МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ
БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«САМАРСКИЙ ГОСУДАРСТВЕННЫЙ
АЭРОКОСМИЧЕСКИЙ УНИВЕРСИТЕТ
ИМЕНИ АКАДЕМИКА С.П. КОРОЛЕВА
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)»

А.П.Котенко

Эконометрика

Интерактивное мультимедийное пособие

Система дистанционного обучения «Moodle»

САМАРА
2012

УДК 330.43(076)
ББК 65
К73
Автор: **Котенко Андрей Петрович**

Рецензенты:

Котенко, А.П. Эконометрика [Электронный ресурс]: интерактив. мультимед. пособие; система дистанц. обучения «Moodle»/ А.П. Котенко; Минобрнауки России, Самар. гос. аэрокосм. ун-т им. С.П. Королева (нац. исслед. ун-т). – Электрон. текстовые и граф. дан. (983 Кбайт). – Самара, 2012. – 1 эл. опт. диск (CD-ROM).

В пособии рассматриваются основы эконометрики и эконометрического моделирования.
Интерактивное мультимедийное пособие разработано на кафедре математических методов в экономике факультета экономики и управления и предназначено для студентов направлений 080100.62 Экономика, 080200.62 Менеджмент, 080500.62 Бизнес-информатика. Представленные материалы могут быть использованы на кафедре математических методов в экономике факультета экономики и управления в дисциплине Эконометрика (5 семестр).
Интерактивные материалы (конспект и презентация лекций, банк тестовых заданий) представлены по электронному адресу http:///feumoodle.ssau.ru//.

© Самарский государственный аэрокосмический университет, 2012

1. Введение.

1.1. Основные задачи и методы эконометрики.

Эконометрика посвящена разработке и применению математических, прежде всего экономико-статистических, методов анализа экономических процессов, для построения их математических моделей.

Эконометрика расположена между экономикой и математической статистикой.

Данные, как правило, не являются экспериментальными, т.к. в экономике нельзя проводить многократные эксперименты.

Эконометрические методы – это методы исследования экономики с количественной стороны.

Экономические данные являются количественными характеристиками экономических объектов и явлений. Они формируются под действием множества факторов, большинство из которых не подвластны внешнему учёту и контролю. Неконтролируемые факторы могут принимать случайные значения из некоторого допустимого множества значений.

Стохастическая природа экономических данных определяет необходимость применения адекватных статистических методов их обработки, разработанных в математической статистике.

1.2. Основные понятия и методы математической статистики.

Математическая статистика изучает методы обработки наблюдений массовых случайных явлений, обладающих статистической устойчивостью и закономерностью, с целью выявления и изучения данной закономерности.

Основным является выборочный метод, который по статистическим свойствам репрезентативной выборки позволяет судить о свойствах генеральной совокупности наблюдений.

1.3. Выборочный метод математической статистики.

Генеральная совокупность – это все возможные (допустимые в реальной жизни или мыслимые) наблюдения исследуемого показателя. Однако, практически всегда, для изучения доступна только малая часть генеральной совокупности, называемая **выборкой.**

Выборку называют **репрезентативной** (**представительной**), если она достаточно верно отражает изучаемые параметры генеральной совокупности. Хотя правила отбора репрезентативной выборки часто включены в ГОСТы, нет гарантии, что в общем случае конкретная выборка будет «правильной». Это объясняется требованием выполнения ряда предположений о генеральной совокупности, проверить которые нет возможности.

Оценка (estimator) – это общее правило (формула или способ оценивания) для получения приближённого численного значения исследуемого параметра выборки.

Значение оценки (estimation) – это число, полученное применением оценки к данной выборке. Оно является случайной величиной, зависящей от заданной выборки. Таким образом, значение оценки само является сложно устроенной случайной величиной, обладающей некоторыми параметрами – мат. ожиданием, дисперсией, средним квадратическим отклонением.

Если мат. ожидание точечной оценки θ_n совпадает со значением параметра θ для генеральной совокупности $M\theta_n=\theta$, то оценку называют **несмещённой**. В противном случае её называют **смещённой.**

Если оценка обладает наименьшей дисперсией среди всех возможных точечных (линейных) оценок величины θ, то её называют **эффективной**. В противном случае оценку называют **неэффективной.**

Если с ростом числа наблюдений $n \to +\infty$ оценка θ_n стремится *по вероятности* $\theta_n \xrightarrow{P} \theta$ (т.е. для подавляющего числа выборок) к истинному значению параметра θ, то её называют **состоятельной**. В противном случае оценку называют **несостоятельной**. Таким образом, точность несостоятельной оценки не возрастает с ростом числа наблюдений.

1.4. Свойства точечных и интервальных оценок.

1.4.1. Точечное оценивание

Математическое ожидание *дискретной* случайной величины (ДСВ) – это взвешенное среднее возможных (**спектральных**) значений с *весовыми коэффициентами* из вероятностей этих исходов. В случае *равновозможных* (**равновероятных**) наблюдений – это среднее арифметическое наблюдаемых значений.

Для ряда распределения ДСВ

x_i	x_1	x_2	…	x_n
p_i	p_1	p_2	…	p_n

получим формулу **мат. ожидания** ДСВ: $MX := x_1 p_1 + x_2 p_2 + \ldots + x_n p_n = \sum_{i=1}^{n} x_i p_i$.

Для *непрерывной* случайной величины (НСВ) с плотностью распределения $f(x)$ получается аналогично: $MX := \int_{-\infty}^{+\infty} x f(x) dx$.

Мат. ожидание СВ является её **средним по генеральной совокупности**.

Среднее по выборке даёт **выборочную** (*несмещённую*) оценку мат. ожидания: $\bar{x} := \frac{1}{n} \sum_{i=1}^{n} x_i$.

Точность выборочного среднего растёт как $\sigma_x / \sqrt{n}$ с ростом объёма выборки n.

Теоретическая (**генеральная**) **дисперсия** – это мера разброса значений СВ относительно её мат. ожидания: $DX = \sigma_x^2 := M(X - MX)^2 = MX^2 - M^2 X$.

Для ДСВ $\sigma_x^2 := \sum_{i=1}^{n} (x_i - MX)^2 p_i = \sum_{i-1}^{n} x_i^2 p_i - \left(\sum_{i=1}^{n} x_i p_i \right)^2$.

Для НСВ $\sigma_x^2 := \int_{-\infty}^{+\infty}(x - MX)^2 f(x)dx = \int_{-\infty}^{+\infty} x^2 f(x)dx - M^2 X$.

Среднее по выборке даёт смещённую (заниженную в среднем) выборочную оценку дисперсии $S^2 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i-1}^{n} x_i^2 - \bar{x}^2$, которую называют **выборочной дисперсией**.

Смещение появляется из-за замены истинного значения *MX* мат. ожидания его приближённым значением $\bar{x}$.

Несмещённая оценка дисперсии $s^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n}{n-1}S^2$ называется **исправленной выборочной дисперсией**.

Корень квадратный из дисперсии $\sigma_x := \sqrt{\sigma_x^2}$ называют **средним квадратическим отклонением** (СКО). Оно так же как дисперсия служит мерой рассеяния СВ относительно её среднего значения, но более удобно, т.к. при смене размерности изменяется пропорционально мат. ожиданию.

Корень из выборочной дисперсии $S := \sqrt{S^2}$ является **точечной (смещённой**) оценкой генерального среднего квадратического отклонения.

Исправленное выборочное среднее квадратическое отклонение $s := \sqrt{s^2} = \sqrt{\frac{n}{n-1}}S$ называют ещё **эмпирическим стандартом**. Он является несмещённой оценкой СКО и даёт более предпочтительную оценку, особенно когда число наблюдений невелико: $n < 18 \div 20$.

Зависимость двух СВ X и Y изучается с помощью **ковариации (ковариационного или корреляционного момента)**

$$\operatorname{cov}(X,Y) := M[(X - MX)(Y - MY)] = MXY - MX \cdot MY.$$

При $X \equiv Y$ ковариация совпадает с дисперсией (**вариацией**):

$$\operatorname{var}(X) := \operatorname{cov}(X, X) = \sigma_x^2.$$

Точечная оценка ковариации даётся формулой **выборочной ковариации**

$$\operatorname{Cov}(X,Y) := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x} \cdot \bar{y} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{1}{n^2}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right),$$

которая смещена, т.к. $M \operatorname{Cov}(X,Y) = \frac{n-1}{n}\operatorname{cov}(X,Y)$.

Точечная оценка вариации совпадает с выборочной (смещённой) дисперсией: $\operatorname{Var}(X) := S^2$.

Исправленная (несмещённая) выборочная ковариация

$$\operatorname{cov}(X,Y) := \frac{n}{n-1}\operatorname{Cov}(X,Y),$$

как и **исправленная (несмещённая) вариация**

$$\operatorname{var}(X) := \frac{n}{n-1}\operatorname{Var}(X)$$

являются несмещёнными оценками соответствующих парметров всей генеральной совокупности.

Для уточнения силы зависимости СВ применяется **коэффициент корреляции** $r_{XY} := \frac{\operatorname{cov}(X,Y)}{\sigma_x \sigma_y} = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X) \cdot \operatorname{var}(Y)}}$.

Выборочный коэффициент корреляции, рассчитанный по выборочным ковариации и дисперсиям $\hat{r}_{XY} := \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X) \cdot \operatorname{Var}(Y)}}$, является несмещённой оценкой **теоретического коэффициента корреляции** генеральной совокупности.

1.4.2. Интервальное оценивание

Формулы для расчёта **доверительных интервалов** исследуемых параметров: $\gamma_a = a \pm \Delta_a$, $\gamma_b = b \pm \Delta_b$.

Если в доверительный интервал попадает 0, то оцениваемый параметр принимается нулевым, т.к. он не может одновременно принимать и положительные и отрицательные значения.

2. Парная регрессия.

2.1. Основные понятия.

Парная регрессия – это уравнение $y = \tilde{f}(x) + \varepsilon$ связи переменных *y* и *x*, где *y* – **зависимая переменная** (**результативный признак**), *x* – **независимая объясняющая переменная** (**признак-фактор**, *регрессор*), ε – **случайный член регрессии** (*остаток*, *невязка*) в *аддитивной* форме записи, учитывающий случайные (стохастические, недетерминированные) воздействия.

Возможна **мультипликативная форма** записи регрессии $y = \tilde{f}(x) \cdot \varepsilon$, а также другие математические способы учёта стохастических влияний.

Идентификация уравнения регрессии сводится к оценке её параметров. Для *линейных по неизвестным параметрам* регрессий применим метод наименьших квадратов (МНК), позволяющий найти оценки, минимизирующие сумму квадратов отклонений **фактических** (**наблюдаемых**, **эксперименталь-**

ных) значений результативного признака *y* от **теоретических** (**гипотетических**, *регрессионных*, **сглаженных**) значений $\tilde{y}$: $\sum(y_x - \tilde{y}_x)^2 \to \min$.

Для линейных, а также нелинейных, *приводимых к линейным по идентифицируемым параметрам*, регрессионных моделей решается следующая, линейная по искомым точечным оценкам идентифицируемых параметров, система алгебраических уравнений в нормальной форме:

$$\begin{cases} an + b\sum x = \sum y, \\ a\sum x + b\sum x^2 = \sum xy. \end{cases}$$

Её решения имеют вид $a = \bar{y} - b\bar{x}$, $b = \dfrac{\operatorname{cov}(x,y)}{\sigma_x^2} = \dfrac{\overline{xy} - \bar{x}\cdot\bar{y}}{\overline{x^2} - \bar{x}^2}$.

Интерпретация коэффициента регрессии *b* – изменение регрессора *x* на 1 единицу влечёт изменение исследуемого фактора *y* на *b* единиц. Т.о. коэффициент *b* – **экономически интерпретируемый** параметр (это иногда позволяет, к примеру, проверить его экспериментально).

