (**<u>если это окажется возможным!</u>**) исходные структурные коэффициенты (т.е. матрицы A и B).

Такой способ оценивания структурных коэффициентов называется ***косвенным методом наименьших квадратов***.

9.3.2. Проблема идентификации

При использовании косвенного метода наименьших квадратов возникает проблема идентификации: удастся ли выразить коэффициенты исходной модели через коэффициенты приведённой модели?

Пример. Рассмотрим модель

$$\begin{cases} Q_t = a_0 + a_1 P_t + \varepsilon_1 \\ Q_t = b_0 + b_1 P_t + b_2 I_t + \varepsilon_2 \end{cases}$$

В приведённой форме она имеет вид

$$\begin{cases} Q_t = \pi_0 + \pi_1 I_t + \xi_{1t} \\ P_t = \pi_2 + \pi_3 I_t + \xi_{2t} \end{cases}$$

Ясно, что **пять** неизвестных параметров исходной модели a_0, a_1, b_0, b_1, b_2 невозможно выразить через **четыре** коэффициента приведённой модели π_0, π_1, π_2, π_3.

Структурный параметр называется

- ***идентифицируемым***, если он может быть однозначно выражен через коэффициенты приведённой формы.
- ***неидентифицируемым***, если его невозможно выразить через коэффициенты приведённой формы.
- ***сверхидентифицируемым***, если он может быть выражен через коэффициенты приведённой формы несколькими способами.

Уравнение называется ***идентифицируемым***, если идентифицируемы все входящие в него коэффициенты.

В случае неиндентифицируемости косвенный метод наименьших квадратов, очевидно, неприменим.

9.3.3. Двухшаговый метод наименьших квадратов

Если система сверхидентифицируема, то косвенный МНК не применим, структурные коэффициенты не выражаются однозначно через коэффициенты приведённой системы.

Пример. Рассмотрим следующую модель в структурной форме.

$$\begin{cases} y_{1t} = a_1 y_{2t} + b_1 x_{1t} + b_2 x_{2t} + \varepsilon_{1t}, \\ y_{2t} = a_2 y_{1t} + b_3 x_{3t} + \varepsilon_{2t}, \end{cases}$$

В приведённой форме она будет иметь вид

$$\begin{cases} y_{1t} = \pi_1 x_{1t} + \pi_2 x_{2t} + \pi_3 x_{3t} + \xi_{1t}, \\ y_{2t} = \pi_4 x_{1t} + \pi_5 x_{2t} + \pi_6 x_{3t} + \xi_{2t}, \end{cases}$$

Пусть получены оценки параметров приведённой формы

$$\begin{cases} y_{1t} = 2 x_{1t} + 3 x_{2t} + 3 x_{3t}, \\ y_{2t} = 3 x_{1t} + 4 x_{2t} + 2 x_{3t}, \end{cases}$$

Попробуем найти оценки структурной формы:

$$\boldsymbol{A\Pi = B}$$

$$\begin{pmatrix} 1 & -a_1 \\ -a_2 & 1 \end{pmatrix} \begin{pmatrix} 2 & 3 & 3 \\ 3 & 4 & 2 \end{pmatrix} = \begin{pmatrix} b_1 & b_2 & 0 \\ 0 & 0 & b_{23} \end{pmatrix}$$

Из шести уравнений есть три с нулевой правой частью, откуда получается три уравнения для нахождения a_1, a_2.

$$\begin{cases} -2a_2 + 3 = 0 \\ -3a_2 + 4 = 0 \\ 3 - 2a_1 = 0 \end{cases}$$

Видно, что a_2 выражается двумя способами, противоречащими друг другу – это и есть сверхидентифицируемость.

Алгоритм оценки коэффициентов структурной формы двухшаговым МНК

1. Оцениваются параметры приведенной формы модели при помощи МНК
2. Оцениваются параметры структурной формы модели, в правую часть которой вместо значений эндогенных переменных подставляются их оценки, рассчитанные по приведенной форме

Пример (продолжение). Пусть в предыдущем примере исходные данные имели вид:

t	y_1	y_2	x_1	x_2	x_3
1	33,4	40,0	2	7	3
2	24,0	30,3	3	5	1
3	21,3	28,6	2	5	1
4	23,5	29,8	3	4	2
5	24,9	34,7	4	5	1
6	21,8	28,9	5	3	1
7	26,1	31,9	4	4	2
8	24,5	32	6	3	1
9	32,7	36,8	7	2	4
10	33,9	37,9	8	1	5

Подставляя в найденные формулы, мы можем найти расчётные значения:

$$\begin{cases} \hat{y}_{1t} = 2x_{1t} + 3x_{2t} + 3x_{3t}, \\ \hat{y}_{2t} = 3x_{1t} + 4x_{2t} + 2x_{3t}, \end{cases}$$

t	y_1	y_2	x_1	x_2	x_3	$\hat{y}_1$	$\hat{y}_2$
1	33,4	40,0	2	7	3	34	40
2	24,0	30,3	3	5	1	24	31
3	21,3	28,6	2	5	1	22	28
4	23,5	29,8	3	4	2	24	29
5	24,9	34,7	4	5	1	26	34
6	21,8	28,9	5	3	1	22	29
7	26,1	31,9	4	4	2	26	32
8	24,5	32	6	3	1	24	32
9	32,7	36,8	7	2	4	32	37
10	33,9	37,9	8	1	5	34	38

Эти рассчитанные значения используем в качестве новых переменных в правой части структурной формы

$$\begin{cases} y_{1t} = a_1 \hat{y}_{2t} + b_1 x_{1t} + b_2 x_{2t} + \varepsilon_{1t}, \\ y_{2t} = a_2 \hat{y}_{1t} + b_3 x_{3t} + \varepsilon_{2t}, \end{cases}$$

Опять обычным методом получаем коэффициенты структурной формы

$$\begin{cases} y_{1t} = 1{,}49\,\hat{y}_{2t} - 2{,}4\,x_{1t} - 3{,}09\,x_{2t} + \varepsilon_{1t}, & R^2 = 0{,}999 \\ y_{2t} = 1{,}38\,\hat{y}_{1t} - 1{,}9\,x_{3t} + \varepsilon_{2t}, & R^2 = 0{,}999 \end{cases}$$

Таким образом, двухшаговый МНК даёт следующий ответ:

$$\begin{cases} y_{1t} = 1{,}49\,y_{2t} - 2{,}4\,x_{1t} - 3{,}09\,x_{2t} + \varepsilon_{1t}, & R^2 = 0{,}999 \\ y_{2t} = 1{,}38\,y_{1t} - 1{,}9\,x_{3t} + \varepsilon_{2t}, & R^2 = 0{,}999 \end{cases}$$

Замечания.

1) В случае точной идентифицируемости двухшаговый МНК даёт те же результаты, что и косвенный МНК.
2) Двухшаговый МНК является частным случаем метода инструментальных переменных, когда “неудобные” эндогенные переменные в правой части заменяются на новые переменные.

Литература

1. Магнус Я.Р., Катышев П.К., Пересецкий А.А. Эконометрика. Начальный курс: Учебник. – М.: Дело, 2004.

2. Эконометрика: Учебник./Под ред. И.И. Елисеевой. – М.: Финансы и статистика, 2005.

3. Практикум по эконометрике: Учеб. пособие. /И.И. Елисеева, С.В. Курышева, Н.М. Гордеенко и др. Под ред. И.И. Елисеевой. – М.: Финансы и статистика, 2005.

4. Кремер Н.Ш., Путко Б.А. Эконометрика: учебник для вузов. /Под ред. проф. Н.Ш. Кремера. – М.: ЮНИТИ–ДАНА, 2002.

5. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников. – М.: ФИЗМАТЛИТ, 2006.

6. Берндт Э. Р. Практика эконометрики: классика и современность: Учебник. – М.: ЮНИТИ–ДАНА, 2005.

7. Вербик М. Путеводитель по современной эконометрике. Пер. с англ. В.А.Банникова. – М.: Научная книга, 2008.

Оглавление

1. Введение	3
2. Корреляционный анализ количественных зависимостей	7
3. Метод наименьших квадратов	16
4. Классическая модель множественной линейной регрессии	19
5. Спецификация модели	38
6. Некоторые дополнительные вопросы	52
7. Нарушения допущений классической линейной модели	61
8. Анализ временных рядов	79
9. Системы эконометрических уравнений	104
Литература	114