Математическая интерпретация свободного члена регрессии *a* – значение *y* при нулевом значении регрессора *x*=0. Однако, в большинстве экономических задач регрессор не может принять нулевое или даже близкое к 0 значение. Поэтому свободный член линейной регрессии *a* относят к **неинтерпретируемым параметрам**, которые почти никогда не удаётся проверить экспериментально.

При интерпретации параметров линейной регрессии отметим 3 факта.

1. Параметры a и b являются лишь выборочными (точечными) оценками параметров линейной регрессии $\tilde{y} = \alpha + \beta x$ генеральной совокупности, поэтому возможно сколь угодно большое их отклонение от «истинных» значений.

2. Уравнение регрессии (не только линейное) отражает лишь **общую тенденцию** выборки, когда каждое отдельное исследование подвержено воздействию сколь угодно больших случайностей.

3. Верность интерпретации зависит от правильности *спецификации* уравнения. «Наилучшая» **МНК-прямая** всегда существует, но не всегда является достаточно хорошей.

2.2. Методы наименьших квадратов и максимального правдоподобия.

Пусть имеется n совместных наблюдений дискретных случайных величин X и Y. Представим их n точками (x_i,y_i) плоскости $\Re^2$. Спроектируем каждую точку параллельно оси Oy на прямую искомой линейной регрессии $\hat{y} = ax + b$. Получим **регрессионные** (**теоретические**, **сглаженные**, **гипотетические**) значения $\hat{y}_i = ax_i + b$.

Подберём теперь параметры a и b искомой регрессии из условия

$$S(a,b) := \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2 \to \min.$$

Для этого воспользуемся **необходимым условием экстремума** функции двух переменных: $\begin{cases} \partial S/\partial a = 0, \\ \partial S/\partial b = 0. \end{cases}$

Получим $\begin{cases} \partial S/\partial a = 2\sum_{i=1}^{2}\left(y_i - ax_i - b\right)\left(-x_i\right) = 0, \\ \partial S/\partial b = 2\sum_{i=1}^{2}\left(y_i - ax_i - b\right)\left(-1\right) = 0. \end{cases}$

Приведём эту систему 2-х линейных неоднородных уравнений с 2-мя неизвестными к **нормальному виду**, когда неизвестные находятся в левой, а известные слагаемые в правой части уравнений:

$$\begin{cases} a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i, \\ a\sum_{i=1}^{n} x_i + b \cdot n = \sum_{i=1}^{n} y_i. \end{cases} \qquad (1)$$

Решить её можно известными методами исключения неизвестных, обратной матрицы или по формулам Крамера. К примеру, по формулам Крамера решение примет вид $a = \Delta_a/\Delta$, $b = \Delta_b/\Delta$, где **определитель системы**

$$\Delta := \begin{vmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^0 \end{vmatrix} = \begin{vmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{vmatrix} \neq 0,\ \Delta_a := \begin{vmatrix} \sum_{i=1}^{n} x_i y_i & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} y_i & n \end{vmatrix},\ \Delta_b := \begin{vmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} y_i \end{vmatrix}.$$

Из первого уравнения СЛАУ (1) получим $a = \mathrm{Cov}(X,Y)/\mathrm{Var}(X)$. Тогда из второго уравнения $b = \bar{y} - a\bar{x}$. Последнее равенство означает, что МНК-прямая всегда проходит через **центр** $(\bar{x},\bar{y})$ диаграммы рассеяния.

2.3. Вывод уравнений МНК для парной полиномиальной регрессии.

Следуя тем же рассуждениям, можно получить соответствующее решение для случая **нелинейной полиномиальной регрессии.**

Так для **квадратической** регрессии $\tilde{y} = ax^2 + bx + c$ получим:

$$S(a,b,c) := \sum_{i=1}^{n}\left(y_i - \tilde{y}_i\right)^2 = \sum_{i=1}^{n}\left(y_i - ax_i^2 - bx_i - c\right)^2 \to \min \ \Rightarrow$$

$$\Rightarrow \begin{cases} \partial S/\partial a = 0, \\ \partial S/\partial b = 0, \\ \partial S/\partial c = 0, \end{cases} \Leftrightarrow \begin{cases} \partial S/\partial a = 2\sum_{i=1}^{n}\left(y_i - ax_i^2 - bx_i - c\right)\left(-x_i^2\right) = 0, \\ \partial S/\partial b = 2\sum_{i=1}^{n}\left(y_i - ax_i^2 - bx_i - c\right)\left(-x_i\right) = 0, \\ \partial S/\partial c = 2\sum_{i=1}^{n}\left(y_i - ax_i^2 - bx_i - c\right)\left(-1\right) = 0, \end{cases} \Leftrightarrow$$

$$\Leftrightarrow \begin{cases} a\sum_{i=1}^{n}x_i^4 + b\sum_{i=1}^{n}x_i^3 + c\sum_{i=1}^{n}x_i^2 &= \sum_{i=1}^{n}x_i^2 y_i, \\ a\sum_{i=1}^{n}x_i^3 + b\sum_{i=1}^{n}x_i^2 + c\sum_{i=1}^{n}x_i &= \sum_{i=1}^{n}x_i y_i, \\ a\sum_{i=1}^{n}x_i^2 + b\sum_{i=1}^{n}x_i + c\cdot n &= \sum_{i=1}^{n}y_i. \end{cases} \qquad (2)$$

Тогда по формулам Крамера $a = \Delta_a/\Delta$, $b = \Delta_b/\Delta$, $c = \Delta_c/\Delta$, где

$$\Delta := \begin{vmatrix} \sum_{i=1}^{n}x_i^4 & \sum_{i=1}^{n}x_i^3 & \sum_{i=1}^{n}x_i^2 \\ \sum_{i=1}^{n}x_i^3 & \sum_{i=1}^{n}x_i^2 & \sum_{i=1}^{n}x_i \\ \sum_{i=1}^{n}x_i^2 & \sum_{i=1}^{n}x_i & n \end{vmatrix} \neq 0,\ \Delta_a := \begin{vmatrix} \sum_{i=1}^{n}x_i^2 y_i & \sum_{i=1}^{n}x_i^3 & \sum_{i=1}^{n}x_i^2 \\ \sum_{i=1}^{n}x_i y_i & \sum_{i=1}^{n}x_i^2 & \sum_{i=1}^{n}x_i \\ \sum_{i=1}^{n}y_i & \sum_{i=1}^{n}x_i & n \end{vmatrix},$$

$$\Delta_b := \begin{vmatrix} \sum_{i=1}^{n}x_i^4 & \sum_{i=1}^{n}x_i^2 y_i & \sum_{i=1}^{n}x_i^3 \\ \sum_{i=1}^{n}x_i^3 & \sum_{i=1}^{n}x_i y_i & \sum_{i=1}^{n}x \\ \sum_{i=1}^{n}x_i^2 & \sum_{i=1}^{n}y_i & n \end{vmatrix},\ \Delta_c := \begin{vmatrix} \sum_{i=1}^{n}x_i^4 & \sum_{i=1}^{n}x_i^3 & \sum_{i=1}^{n}x_i^2 y_i \\ \sum_{i=1}^{n}x_i^3 & \sum_{i=1}^{n}x_i^2 & \sum_{i=1}^{n}x_i y_i \\ \sum_{i=1}^{n}x_i^2 & \sum_{i=1}^{n}x_i & \sum_{i=1}^{n}y_i \end{vmatrix}.$$

Из третьего уравнения СЛАУ (2) вновь легко получить равенство $\bar{y} = a\overline{x^2} + b\bar{x} + c$. Поскольку $\overline{x^2} \neq \bar{x}^2$, **МНК-парабола** в общем случае уже не проходит через центр $(\bar{x}, \bar{y})$ диаграммы рассеяния.

Аналогично, для регрессии $\tilde{y} = a_3x^3 + a_2x^2 + a_1x + a_0$ получим

$$S(a_3, a_2, a_1, a_0) := \sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 = \sum_{i=1}^{n}\left(y_i - a_3x_i^3 - a_2x_i^2 - a_1x_i - a_0\right)^2 \to \min \Rightarrow$$

$$\Rightarrow \begin{cases} \partial S/\partial a_3 = 0, \\ \partial S/\partial a_2 = 0, \\ \partial S/\partial a_1 = 0, \\ \partial S/\partial a_0 = 0, \end{cases} \Leftrightarrow \begin{cases} \partial S/\partial a_3 = 2\sum_{i=1}^{n}\left(y_i - a_3x_i^3 - a_2x_i^2 - a_1x_i - a_0\right)\left(-x_i^3\right) = 0, \\ \partial S/\partial a_2 = 2\sum_{i=1}^{n}\left(y_i - a_3x_i^3 - a_2x_i^2 - a_1x_i - a_0\right)\left(-x_i^2\right) = 0, \\ \partial S/\partial a_1 = 2\sum_{i=1}^{n}\left(y_i - a_3x_i^3 - a_2x_i^2 - a_1x_i - a_0\right)\left(-x_i\right) = 0, \\ \partial S/\partial a_0 = 2\sum_{i=1}^{n}\left(y_i - a_3x_i^3 - a_2x_i^2 - a_1x_i - a_0\right)\left(-1\right) = 0, \end{cases} \Leftrightarrow$$

$$\Leftrightarrow \begin{cases} a_3\sum_{i=1}^n x_i^6 + a_2\sum_{i=1}^n x_i^5 + a_1\sum_{i=1}^n x_i^4 + a_0\sum_{i=1}^n x_i^3 &= \sum_{i=1}^n x_i^3 y_i, \\ a_3\sum_{i=1}^n x_i^5 + a_2\sum_{i=1}^n x_i^4 + a_1\sum_{i=1}^n x_i^3 + a_0\sum_{i=1}^n x_i^2 &= \sum_{i=1}^n x_i^2 y_i, \\ a_3\sum_{i=1}^n x_i^4 + a_2\sum_{i=1}^n x_i^3 + a_1\sum_{i=1}^n x_i^2 + a_0\sum_{i=1}^n x_i &= \sum_{i=1}^n x_i y_i, \\ a_3\sum_{i=1}^n x_i^3 + a_2\sum_{i=1}^n x_i^2 + a_1\sum_{i=1}^n x_i + a_0 \cdot n &= \sum_{i=1}^n y_i. \end{cases} \quad (3)$$

Тогда по формулам Крамера $a_3 = \Delta_3/\Delta$, $a_2 = \Delta_2/\Delta$, $a_1 = \Delta_1/\Delta$, $a_0 = \Delta_0/\Delta$, где

$$\Delta := \begin{vmatrix} \sum_{i=1}^n x_i^6 & \sum_{i=1}^n x_i^5 & \sum_{i=1}^n x_i^4 & \sum_{i=1}^n x_i^3 \\ \sum_{i=1}^n x_i^5 & \sum_{i=1}^n x_i^4 & \sum_{i=1}^n x_i^3 & \sum_{i=1}^n x_i^2 \\ \sum_{i=1}^n x_i^4 & \sum_{i=1}^n x_i^3 & \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i^3 & \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i & n \end{vmatrix} \neq 0, \quad \Delta_3 := \begin{vmatrix} \sum_{i=1}^n x_i^3 y_i & \sum_{i=1}^n x_i^5 & \sum_{i=1}^n x_i^4 & \sum_{i=1}^n x_i^3 \\ \sum_{i=1}^n x_i^2 y_i & \sum_{i=1}^n x_i^4 & \sum_{i=1}^n x_i^3 & \sum_{i=1}^n x_i^2 \\ \sum_{i=1}^n x_i y_i & \sum_{i=1}^n x_i^3 & \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n y_i & \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i & n \end{vmatrix},$$

$$\Delta_2 := \begin{vmatrix} \sum_{i=1}^n x_i^6 & \sum_{i=1}^n x_i^3 y_i & \sum_{i=1}^n x_i^4 & \sum_{i=1}^n x_i^3 \\ \sum_{i=1}^n x_i^5 & \sum_{i=1}^n x_i^2 y_i & \sum_{i=1}^n x_i^3 & \sum_{i=1}^n x_i^2 \\ \sum_{i=1}^n x_i^4 & \sum_{i=1}^n x_i y_i & \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i^3 & \sum_{i=1}^n y_i & \sum_{i=1}^n x_i & n \end{vmatrix}, \quad \Delta_1 := \begin{vmatrix} \sum_{i=1}^n x_i^6 & \sum_{i=1}^n x_i^5 & \sum_{i=1}^n x_i^3 y_i & \sum_{i=1}^n x_i^3 \\ \sum_{i=1}^n x_i^5 & \sum_{i=1}^n x_i^4 & \sum_{i=1}^n x_i^2 y_i & \sum_{i=1}^n x_i^2 \\ \sum_{i=1}^n x_i^4 & \sum_{i=1}^n x_i^3 & \sum_{i=1}^n x_i y_i & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i^3 & \sum_{i=1}^n x_i^2 & \sum_{i=1}^n y_i & n \end{vmatrix},$$

$$\Delta_0 := \begin{vmatrix} \sum_{i=1}^n x_i^6 & \sum_{i=1}^n x_i^5 & \sum_{i=1}^n x_i^4 & \sum_{i=1}^n x_i^3 y_i \\ \sum_{i=1}^n x_i^5 & \sum_{i=1}^n x_i^4 & \sum_{i=1}^n x_i^3 & \sum_{i=1}^n x_i^2 y_i \\ \sum_{i=1}^n x_i^4 & \sum_{i=1}^n x_i^3 & \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n x_i^3 & \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i & \sum_{i=1}^n y_i \end{vmatrix}.$$

Из последнего уравнения СЛАУ (3) вновь легко получить равенство

$$\bar{y} = a_3\overline{x^3} + a_2\overline{x^2} + a_1\bar{x} + a_0.$$

Из него следует, что и **МНК-парабола 3-ей степени** в общем случае не проходит через центр $(\bar{x}, \bar{y})$ диаграммы рассеяния.

В общем случае **полиномиальной регрессии** $\tilde{y} = a_k x^k + a_{k-1}x^{k-1} + a_1 x + a_0$ получим

$$S(a_k, a_{k-1}, \ldots, a_0) := \sum_{i=1}^n (y_i - \tilde{y}_i)^2 = \sum_{i=1}^n \left(y_i - a_k x_i^k - a_{k-1}x_i^{k-1} - \ldots - a_1 x_i - a_0\right)^2 \to \min \Rightarrow$$

$$\Rightarrow \begin{cases} \partial S/\partial a_k = 0, \\ \partial S/\partial a_{k-1} = 0, \\ \dots \\ \partial S/\partial a_0 = 0, \end{cases} \Leftrightarrow \begin{cases} \partial S/\partial a_k = 2\sum_{i=1}^{n}\left(y_i - a_k x_i^k - a_{k-1}x_i^{k-1} - \dots - a_1 x_i - a_0\right)\left(-x_i^k\right) = 0, \\ \partial S/\partial a_{k-1} = 2\sum_{i=1}^{n}\left(y_i - a_k x_i^k - a_{k-1}x_i^{k-1} - \dots - a_1 x_i - a_0\right)\left(-x_i^{k-1}\right) = 0, \\ \dots \\ \partial S/\partial a_0 = 2\sum_{i=1}^{n}\left(y_i - a_k x_i^k - a_{k-1}x_i^{k-1} - \dots - a_1 x_i - a_0\right)(-1) = 0, \end{cases} \Leftrightarrow$$

$$\Leftrightarrow \begin{cases} a_k\sum_{i=1}^{n}x_i^{2k} + a_{k-1}\sum_{i=1}^{n}x_i^{2k-1} + \dots + a_1\sum_{i=1}^{n}x_i^{k+1} + a_0\sum_{i=1}^{n}x_i^{k} = \sum_{i=1}^{n}x_i^k y_i, \\ a_k\sum_{i=1}^{n}x_i^{2k-1} + a_{k-1}\sum_{i=1}^{n}x_i^{2k-2} + \dots + a_1\sum_{i=1}^{n}x_i^{k} + a_0\sum_{i=1}^{n}x_i^{k-1} = \sum_{i=1}^{n}x_i^{k-1} y_i, \\ \dots \quad \dots \quad \dots \\ a_k\sum_{i=1}^{n}x_i^{k} + a_{k-1}\sum_{i=1}^{n}x_i^{k-1} + \dots + a_1\sum_{i=1}^{n}x_i + a_0 \cdot n = \sum_{i=1}^{n}y_i. \end{cases} \quad (4)$$

Тогда по формулам Крамера $a_k = \Delta_k/\Delta$, $a_{k-1} = \Delta_{k-1}/\Delta, \dots$, $a_1 = \Delta_1/\Delta$, $a_0 = \Delta_0/\Delta$, где

$$\Delta := \begin{vmatrix} \sum_{i=1}^{n}x_i^{2k} & \sum_{i=1}^{n}x_i^{2k-1} & \dots & \sum_{i=1}^{n}x_i^{k} \\ \sum_{i=1}^{n}x_i^{2k-1} & \sum_{i=1}^{n}x_i^{2k-2} & \dots & \sum_{i=1}^{n}x_i^{k-1} \\ \dots & \dots & \dots & \dots \\ \sum_{i=1}^{n}x_i^{k} & \sum_{i=1}^{n}x_i^{k-1} & \dots & n \end{vmatrix} \neq 0, \dots$$

Из последнего уравнения СЛАУ (4) вновь легко получить равенство

$$\bar{y} = a_k\overline{x^k} + a_{k-1}\overline{x^{k-1}} + \dots + a_1\bar{x} + a_0.$$

Вновь заключаем, что полученная кривая регрессии не проходит в общем случае через центр $(\bar{x}, \bar{y})$ диаграммы рассеяния.

2.4. Линейная и нелинейная регрессия.

Различают **линейные** и **нелинейные** регрессии.

Линейная регрессия: $y=a+bx+\varepsilon$.

Нелинейные регрессии делятся на регрессии, нелинейные относительно регрессоров, и регрессии, нелинейные по оцениваемым параметрам.

Регрессии, *нелинейные по регрессорам*:

полиномы $y = a + b_1 x + b_2 x^2 + \ldots + b_k x^k + \varepsilon$;

(данная зависимость **линеаризуется** сведением к множественной линейной регрессии путём введения **новых регрессоров** $X_1 := x$, $X_2 := x^2$, …, $X_k := x^k$);

равносторонняя гипербола $y = a + b/x + \varepsilon$;

(данная зависимость линеаризуется заменой $X := 1/x$);

и т.п.

Регрессии, *нелинейные по оцениваемым параметрам*:

показательная зависимость $y = ab^x \cdot \varepsilon$;

(данная зависимость линеаризуется заменами $Y := \ln y$, $A := \ln a$, $B := \ln b$, $E := \ln \varepsilon$);

степенная зависимость $y = ax^b \cdot \varepsilon$;

(данная зависимость линеаризуется заменами $Y := \ln y$, $X := \ln x$, $A := \ln a$, $E := \ln \varepsilon$);

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

(данная зависимость линеаризуется заменами $Y := \ln y$, $E := \ln \varepsilon$);

и т.п.

Регрессии, которые можно свести к линейным, называют **нелинейными внутренне линейными**. К ним применим МНК.

Регрессии, которые не сводятся к линейным, называют **нелинейными внутренне нелинейными**. К ним МНК не применим, т.к. система уравнений для идентификации **МНК-параметров** является нелинейной и требует специальных приёмов нахождения точного или приближённого решения.

2.5. Метод Зарембки

Метод Зарембки позволяет осуществить выбор между линейной и **логарифмической** моделями. Его алгоритм состоит из 4-х шагов.

1. Найти **среднее геометрическое по выборке** $\overline{Y}_{\text{геом}} := \sqrt[n]{y_1 y_2 y_3 \dots y_n}$.

2. Ввести новые переменные $y_i^* := y_i / \overline{Y}_{\text{геом}}$.

3. Найти МНК-регрессии с наблюдениями y_i^* вместо y_i и логарифмическую регрессию с наблюдениями $\ln y_i^*$ вместо $\ln y_i$.

4. Вычислить остаточные суммы квадратов полученных регрессий RSS_1 и RSS_2.

5. Если **статистика** $\left|\frac{n}{2}\ln\frac{RSS_1}{RSS_2}\right|$ превышает **критическое значение** распределения χ^2 с одной степенью свободы, то выбирается модель $y=ax^b\varepsilon$, если не превышает – выбирается модель $y=a+bx+\varepsilon$.

2.6. Тест Бокса–Кокса (решётчатый поиск)

Дж.Бокс и Д.Кокс заметили, что из функции $(y^\lambda-1)\lambda$ при $\lambda=1$ получается функция $y-1$, а при $\lambda \to 0$ получается функция $\ln y$. Есть смысл подобрать значение $\lambda \in (0,1]$, минимизирующее СКО. Эту процедуру называют **решётчатым поиском**, или **поиском по сетке**.

2.7. Коэффициент и индекс корреляции.

Коэффициент парной линейной корреляции

$$r_{xy} := b\frac{\sigma_x}{\sigma_y} = \frac{\operatorname{cov}(x,y)}{\sigma_x\sigma_y} = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2}\cdot\sqrt{\overline{y^2} - \bar{y}^2}},\ -1 \le r_{xy} \le 1,$$

оценивает тесноту *линейной* связи случайных величин, а **индекс корреляции**

$$\rho_{xy} := \sqrt{1 - \frac{\sigma^2_{\text{ост}}}{\sigma^2_y}} = \sqrt{1 - \frac{\sum(y_x - \tilde{y}_x)^2}{\sum(y_x - \bar{y})^2}},\ 0 \le \rho_{xy} \le 1,$$

оценивает тесноту *нелинейной* связи.

2.8. Коэффициент и индекс детерминации.

Оценку **адекватности** построенной модели экспериментальным наблюдениям даёт коэффициент (индекс) детерминации, а также средняя ошибка аппроксимации.

Долю дисперсии, объясняемую регрессией, в общей дисперсии результативного признака y характеризует **коэффициент (индекс) детерминации**

$$R^2 := \frac{\operatorname{Var}(\tilde{y})}{\operatorname{Var}(y)} = \frac{\sum(\tilde{y}_x - \bar{y})^2}{\sum(y_x - \bar{y})^2},\ \text{или}\ R^2 := 1 - \frac{\operatorname{Var}(\varepsilon)}{\operatorname{Var}(y)} = 1 - \frac{\sum(y_x - \tilde{y}_x)^2}{\sum(y_x - \bar{y})^2}.$$

Коэффициент (индекс) *детерминации* равен квадрату коэффициента (индекса) линейной парной *корреляции*.

Средняя ошибка аппроксимации $\overline{A} := \frac{1}{n}\sum\left|\frac{y_x - \tilde{y}_x}{y_x}\right| \cdot 100\%$ оценивает **среднее отклонение** расчётных значений от фактических. Допустимый предел значений $\overline{A}$ – не более 8-10%.

2.9. Расчёт коэффициента детерминации линейной регрессии и индекса детерминации нелинейной регрессии; их сравнение.

Задача **дисперсионного анализа** заключается в анализе дисперсии зависимой переменной:

$$TSS := \sum(y_x - \bar{y})^2 = \sum(\tilde{y}_x - \bar{y})^2 + \sum(y_x - \tilde{y}_x)^2,$$

где $\sum(y_x - \bar{y})^2$ – **общая сумма квадратов отклонений**,

$ESS := \sum(\tilde{y}_x - \bar{y})^2$ – сумма квадратов отклонений, обусловленная (**объяснённая**) наличием регрессии,

$RSS := \sum(y_x - \tilde{y}_x)^2$ – **остаточная** (**необъяснённая**) сумма квадратов отклонений.

Для *линейной* регрессии *TSS=ESS+RSS*, для *нелинейной* регрессии такое свойство отсутствует, поэтому *коэффициент* детерминации заменяется на **индекс детерминации** $R^2 := 1 - \frac{RSS}{TSS}$.

С увеличением **объяснённой доли разброса** $R^2 \to 1 - 0$.

Если заметная связь между *x* и *y* отсутствует, то $R^2 \approx 0$.

Корреляция между y и $\tilde{y}$:

$$r_{y,\tilde{y}} = \frac{\mathrm{Cov}(y,\tilde{y})}{\sqrt{\mathrm{Var}(y)\cdot\mathrm{Var}(\tilde{y})}} = \frac{\mathrm{Cov}(\tilde{y}+\varepsilon,\tilde{y})}{\sqrt{\mathrm{Var}(y)\cdot\mathrm{Var}(\tilde{y})}} = \frac{\mathrm{Cov}(\tilde{y},\tilde{y})+\mathrm{Cov}(\varepsilon,\tilde{y})}{\sqrt{\mathrm{Var}(y)\cdot\mathrm{Var}(\tilde{y})}} = \frac{\mathrm{Var}(\tilde{y})}{\sqrt{\mathrm{Var}(y)\cdot\mathrm{Var}(\tilde{y})}} =$$

$$= \sqrt{\frac{Var(\tilde{y})}{Var(y)}} = \sqrt{R^2} \;\Leftrightarrow\; R^2 = r^2_{y,a+bx}.$$

2.10. Линеаризация нелинейных моделей.

2.10.1. Степенная модель регрессии.

Применению МНК для степенной модели регрессии $y = ax^b \cdot \varepsilon$ предшествует **линеаризация** с помощью логарифмирования обеих частей уравнения:

$$\ln y = \ln a + b\ln x \Leftrightarrow Y = A + bX,$$

где Y:=ln$y$, X:=ln$x$, A:=ln$a$.

2.10.2. Показательная модель.

Линеаризация показательной кривой $y = ab^x \cdot \varepsilon$ также производится логарифмированием: $\ln y = \ln a + x\ln b \Leftrightarrow Y = A + Bx$, где Y:=ln$y$, B:=ln$b$, A:=ln$a$.

2.10.3. Модель равносторонней гиперболы.

Уравнение равносторонней гиперболы $y = a + b/x + \varepsilon$ заменой $z := 1/x$ **приводится к линейному виду** $y=a+bz$.

2.11. *F*-критерий Фишера.

Расчёт *F*-критерия Фишера парной регрессии.

Оценка качества уравнения регрессии состоит в проверке *нулевой гипотезы* H_0 о **статистической незначимости** уравнения регрессии и показателя тесноты связи. Для этого *фактическое* значение статистики $F_{\text{факт}}$ сравнивается с *критическим* (*табличным*) значением $F_{\text{табл}}$ критерия Фишера.

Фактическое значение статистики Фишера равно отношению удельных (рассчитанных на одну степень свободы) факторной и остаточной дисперсий:

$$F_{\text{факт}} := \frac{\sum\left(\tilde{y}_x - \bar{y}\right)^2 / m}{\sum\left(y_x - \tilde{y}_x\right)^2 / \left(n-m-1\right)} = \frac{r_{xy}^2}{1-r_{xy}^2}\left(n-2\right),$$

где n – объём выборки, m – число параметров при регрессорах.

Табличное значение $F_{\text{табл}}$ – это максимально возможное значение *F*-статистики при действии случайных факторов при заданных уровне значимости α и числе степеней свободы.

F-статистика Фишера имеет 2 числа **степеней свободы**:

– **верхнее число** = числу объясняющих переменных (в парной регрессии =1);

– **нижнее число** = числу наблюдений в выборке минус число оцениваемых параметров (в парной регрессии =n–2).

$$F_{\text{факт}} := \frac{ESS/p}{RSS/q},$$

где p – верхнее число степеней свободы, q – нижнее число степеней свободы.

Для *парной* регрессии $F_{\text{факт}} := \frac{ESS}{RSS} \cdot (n-2)$.

Уровень значимости α – вероятность *отвергнуть верную* гипотезу. Стандартное значение α=0,05; часто табулируются значения 0,01 и 0,1.

При $F_{\text{табл}}<F_{\text{факт}}$ нулевая гипотеза H_0 о случайном характере оцениваемых характеристик отклоняется и принимается их **статистическая значимость и надёжность**. В противном случае $F_{\text{табл}}>F_{\text{факт}}$ гипотеза H_0 принимается, т.е. признаётся **статистическая незначимость и ненадёжность** найденного уравнения регрессии.

Если $1 \le F < +\infty$, следует рассмотреть F^{-1}.

2.12. Использование *t*-критерия Стьюдента.

Для оценки **статистической значимости** коэффициентов регрессии и корреляции рассчитывается *t*-критерий Стьюдента и доверительные интервалы каждого показателя. *Нулевая гипотеза* H_0 говорит о **случайной природе показателей**, т.е. незначимом их отклонении от 0. С табличным значением сравнивается отношение значения оценки и её случайной ошибки

$$t_a := a/m_a\ ,\ t_b := b/m_b\ ,\ t_r := r/m_r\ ,$$

где **случайные ошибки** точечных оценок параметров определяются формулами

$$m_a := \sqrt{\frac{\sum (y_x - \tilde{y}_x)^2}{n-2} \cdot \frac{\sum x^2}{n\sum (x-\bar{x})^2}} = \sqrt{S_{\text{ост}}^2 \frac{\sum x^2}{n^2\sigma_x^2}} = \frac{S_{\text{ост}}}{n\sigma_x}\sqrt{\sum x^2}\ ,$$

$$m_b := \sqrt{\frac{\sum (y_x - \tilde{y}_x)^2}{(n-2)\sum (x-\bar{x})^2}} = \sqrt{\frac{S_{ост}^2}{\sum (x-\bar{x})^2}} = \frac{S_{ост}}{\sigma_x \sqrt{n}},\ m_{r_{xy}} := \sqrt{\frac{1-r_{xy}^2}{n-2}}.$$

Сравнивая *фактическое* $t_{факт}$ и *критическое* (табличное) $t_{табл}$ значения, принимаем или отвергаем гипотезу H_0:

при $t_{табл} < t_{факт}$ – отвергаем (т.е. a, b и r_{xy} под действием систематически действующего фактора x отличаются от 0 **не случайно**);

при $t_{табл} > t_{факт}$ – принимаем (т.е. a, b и r_{xy} формируются **случайно**).

Статистики Фишера и Стьюдента связаны соотношением $t_b^2 = t_r^2 = \sqrt{F}$.

Предельные ошибки для расчёта доверительных интервалов найдём по формулам $\Delta_a := t_{табл} m_a$, $\Delta_b := t_{табл} m_b$.

Для проверки **гипотезы о линейной зависимости** между СВ X и Y справедлива следующая процедура:

1. Вычислить t-статистику $t := r \cdot \sqrt{\frac{n-2}{1-r^2}}$.

2. По заданному уровню значимости α найти критическое значение t-статистики с (n–2) степенями свободы.

3. Если величина t превосходит критическое значение в положительную или отрицательную сторону, то нулевая гипотеза H_0:r=0 отклоняется, т.е. принимается *альтернативная* гипотеза H_1 о **наличии линейной положительной или отрицательной зависимости** соответственно.

Замечание: тест справедлив лишь для проверки нулевой гипотезы об отсутствии линейной зависимости; для проверки какой-либо другой гипотезы

(например, о равенстве коэффициента корреляции заданному ненулевому числу) требуется иная, более сложная процедура.

Для *парного* (!) регрессионного анализа *t*-критерий для нулевой гипотезы H_0:r_{xy}=0, *F*-критерий для гипотезы H_0:R^2=0 и *t*-критерий для H_0:β=0 эквивалентны.

2.13. Прогнозирование.

Прогноз $\tilde{y}_p$ определяется подстановкой прогнозируемого значения регрессора x_p в регрессионное уравнение $\tilde{y}_p := a + bx_p$. При этом **средняя стандартная ошибка прогноза** имеет вид

$$m_{\tilde{y}_p} := \sigma_{\text{ост}} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x - \bar{x})^2}}, \text{ где } \sigma_{ост} := \sqrt{\frac{\sum (y_x - \tilde{y}_x)^2}{n - m - 1}}.$$

Тогда **доверительный интервал прогноза** имеет вид

$$(\tilde{y}_p - \Delta_{\tilde{y}_p}; \tilde{y}_p + \Delta_{\tilde{y}_p}),$$

где $\Delta_{\tilde{y}_p} := t_{\text{табл}} \cdot m_{\tilde{y}_p}$.

3. Множественная регрессия.

Множественная регрессия $y=f(x_1,x_2,\ldots,x_k)$ – уравнение связи с несколькими регрессорами.

Например,

линейная модель $y = a + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k + \varepsilon$; (Плоскость регрессии является k-мерной гиперплоскостью в (k+1)-мерном пространстве.)

линейная модель без свободного члена $y = b_1 x_1 + b_2 x_2 + \ldots + b_k x_k + \varepsilon$;

степенная $y = a x_1^{b_1} x_2^{b_2} \ldots x_k^{b_k} \cdot \varepsilon$;

экспонента $y = e^{a + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k} \cdot \varepsilon$;

гипербола $y = \left(a + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k + \varepsilon\right)^{-1}$ и т.п.

Для *идентификации* неизвестных параметров множественной регрессии, *линейно зависящей от неизвестных параметров* вновь применим МНК. **Нормальная форма** системы линейных уравнений примет вид:

$$\begin{cases} an + b_1 \sum x_1 + b_2 \sum x_2 + \ldots + b_k \sum x_k = \sum y, \\ a \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2 + \ldots + b_k \sum x_k x_1 = \sum x_1 y, \\ \ldots \\ a \sum x_k + b_1 \sum x_1 x_k + b_2 \sum x_2 x_k + \ldots + b_k \sum x_k^2 = \sum x_k y. \end{cases}$$

Её решения по формулам Крамера: $a = \Delta_a / \Delta$, $b_1 = \Delta_{b_1} / \Delta$, …, $b_k = \Delta_{b_k} / \Delta$,

где определитель системы $\Delta := \begin{vmatrix} n & \sum x_1 & \sum x_2 & \cdots & \sum x_k \\ \sum x_1 & \sum x_1^2 & \sum x_2 x_1 & \cdots & \sum x_k x_1 \\ \sum x_2 & \sum x_1 x_2 & \sum x_2^2 & \cdots & \sum x_k x_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum x_k & \sum x_1 x_k & \sum x_2 x_k & \cdots & \sum x_k^2 \end{vmatrix} \neq 0$.

Для случая 2-х регрессоров x_1и x_2 можно преобразовать МНК-параметры к виду

$$b_1 = \frac{\mathrm{Cov}(x_1, y)\mathrm{Var}(x_2) - \mathrm{Cov}(x_2, y)\mathrm{Cov}(x_1, x_2)}{\mathrm{Var}(x_1)\mathrm{Var}(x_2) - \mathrm{Cov}^2(x_1, x_2)},\ b_2 = \frac{\mathrm{Cov}(x_2, y)\mathrm{Var}(x_1) - \mathrm{Cov}(x_1, y)\mathrm{Cov}(x_1, x_2)}{\mathrm{Var}(x_1)\mathrm{Var}(x_2) - \mathrm{Cov}^2(x_1, x_2)},$$

$$a = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2 .$$

Для большего числа регрессоров формулы для коэффициентов значительно усложняются, но формула свободного члена регрессии остаётся практически той же: $a = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 - \ldots - b_k\bar{x}_k$. Как и в случае парной регрессии, она гарантирует, что найденная **МНК-плоскость** линейной выборочной регрессии всегда проходит через **центр диаграммы рассеяния** $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k, \bar{y}) \in \Re^{k+1}$.

4. Условия применимости МНК.

4.1. Условия теоремы Гаусса-Маркова.

1. Случайный член регрессии ε_i в каждом наблюдении имеет нулевое мат. ожидание – $\forall i \in \overline{1,n} \Rightarrow M\varepsilon_i = 0$.

2. Дисперсия остатков постоянна (**гомоскедастичность**)

$$\forall i \in \overline{1,n} \Rightarrow \sigma^2_{\varepsilon_i} = \sigma^2_0 = \text{Const}.$$

Если это условие не выполняется, то найденные МНК-параметры будут *неэффективны*; в этом случае применяют обобщённый МНК.

3. **Автокорреляция остатков** отсутствует – $\forall i \neq j \Rightarrow \text{cov}(\varepsilon_i, \varepsilon_j) = 0$.

Если это условие не выполняется, то найденные МНК-параметры будут *неэффективны*; в этом случае допущена ошибка спецификации модели регрессии, необходимо выбрать другую модель регрессии, при которой автокорреляции в остатках не будет.

4. Случайный член распределён независимо от объясняющей переменной

$$\forall i \in \overline{1,n} \Rightarrow \text{cov}(x_i, \varepsilon_i) = 0.$$

5. Остаток регрессии ε_i имеет нормальное распределение.

Это предположение основано на **центральной предельной теореме**, утверждающей, что СВ распределена *асимптотически нормально*, если она является результатом взаимодействия большого числа СВ, ни одна из которых не доминирует над остальными.

4.2. Теорема Гаусса–Маркова.

Теорема Гаусса–Маркова. При выполнении условий 1-5 МНК-оценки параметров линейной регрессии несмещённые, эффективные, состоятельные.

Замечания:

1. При нарушении условий 2-3 сохраняются несмещённость и состоятельность, но эффективность теряется.

2. При нарушении условия 4 может нарушаться и несмещённость оценок.

Для *парной линейной* регрессии **стандартные ошибки параметров** *a* и *b* имеют вид: $s.d.(a)=\sqrt{\frac{s_\varepsilon^2}{n}\left(1+\frac{\bar{x}^2}{\mathrm{Var}(x)}\right)}$, $s.d.(b)=\sqrt{\frac{s_\varepsilon^2}{n\,\mathrm{Var}(x)}}$, где $s_\varepsilon^2 := \frac{n}{n-2}\mathrm{Var}(\varepsilon)$ – несмещённая оценка **дисперсии остатка** регрессии σ_ε^2.

Тогда можно сделать следующие выводы о **точности** параметров линейной регрессии:

1. МНК-оценка *a* генерального значения α нормально распределена с мат. ожиданием *a* и стандартной ошибкой $\sqrt{\frac{s_\varepsilon^2}{n}\left(1+\frac{\bar{x}^2}{\mathrm{Var}(x)}\right)}$.

2. МНК-оценка b генерального значения β нормально распределена с мат. ожиданием b и стандартной ошибкой $\sqrt{\frac{s_\varepsilon^2}{n\,\mathrm{Var}(x)}}$.

3. Точность указанных МНК-оценок растёт с ростом объёма выборки n, расширением диапазона наблюдений Var(x) и увеличением точности измерений (т.е. уменьшением s_ε^2).

Несмещённой оценкой дисперсии σ_ε^2 остатка модели множественной регрессии является $s_\varepsilon^2 := \frac{RSS}{n-k-1}$.

Для 2-х регрессоров x_1 и x_2 стандартные ошибки выборочных коэффициентов множественной линейной регрессии примут вид

$$s.d.(b_1) = \sqrt{\frac{s_\varepsilon^2}{n\,\mathrm{Var}(x_1)} \cdot \frac{1}{1-r_{x_1x_2}^2}},\ s.d.(b_2) = \sqrt{\frac{s_\varepsilon^2}{n\,\mathrm{Var}(x_2)} \cdot \frac{1}{1-r_{x_1x_2}^2}}. \qquad (4)$$

Отсюда следует, что точность оценок растёт с ростом объёма выборки n пропорционально $\sqrt{n}$. Зависимость от выборочного СКО $S := \sqrt{\mathrm{Var}(x_i)}$ **обратная пропорциональная**, причём вариация одного регрессора x_i не влияет на точность другого коэффициента b_j, $i \neq j$. Поэтому влиять на точность оценки каждого *коэффициента* (но не свободного члена регрессии!) можно, подбирая необходимую вариацию соответствующего регрессора $\mathrm{Var}(x_i) \to +\infty$. При этом выбор равномерной сетки, обычный для большинства экспериментов, не будет оптимальным при фиксированном объёме выборки n. Гораздо полезнее взять все значения на краях допустимого размаха каждого регрессора.

Очевидно, указанная рекомендация верна и в случае любого числа регрессоров, в том числе и для парной регрессии.

Минимальное число наблюдений n для однозначного определения гиперплоскости в (k+1)-мерном пространстве равно k+1. Например, в $\Re^3$ для определения плоскости нужны 3 точки, не лежащие на одной прямой. Следовательно, если $n>k+1$ появляются «лишних» ($n–k–1$) степени свободы, требующих оптимизации выбора МНК-гиперплоскости.

Если число степеней свободы мало, то статистическая надёжность выборочной оценки невысока. Так, проведённая через 3 точки плоскость в $\Re^3$ не будет проходить через следующее 4-е наблюдение, причём сколь угодно далеко. Обычно для обеспечения статистической надёжности выборочных оценок параметров множественной регрессии требуется, чтобы число наблюдений по крайней мере в 3 раза превосходило число оцениваемых параметров.

Тесты Стьюдента в множественном регрессионном анализе такие же, как и в парном. Критический уровень t-распределения зависит от $n–k–1$ степени свободы. Доверительные интервалы рассчитываются с учётом данного числа степеней свободы.

Зависимость стандартных ошибок (4) от коэффициента парной корреляции $r_{x_1x_2}$ регрессоров показывает неограниченный рост ошибки при сильной зависимости $|r_{x_1x_2}| \to 1$, возникающей при **коллинеарности** векторов наблюдений регрессоров. Следовательно, линейно зависимые вектора наблюдений регрессоров нельзя включать в одну линейную модель множественной регрессии. Так про-

является необходимость *отбора факторов* линейной регрессии на этапе, предшествующем этапу *спецификации* (а значит, и этапу *идентификации*) модели множественной регрессии. Такого этапа нет при моделировании парной регрессии.

Особое внимание среди **условий применимости МНК** (предпосылки теоремы Гаусса–Маркова) уделяется проверке гомоскедастичности и отсутствию автокорреляции в остатках.

4.3. Неэффективность МНК в случае гетероскедастичности.

В предпосылках теоремы Гаусса–Маркова требуется постоянство дисперсии случайного члена регрессии в каждом наблюдении (**гомоскедастичность остатков**).

Гетероскедастичность – это нарушение предположения о постоянстве дисперсии остатка $\sigma_{\varepsilon_i}^2 = \sigma_i^2 \neq \text{Const}$.

Хотя гетероскедастичность не влияет на несмещённость выборочных оценок параметров линейной регрессии, она приводит к их неэффективности. При гомоскедастичности, МНК-коэффициенты обладают наименьшей дисперсией среди всех несмещённых оценок, являющихся линейными функциями наблюдений y_i.

Обычный МНК не делает различия между качеством отдельных наблюдений, в то время как наблюдения с малой дисперсией обладают заведомо мень-

шей погрешностью по сравнению с наблюдениями с большой дисперсией. Придавая разные веса таким наблюдениям можно повысить общую точность результата применения МНК, а значит, получить оценку с меньшей дисперсией. Это и означает потерю эффективности оценок, полученных обычным МНК при гетероскедастичности остатков.

Экономической причиной гетероскедастичности часто является разномасштабность наблюдений, например, при исследовании временных рядов со значительным трендом. Такими являются, к примеру, макроэкономические показатели на больших временны***х*** интервалах исследования.

Установлению гетероскедастичности служат тест ранговой корреляции Спирмена, тест Голдфелда–Квандта, тест Глейзера и др.

4.3.1. Тест ранговой корреляции Спирмена.

Ранг наблюдения переменной – это номер наблюдения в ранжированной по возрастанию последовательности.

Тест ранговой корреляции Спирмена позволяет обнаружить гетероскедастичность в виде нестрогой линейной зависимости стандартного отклонения остатка от объясняющей переменной. При этом предполагается увеличение либо уменьшение дисперсии остатка при увеличении регрессора *x*.

Коэффициент ранговой корреляции для теста Спирмена

$$r_{x\varepsilon} := 1 - \frac{6\sum_{i=1}^{n} D_i}{n(n-1)} \qquad (5)$$

связывает регрессор x, с которым по предположению коррелирует остаток ε, и этот остаток. Здесь D_i – разность между рангом i-го наблюдения x_i и рангом модуля остатка $|\varepsilon_i|$ при этом наблюдении.

При отсутствии гетероскедастичности $r_{x\varepsilon}=0$ для генеральной совокупности, что влечёт нормальное распределение коэффициента ранговой корреляции Спирмена (5) с мат. ожиданием 0 и дисперсией $1/(n-1)$ при большом n.

Следовательно, соответствующая тестовая статистика равна $r_{x\varepsilon}\sqrt{n-1}$, то есть при *двустороннем* критерии на уровне значимости $\alpha=0{,}05$ нулевая гипотеза об отсутствии гетероскедастичности будет отклонена, если тестовая статистика превысит 1,96.

Для $\alpha=0{,}01$ соответствующее критическое значение равно 2,58.

При множественной регрессии проверка гипотезы может быть проведена с любым регрессором.

4.3.2. Тест Голдфелда–Квандта.

Используется наиболее часто. Предполагается, что стандартное отклонение остатка σ_{ε_i} прямо (или обратно) пропорционально значению x_i в этом наблюдении. Также предполагается, что случайный член ε_i распределён нормально и не подвержен автокорреляции.

Алгоритм теста Голдфелда–Квандта:

1. Все наблюдения в выборке упорядочиваются по возрастанию регрессора *x*.

2. Извлекаются первые и последние $n' \approx n/3$, $n' > k+1$, наблюдений, для которых рассчитываются остаточные суммы квадратов отклонений RSS_1 и RSS_2 соответственно; k – число регрессоров.

3. Для отношения RSS_2/RSS_1 проводят тест Фишера с $(n'-k-1)$ верхними и нижними степенями свободы.

Если RSS_2/RSS_1 попадает за пределы области принятия нулевой гипотезы, то принимается решение о наличии гетероскедастичности.

Для проверки гипотезы об *обратной пропорциональной* зависимости от *x*, аналогично рассматривается отношение RSS_1/RSS_2.

Если в модели более одного регрессора, то наблюдения упорядочиваются по той из них, от которой ожидается более чёткая зависимость дисперсии σ_ε^2.

4.3.3. Тест Глейзера.

Тест Глейзера наиболее точно улавливает связь (в т.ч. нелинейную) между **стандартным отклонением остатка регрессии** и объясняющей переменной. При этом ищется зависимость вида $\sigma_i = \alpha + \beta x_i^\gamma$.

Вначале обычным МНК оценивается регрессия *y* на *x*, а затем идентифицируются параметры α и β при фиксированных значениях параметра γ для пар-

ной линейной регрессии $|\varepsilon_i| = \alpha + \beta x_i^{\gamma}$. В каждом случае нулевая гипотеза об **отсутствии гетероскедастичности** будет отклонена, если выборочная оценка коэффициента β значимо отличается от 0.

Выбирается такое значение показателя γ, при котором значимость отклонения оценки β от 0 выше.

4.4.1. Последствия автокорреляции.

При **автокорреляции** $\mathrm{cov}(\varepsilon_i, \varepsilon_j) \neq 0$ при $i \neq j$. Последствия автокорреляции сходны с последствиями гетероскедастичности: оценки параметров регрессии остаются несмещёнными, но становятся неэффективными, их стандартные ошибки считаются неправильно (как правило, занижаются).

Обычно, автокорреляция встречается в остатках *временных рядов*.

Положительная автокорреляция соответствует случаю $r_{\varepsilon_k \varepsilon_{k+1}} \approx \mathrm{Const} > 0$.

Причиной положительной автокорреляции чаще всего является воздействие не включённых в уравнение регрессии объясняющих переменных, которые **сонаправленно** искажают случайный член регрессии в разные моменты времени.

Автокорреляция представляет тем большую проблему, чем меньше интервал между наблюдениями, поскольку при увеличении этого интервала менее вероятно сохранение однонаправленного искажения замеров.

На графике остатков парной регрессии наличие положительной автокорреляции отмечается появлением **долгопериодных (=низкочастотных) квазипериодических колебаний**. Визуально такие колебания обнаруживаются достаточно легко.

Отрицательная автокорреляция соответствует случаю $r_{\varepsilon_k \varepsilon_{k+1}} \approx \text{Const} < 0$.

В экономике встречается сравнительно редко, хотя в теории автоматического управления, из которой эконометрика заимствовала многие методы и модели, считается предпочтительной характеристикой практически осуществимых **систем с обратной связью**.

На графике остатков парной регрессии наличие отрицательной автокорреляции отмечается появлением **короткопериодных (=высокочастотных**) квазипериодических колебаний. В отличие от низкочастотных колебаний визуально обнаружить такие колебания сложно, поэтому остаётся лишь аналитический метод обнаружения таких автокорреляций (например, критерий Дарбина–Уотсона).

4.4.2. Автокорреляция первого порядка.

Автокорреляция 1-го порядка – это наличие корреляции в соседних наблюдениях: $r_{\varepsilon_k \varepsilon_{k+1}} \approx \text{Const} \neq 0$.

Авторегрессионная модель 1-го порядка – это частный случай автокорреляции 1-го порядка с парной линейной зависимостью соседних остатков $\varepsilon_{k+1}=\rho\varepsilon_k+\delta_k$, где ρ – константа, δ_k – новый случайный член новой регрессии.

Обычным МНК получается **выборочная оценка коэффициента авторегрессии** $\hat{\rho} := \frac{Cov(\varepsilon_{k-1}, \varepsilon_k)}{Var(\varepsilon_{k-1})} = \frac{\sum_{t=2}^{k} \varepsilon_{t-1}\varepsilon_t}{\sum_{t=1}^{k} \varepsilon_t^2}$.

Если ρ=0, то соответствующее условие теоремы Гаусса–Маркова выполнено.

Если ρ>0, то имеется **положительная автокорреляция остатков 1-го порядка**; если же ρ<0, то **автокорреляция 1-го порядка отрицательная**.

Критерий Дарбина–Уотсона – метод обнаружения автокорреляции 1-го порядка с помощью **статистики Дарбина–Уотсона**

$$DW := \frac{\sum_{k=2}^{n} (\varepsilon_k - \varepsilon_{k-1})^2}{\sum_{k=1}^{n} \varepsilon_k^2}.$$

Условие применимости критерия Дарбина–Уотсона: отсутствие **лаговых переменных** в регрессии.

Так как $|\rho| \le 1$, то $0 \le d \le 4$, где d – **расчётное значение критерия** DW. При этом $\rho \approx 0 \Leftrightarrow d \approx 2$.

1. При наличии **положительной** автокорреляции ρ>0 $\Rightarrow$ 0<d<2 и чем положительная автокорреляция сильнее (т.е. $\rho \to 1$), тем значение d ближе к 0.

2. При **отрицательной** автокорреляции ρ<0 $\Rightarrow$ 2<d<4 и чем отрицательная автокорреляция сильнее (т.е. $\rho \to -1$), тем значение d ближе к 4.

Точно определить **критические значения** $d_{крит}$ статистики *DW* не удаётся, поэтому в таблице *DW*-распределения даны лишь их **нижние** d_L и **верхние** d_U **оценки**: $d_L<d_{крит}<d_U$**.** Таблицы, приводимые в литературе, дают значения этих оценок лишь для проверки гипотезы о наличии *положительной* автокорреляции. Для исследования *отрицательной* автокорреляции можно воспользоваться симметрией расположения критического значения и его оценок относительно 2.

В зависимости от величины расчётного значения *d* возможны следующие случаи.

1. Если $d<d_L$, то $d<d_{крит}$. Следовательно, можно сделать вывод о **наличии положительной автокорреляции остатков**.

2. Если $d>4-d_L$, то $d>d_{крит}$. Следовательно, можно сделать вывод о **наличии отрицательной автокорреляции остатков**.

3. Если $d_U<d<4-d_U$, то $d>d_{крит}$ при $d<2$ и $d<d_{крит}$ при $d>2$. В первом случае можно сделать вывод об **отсутствии** положительной, а во втором – об **отсутствии** отрицательной автокорреляции остатков.

4. Если *d* попадает в **зону неопределённости критерия**

$$[d_L, d_U] \cup [4-d_U, 4-d_L],$$

то достаточных оснований принять или отвергнуть гипотезу о наличии автокорреляции в остатках нет. (На практике это означает, что недостаточен объём *n* наблюдений.)

Наилучший (хотя и не всегда возможный) способ устранения автокорреляции в остатках – установление экономического фактора, ответственного за неё, и включение соответствующего регрессора в модель.

4.5. Общий и частный *F*-критерии Фишера.

Значимость уравнения множественной регрессии **в целом** оценивается **общим** *F*-критерием Фишера $F := \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m}$.

Частный *F*-критерий $F_{x_i} := \frac{R^2_{yx_1\ldots x_{i-1}x_ix_{i+1}\ldots x_k} - R^2_{yx_1\ldots x_{i-1}x_{i+1}\ldots x_k}}{1-R^2_{yx_1\ldots x_{i-1}x_ix_{i+1}\ldots x_k}} \cdot \frac{n-m-1}{1}$ позволяет оценить статистическую значимость **вклада в уравнение регрессии фактора** x_i.

4.6. Значимость выборочного уравнения и значимость выборочных оценок коэффициентов множественной регрессии.

Оценка значимости коэффициентов регрессии с помощью *t*-критерия Стьюдента сводится к вычислению **статистики** $t_{b_i} := b_i / m_{b_i} = \sqrt{F_{x_i}}$, где **средняя квадратическая ошибка коэффициента регрессии** b_i определяется формулой

$$m_{b_i} := \frac{\sigma_y \sqrt{1-R^2_{yx_1\ldots x_k}}}{\sigma_{x_i}\sqrt{1-R^2_{x_ix_1\ldots x_{i-1}x_{i+1}\ldots x_k}}} \cdot \frac{1}{\sqrt{n-m-1}}.$$

4.7. Мультиколлинеарность.

Строгая линейная зависимость между переменными – это случай равенства выборочной корреляции +1 (**сонаправленное** изменение) или –1 (**противоположное**, **обратное** изменение) переменных.

При **нестрогой линейной зависимости** между переменными теоретическая корреляция близка по модулю к единице: $\left|r_{x_i x_j}\right| \approx 1$.

Считается, что **две** случайные величины *сильно коллинеарны*, т.е. связаны **сильной линейной зависимостью**, если $\left|r_{x_i x_j}\right| \geq 0{,}7$.

Мультиколлинеарность – это сильная линейная зависимость между объясняющими переменными, приводящая к получению ненадёжных выборочных оценок параметров множественной линейной регрессии. В модели с 2-мя регрессорами это проявляется множителем $\left(1 - r^2_{x_1 x_2}\right)^{-1}$ в стандартной ошибке выборочных оценок коэффициентов линейной регрессии.

Мультиколлинеарность в той или иной степени присутствует в каждой модели множественной регрессии, но наиболее негативно проявляется при полной (строгой) (мульти)коллинеарности, когда применение МНК становится невозможным.

Для оценки **мультиколлинеарности**, т.е. линейной зависимости большего числа случайных величин, рассматривают определитель **матрицы парных коэффициентов линейной корреляции**. Чем определитель **матрицы межфак-**

торной корреляции ближе к 0, тем сильнее мультиколлинеарность. И наоборот, чем он ближе к 1, тем слабее мультиколлинеарность.

Более точно мультиколлинеарность можно проверить с помощью нулевой гипотезы $H_0 : \det\|R\| = 1$, так как известно, что случайная величина

$$\left[n - 1 - \frac{2m+5}{6} \ln \det\|R\| \right]$$

имеет приближённое распределение χ^2 с $n(n-1)/2$ степенями свободы.

Если фактическое значение статистики больше критического (табличного) $\chi^2_{\text{факт}} > \chi^2_{\text{табл}}(df, \alpha)$, то гипотеза H_0 отклоняется. Следовательно, $\det\|R\| \neq 1$ и недиагональные ненулевые коэффициенты корреляции означают коллинеарность регрессоров. Тем самым мультиколлинеарность доказана.

Строгая линейная зависимость между регрессорами может происходить из-за теоретической линейной зависимости и случайно, в силу свойств полученной выборки.

В первом случае линейная зависимость присутствует во всех выборках данной генеральной совокупности. Во втором случае при переходе к другой выборке она может превратиться в нестрогую зависимость. Однако, применение МНК невозможно в обоих случаях.

Для устранения полной (строгой) коллинеарности нужно уменьшить число объясняющих переменных, устраняя их из модели до исчезновения строгой зависимости. Для определения **попарно связанных регрессоров** можно использовать матрицу парных корреляций, пересчитывая её заново каждый раз после

устранения очередного регрессора, имеющего максимальные по модулю коэффициенты корреляции.

При нестрогой мультиколлинеарности выборочные оценки могут оказаться вполне удовлетворительными, если число наблюдений *n* и вариации объясняющих переменных Var(x_i) велики, а дисперсия случайного остатка s_ε^2 мала.

Два варианта **отбора регрессоров**:

1. Удаление лишних переменных из первоначального (максимального) множества объясняющих переменных. Проверка величины уменьшения объяснённой суммы квадратов отклонений *ESS*.

2. Добавление необходимых переменных к первоначальному (минимальному) множеству объясняющих переменных. Проверка величины роста объяснённой суммы квадратов отклонений *ESS*.

4.8. Стандартизованное уравнение регрессии.

Для ранжирования значимости регрессоров x_i перейдём к уравнению множественной регрессии в **стандартизованном масштабе** $t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2} + \ldots + \beta_k t_{x_k}$, где $t_y := \frac{y_x - \bar{y}}{\sigma_y}$, $t_{x_i} := \frac{x_i - \bar{x}_i}{\sigma_{x_i}}$ – **стандартизованные переменные**, β_i – **стандартизованные коэффициенты** линейной регрессии.

Применение МНК для идентификации стандартизованных коэффициентов даёт следующую систему линейных уравнений в нормальной форме:

$$\begin{cases} \beta_1 + \beta_2 r_{x_2x_1} + \beta_3 r_{x_3x_1} + \ldots + \beta_k r_{x_kx_1} = r_{yx_1}, \\ \beta_1 r_{x_1x_2} + \beta_2 + \beta_3 r_{x_3x_2} + \ldots + \beta_k r_{x_kx_2} = r_{yx_2}, \\ \ldots \\ \beta_1 r_{x_1x_k} + \beta_2 r_{x_2x_k} + \beta_3 r_{x_3x_k} + \ldots + \beta_k = r_{yx_k}. \end{cases}$$

Коэффициенты b_i линейной множественной регрессии в исходных (**натуральных**) показателях связаны со стандартизованными коэффициентами β_i соотношением $b_i = \beta_i \sigma_y / \sigma_{x_i}$. Свободный член a уравнения в натуральных показателях определяется как $a = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 - \ldots - b_k\bar{x}_k$.

5.1. Индекс множественной корреляции.

Тесноту совместного влияния регрессоров на результат при *нелинейной* регрессии оценивает **индекс множественной корреляции**

$$R_{yx_1x_2\ldots x_k} := \sqrt{1 - \sigma^2_{y\,\text{ост}} / \sigma^2_y}\ ,\ 0 \le R_{yx_1x_2\ldots x_k} \le 1.$$

При этом справедлива **нижняя оценка** $R_{yx_1x_2\ldots x_k} \ge \max\limits_{1\le i\le k} r_{yx_i}$, т.е. индекс множественной корреляции мажорирует коэффициенты парной линейной корреляции между регрессорами x_i и результирующим признаком y.

Через стандартизованные коэффициенты и парные коэффициенты линейной корреляции его можно выразить в виде $R_{yx_1x_2\ldots x_k} = \sqrt{\sum \beta_i r_{yx_i}}$.

5.2. Коэффициент множественной корреляции.

При *линейной* множественной регрессии индекс превращается в **коэффициент множественной линейной корреляции** $R_{yx_1x_2\ldots x_k} := \sqrt{1-\Delta_r/\Delta_{r_{11}}}$, где

$$\Delta_r := \begin{vmatrix} 1 & r_{yx_1} & r_{yx_2} & \cdots & r_{yx_k} \\ r_{yx_1} & 1 & r_{x_2x_1} & \cdots & r_{x_kx_1} \\ r_{yx_2} & r_{x_1x_2} & 1 & \cdots & r_{x_kx_2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{yx_k} & r_{x_1x_k} & r_{x_2x_k} & \cdots & 1 \end{vmatrix}$$ – определитель матрицы парных коэффициентов линейной корреляции;

$$\Delta_{r_{11}} := \begin{vmatrix} 1 & r_{x_1x_2} & r_{x_1x_3} & \cdots & r_{x_1x_k} \\ r_{x_2x_1} & 1 & r_{x_2x_3} & \cdots & r_{x_2x_k} \\ r_{x_3x_1} & r_{x_3x_2} & 1 & \cdots & r_{x_3x_k} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{x_kx_1} & r_{x_kx_2} & r_{x_kx_3} & \cdots & 1 \end{vmatrix}$$ – определитель матрицы парной межфакторной линейной корреляции.

5.3. Частный индекс множественной корреляции.

Частный индекс множественной корреляции

$$r_{yx_i\bullet x_1x_2\ldots x_{i-1}x_{i+1}\ldots x_k} := \pm\sqrt{1-\frac{1-R^2_{yx_1x_2\ldots x_{i-1}x_ix_{i+1}\ldots x_k}}{1-R^2_{yx_1x_2\ldots x_{i-1}x_{i+1}\ldots x_k}}},\ -1 \le r_{yx_i\bullet x_1x_2\ldots x_{i-1}x_{i+1}\ldots x_k} \le 1,$$

измеряет влияние на *y* регрессора x_i при неизменном уровне других регрессоров.

Рекуррентное определение:

$$r_{yx_i\bullet x_1x_2\ldots x_{i-1}x_{i+1}\ldots x_k} := \frac{r_{yx_i\bullet x_1x_2\ldots x_{i-1}x_{i+1}\ldots x_{k-1}} - r_{yx_k\bullet x_1x_2\ldots x_{i-1}x_{i+1}\ldots x_{k-1}} \times r_{x_ix_k\bullet x_1x_2\ldots x_{i-1}x_{i+1}\ldots x_{k-1}}}{\sqrt{\left(1-r^2_{yx_k\bullet x_1x_2\ldots x_{k-1}}\right)\cdot\left(1-r^2_{x_ix_k\bullet x_1x_2\ldots x_{i-1}x_{i+1}\ldots x_{k-1}}\right)}}.$$

Качество построенной модели в целом оценивает **коэффициент (индекс) множественной детерминации**, равный квадрату коэффициента (индекса) множественной *корреляции* $R^2_{yx_1x_2\ldots x_k}$.

Скорректированный (исправленный) индекс (коэффициент) множественной детерминации содержит поправку на число степеней свободы:

$$\tilde{R}^2 := 1-\left(1-R^2\right)\frac{n-1}{n-m-1} = R^2 - \frac{m}{n-m-1}\left(1-R^2\right),$$

где n – объём выборки, m – число регрессоров.

Скорректированный (adjusted) коэффициент детерминации $\tilde{R}^2$ компенсирует рост коэффициента детерминации R^2 при добавлении объясняющих факторов путём наложения «штрафа» за увеличение числа степеней свободы. По мере роста m отношение $m/(n–m–1)$ увеличивается, и, следовательно, возрастает коррекция коэффициента детерминации R^2 в сторону уменьшения.

В пределе $\lim\limits_{m\to+\infty}\dfrac{m}{n-m-1} = -1 \;\Rightarrow\; \lim\limits_{m\to+\infty}\tilde{R}^2 = R^2 + 1\cdot\left(1-R^2\right) = 1.$

С другой стороны, $\lim\limits_{n\to+\infty}\dfrac{m}{n-m-1} = 0 \;\Rightarrow\; \lim\limits_{n\to+\infty}\tilde{R}^2 = R^2 + 0\cdot\left(1-R^2\right) = R^2.$

Известно, что добавление нового регрессора приведёт к росту $\tilde{R}^2$ тогда и только тогда, когда соответствующая t-статистика по модулю больше 1. Следовательно, рост $\tilde{R}^2$ при этом не обязательно означает, что коэффициент при добавленной переменной значим. Поэтому вопреки ожиданиям увеличение $\tilde{R}^2$ *не обязательно* означает улучшение спецификации уравнения модели регрессии.

По этой причине исправленный коэффициент детерминации не стал распространённой **диагностической характеристикой**. Другая причина заключа-

ется в общем уменьшении внимания к коэффициенту детерминации R^2. Ранее коэффициент R^2 рассматривался как основной индикатор **качества спецификации модели.** Однако практика показала, что иногда плохо определённая модель регрессии может дать высокое значение коэффициента детерминации.

В настоящее время R^2 (а значит, и его поправка $\tilde{R}^2$) рассматривается лишь как один из ряда *диагностических показателей*, проверка которых необходима для оценки качества модели регрессии, причём не самый важный показатель. Следовательно, и корректировка данного коэффициента мало что даёт.

6. Замещающие переменные.

Замещающая переменная – это объясняющий регрессор, используемый вместо трудноизмеримой, но важной переменной.

Она требуется, если отсутствие важной переменной вызывает смещение выборочных оценок параметров регрессии или если включение замещающей переменной может дать косвенную информацию о замещённой переменной.

7. Фиктивные переменные.

Фиктивные переменные включают в регрессию, если необходимо оценить влияние *дискретных* факторов.

В модели множественной регрессии всегда желательно присутствие хотя бы одной **нефиктивной** переменной, т.к. дисперсия фиктивной переменной мала, что отрицательно сказывается на достоверности оценок.

В модели с фиктивными переменными коэффициент детерминации R^2 часто очень мал, а значения *t*-статистики для коэффициентов при фиктивных переменных незначимо отличаются от 0. Однако, исключать фиктивные переменные из модели не нужно, т.к. они описывают небольшие, но важные поправки к **главной** (нефиктивной) объясняющей переменной.

Категория – событие, про которое в каждом наблюдении можно сказать определённо – произошло оно или нет.

Набор категорий – конечное число взаимоисключающих событий, исчерпывающих все возможности.

Совокупность фиктивных переменных – некоторое число фиктивных переменных, достаточное для описания всего набора категорий. Таким образом, по набору значений фиктивных переменных однозначно определяется одна категория.

Эталонная категория – категория, с которой сравниваются все другие категории (обычно в этой категории все фиктивные переменные принимаются равными 0).

Сезонные фиктивные переменные – совокупность фиктивных переменных для обозначения различных лет, месяцев, времён года, кварталов и т.п.

Пример **поквартального описания**: $y=\alpha+\beta t+\delta_2 D_2+\delta_3 D_3+\delta_4 D_4+\varepsilon$, где D_i – фиктивные переменные, принимающие значение 1, если наблюдение относится к i-му кварталу, или 0 – если оно относится к другому кварталу, i=2,3,4.

Отметим, что четвёртая фиктивная переменная D_1, относящаяся к 1-му кварталу не вводится, т.к. имеется *строгая мультиколлинеарная зависимость* $D_1+D_2+D_3+D_4=1$. Такая ситуация называется **ловушкой dummy trap** – выбор совокупности фиктивных переменных, сумма которых тождественно равна константе.

Для определения влияния категории на коэффициент линейной регрессии при нефиктивной переменной вводится **фиктивная переменная для коэффициента наклона** – произведение фиктивной переменной, отвечающей за данную категорию, на соответствующую нефиктивную переменную.

Коэффициент регрессии при фиктивной переменной интерпретируется как среднее изменение зависимой переменной при переходе от одного значения дискретного регрессора к другому. На основе t-критерия Стьюдента делается вывод о значимости влияния фиктивной переменной, т.е. существенности расхождения между категориями.

8. Сравнение стандартизованного и натурального уравнений множественной регрессии.

Пример. Линейное уравнение множественной регрессии y от x_1 и x_2 в натуральных показателях имеет вид $y=a+b_1x_1+b_2x_2$.

Перейдём к стандартизованному уравнению $t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2}$, коэффициенты которого найдём по формулам

$$\beta_1 = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{1 - r^2_{x_1x_2}},\ \beta_2 = \frac{r_{yx_2} - r_{yx_1} r_{x_1x_2}}{1 - r^2_{x_1x_2}}.$$

По стандартизованным β-коэффициентам можно определить МНК-оценки параметров исходного уравнения регрессии:

$$b_1 = \beta_1 \sigma_y / \sigma_{x_1},\ b_2 = \beta_2 \sigma_y / \sigma_{x_2},\ a = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2.$$

Линейные коэффициенты частной корреляции найдём по формулам

$$r_{yx_1 \bullet x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{\sqrt{\left(1 - r^2_{yx_2}\right)\left(1 - r^2_{x_1x_2}\right)}},\ r_{yx_2 \bullet x_1} = \frac{r_{yx_2} - r_{yx_1} r_{x_1x_2}}{\sqrt{\left(1 - r^2_{yx_1}\right)\left(1 - r^2_{x_1x_2}\right)}},\ r_{x_1x_2 \bullet y} = \frac{r_{x_1x_2} - r_{yx_1} r_{yx_2}}{\sqrt{\left(1 - r^2_{yx_1}\right)\left(1 - r^2_{yx_2}\right)}}.$$

Расчёт линейного коэффициента множественной корреляции с использованием коэффициентов r_{yx_i} и β_i даёт формулу $R_{yx_1x_2} = \pm\sqrt{\beta_1 r_{yx_1} + \beta_2 r_{yx_2}}$.

Общий F-критерий

$$F_{\text{факт}} = \frac{R^2_{yx_1x_2}}{1 - R^2_{yx_1x_2}} \cdot \frac{n - m - 1}{m}$$

проверяет нулевую гипотезу H_0 о статистической значимости уравнения регрессии в целом и значимости показателя тесноты связи $R^2 \neq 0$. Сравнивая $F_{\text{табл}}$ и $F_{\text{факт}}$ при заданном уровне значимости α можно сделать вывод о статистической

(не)значимости уравнения регрессии в целом и показателя тесноты связи $R_{yx_1x_2}$, сформированных неслучайным воздействием регрессоров x_1 и x_2.

Частные *F*-критерии

$$F_{x_1\,\text{факт}} = \frac{R^2_{yx_1x_2} - r^2_{yx_2}}{1 - R^2_{yx_1x_2}} \cdot \frac{n-m-1}{1},\ F_{x_2\,\text{факт}} = \frac{R^2_{yx_1x_2} - r^2_{yx_1}}{1 - R^2_{yx_1x_2}} \cdot \frac{n-m-1}{1}$$

оценивают статистическую значимость присутствия регрессоров x_1 и x_2 в уравнении множественной линейной регрессии, т.е. оценивают **целесообразность включения в уравнение одного фактора после другого**. Если $F_{x_i\,\text{факт}} > F_{\text{табл}}$, то включение переменной x_i в модель целесообразно, т.к. приводит к заметному росту $R^2_{yx_1x_2}$. В противном случае, неравенство $F_{x_i\,\text{факт}} < F_{\text{табл}}$ подтверждает нулевую гипотезу H_0 о нецелесообразности включения фактора x_i после включения другого.

Задача дисперсионного анализа состоит в проверке нулевой гипотезы H_0 о статистической незначимости уравнения регрессии в целом и показателя тесноты связи сравнением фактического $F_{\text{факт}}$ и табличного (критического) $F_{\text{табл}}$ значений критерия. **Фактическое значение статистики** равно отношению удельных (рассчитанных на 1 степень свободы) факторной и остаточной дисперсий:

$$F_{\text{факт}} := \frac{\sum\left(\tilde{y}_{x_1x_2} - \bar{y}\right)^2}{m} : \frac{\sum\left(y_{x_1x_2} - \tilde{y}_{x_1x_2}\right)^2}{n-m-1} = \frac{S_{\text{факт}}}{S_{\text{ост}}} \cdot \frac{n-m-1}{m},$$

где n – объём выборки, m – число регрессоров.

Можно выразить: $S_{\text{факт}} = \sigma_y^2 \cdot n \cdot R^2_{yx_1x_2}$, $S_{\text{ост}} = \sigma_y^2 \cdot n \cdot \left(1 - R^2_{yx_1x_2}\right)$, $S_{\text{общ}} = \sigma_y^2 \cdot n$.

Частный *F*-критерий Фишера оценивает **статистическую целесообразность включения в модель регрессора** x_i после того, как в неё включены ос-

тальные факторы. Он равен отношению удельного (в расчёте на 1 степень свободы) прироста факторной (регрессионной, объяснённой) дисперсии за счёт дополнительно включённого регрессора x_i к удельной остаточной дисперсии модели, содержащей все регрессоры. Например, для *двухфакторной* множественной регрессии:

$$F_{x_1} := \frac{S_{\text{факт}_{yx_1x_2}} - S_{\text{факт}_{yx_2}}}{1} : \frac{S_{\text{ост}_{yx_1x_2}}}{n-m-1} = \frac{S_{\text{факт}_{yx_1x_2}} - S_{\text{факт}_{yx_2}}}{S_{\text{ост}_{yx_1x_2}}} \cdot (n-m-1),$$

$$F_{x_2} := \frac{S_{\text{факт}_{yx_1x_2}} - S_{\text{факт}_{yx_1}}}{1} : \frac{S_{\text{ост}_{yx_1x_2}}}{n-m-1} = \frac{S_{\text{факт}_{yx_1x_2}} - S_{\text{факт}_{yx_1}}}{S_{\text{ост}_{yx_1x_2}}} \cdot (n-m-1),$$

$$S_{\text{факт}_{yx_1x_2}} = \sigma_y^2 \cdot n \cdot R^2_{yx_1x_2},\ S_{\text{ост}_{yx_1x_2}} = \sigma_y^2 \cdot n \cdot \left(1 - R^2_{yx_1x_2}\right),\ S_{\text{общ}_{yx_1x_2}} = S_{\text{факт}_{yx_1x_2}} + S_{\text{ост}_{yx_1x_2}} = \sigma_y^2 \cdot n,$$

$$S_{\text{факт}_{yx_2}} = \sigma_y^2 \cdot n \cdot r^2_{yx_2},\ S_{\text{факт}_{yx_1}} = \sigma_y^2 \cdot n \cdot r^2_{yx_1} = S_{\text{факт}_{yx_1x_2}} - S_{\text{факт}_{yx_2}}.$$

Другой вариант вычисления:

$$F_{x_1} := \frac{R^2_{yx_1x_2} - r^2_{yx_2}}{1 - R^2_{yx_1x_2}} \cdot (n-m-1),\ F_{x_2} := \frac{R^2_{yx_1x_2} - r^2_{yx_1}}{1 - R^2_{yx_1x_2}} \cdot (n-m-1).$$

Оценка значимости коэффициентов линейной регрессии b_i с помощью *t*-критерия Стьюдента связана с трудоёмким вычислением их случайных ошибок m_{b_i}. Более простой путь – расчёт по формуле $t_{b_i} = \sqrt{F_{x_i}}$. Табличные (критические) значения *t*-критерия зависят от уровня значимости α и числа степеней свободы (n–m–1), где n – объём выборки, m – число регрессоров.

9. Отбор факторов, спецификация и идентификация моделей множественной и парной регрессии.

Лишняя переменная – это объясняющая переменная, включение которой в модель множественной регрессии не даёт улучшения объяснений экономических свойств модели.

Отсутствующая переменная – необходимая по экономическим причинам объясняющая переменная, не учтённая в математической модели.

Спецификация переменных – это отбор необходимых объясняющих переменных, обеспечивающий максимальную адекватность математической модели имеющимся наблюдениям.

В результате *неправильной спецификации* переменных:

1. При отсутствии необходимой переменной выборочные оценки параметров регрессии как правило (хотя и не всегда!) смещены:

– оценка b_1 смещена на слагаемое $\beta_2 \mathrm{Cov}(x_1,x_2)/\mathrm{Var}(x_1)$;

– оценка b_2 смещена на слагаемое $\beta_1 \mathrm{Cov}(x_1,x_2)/\mathrm{Var}(x_2)$.

Оценка b_i может быть как завышенной, так и заниженной в зависимости от знаков соответствующих коэффициентов генеральной линейной регрессии и ковариаций регрессоров. Расчётные значения стандартных ошибок b_i и соответствующие t-тесты некорректны.

Оценка может остаться несмещённой только если $\mathrm{Cov}(x_1,x_2)=0$.

Смещение было бы нулевым и в случае β_i=0, но в этом случае *неправильной спецификации не возникает*.

2. Включение лишней переменной не приводит к смещению выборочных оценок, но обычно (хотя и не всегда!) делает их неэффективными. Расчётные значения их стандартных ошибок приемлемы, но из-за неэффективности **заведомо завышены**.

Спецификация формы модели – выбор аналитической зависимости объясняемой переменной (y) от объясняющих (x_i).

10. Временные ряды.

Временной ряд – это совокупность наблюдений какого-либо показателя $x(t_1)$, $x(t_2)$, …, $x(t_N)$ за несколько последовательных **моментов** или **периодов** времени.

Каждый его **уровень** формируется из **трендовой** (T), **циклической** (S) и случайной (E) компонент.

Аддитивная модель: $Y:=T+S+E$; **мультипликативная** модель: $Y=T\times S\times E$.

10.1. Стационарность и нестационарность временного ряда.

Интуитивное представление – ряд имеет постоянное среднее значение и постоянную дисперсию. Однако необходимо учесть и менее очевидную внут-

реннюю связь между наблюдениями временного ряда в разные моменты времени. Поэтому необходимо добавить требование **постоянства автокорреляционной функции по времени.**

Ряд $x(t)$ называется **строго стационарным** (**сильно стационарным**, стационарным **в узком смысле**), если совместное распределение вероятностей m наблюдений $x(t_1)$, $x(t_2)$, …, $x(t_m)$ такое же, как и совместное распределение наблюдений $x(t_1+\tau)$, $x(t_2+\tau)$, …, $x(t_m+\tau)$ при любых m, t_1, t_2, …, t_m, τ.

Таким образом, свойства строго стационарного временного ряда не меняются при изменении **начала отсчёта времени**. В частности, при m=1 закон распределения $x(t)$ не зависит от t, а значит, не зависят от t мат. ожидание $Mx(t)=a$=Const и дисперсия $Dx(t)=M(x(\mathrm{t})-a)^2=\sigma^2$=Const.

Выборочными оценками этих моментов являются **выборочное среднее**

$$\hat{a} := \frac{1}{N}\sum_{t=1}^{N} x(t)$$

и **выборочная дисперсия**

$$\hat{\sigma}^2 := \frac{1}{N}\sum_{t=1}^{N}\left(x(t)-\hat{a}\right)^2 = \frac{1}{N}\sum_{t=1}^{N} x^2(t) - \hat{a}^2$$

или **исправленная выборочная дисперсия**

$$s^2 := \frac{1}{N-1}\sum_{t=1}^{N}\left(x(t)-\hat{a}\right)^2 = \frac{N}{N-1}\hat{\sigma}^2.$$

Проверка *строгой* стационарности на практике невозможна. Поэтому вводится ослабленное с точки зрения математической теории, но экономически обоснованное понятие *слабой* стационарности.

Ряд $x(t)$ называют **слабо стационарным** (стационарным **в широком смысле**), если не зависят от времени его среднее значение и дисперсия.

Не удовлетворяющие этим определениям ряды называют **нестационарными**.

Из строгой стационарности очевидно следует слабая; обратное в общем случае не верно.

10.2. Автоковариация и автокорреляция.

Из предположения о строгой стационарности временного ряда $x(t)$ при m=2 следует совпадение совместных двумерных распределений пар СВ $(x(t_1),x(t_2))$ и $(x(\tau),x(t_2-t_1+\tau))$. Они зависят лишь от разности t_2-t_1, но не от начала отсчёта τ.

Тогда ковариация СВ $x(t)$ и $x(t\pm\tau)$ зависит только от сдвига по времени τ, но не от t.

Соответственно, **автоковариационная функция** $\gamma(\tau):=\mathrm{cov}(x(t),x(t+\tau))$ будет зависеть только от сдвига τ и будет чётна: $\gamma(-\tau)=\gamma(\tau)$.

Выборочная автоковариационная функция:

$$\hat{\gamma}(\tau):=\frac{1}{N-\tau}\sum_{t=1}^{N-\tau}\left(x(t)-\hat{a}\right)\left(x(t+\tau)-\hat{a}\right)=\frac{1}{N-\tau}\sum_{t=1}^{N-\tau}x(t)x(t+\tau)-\hat{a}^2,\ \tau\in\overline{1,\ldots,N-1}.$$

При $\tau=0\Rightarrow\gamma(0)=\sigma^2=\mathrm{var}(x(t))$.

Сила зависимости между разнесёнными по времени значениями временного ряда $x(t)$ определяется **автокорреляционной функцией** $r(\tau):=\gamma(\tau)/\gamma(0)$.

Очевидно, $r(0)=1$.

Её график называют **коррелограммой** временного ряда.

Как и для коэффициента корреляции $|r(\tau)| \le 1$. Кроме того, для стационарного временного ряда $r(-\tau)=r(\tau)$.

Выборочная автокорреляционная функция:

$$\hat{r}(\tau) := \frac{\frac{1}{N-\tau}\sum_{t=1}^{N-\tau}(x(t)-\hat{a})(x(t+\tau)-\hat{a})}{\frac{1}{N}\sum_{t=1}^{N}(x(t)-\hat{a})^2} = \frac{\hat{\gamma}(\tau)}{\hat{\gamma}(0)}.$$

Для стационарных рядов $r(\tau) \underset{\tau\to\infty}{\to} 0$, но не обязательно монотонно для конечного отрезка времени.

10.3. Частная (очищенная) автокорреляция.

Частная автокорреляционная функция определяется как парная корреляция двух членов временного ряда при фиксированных (обычно, на среднем уровне) значениях промежуточных членов этого ряда. По числу фиксированных промежуточных членов ряда определяется **порядок частной автокорреляции**.

Так, **частная автокорреляция 1-го порядка**

$$r_{част}(2) := r[x(t), x(t+2) \mid x(t+1) = a] = \frac{r_{02} - r_{01}r_{12}}{\sqrt{(1-r_{01}^2)(1-r_{12}^2)}} = \frac{r(2)-r^2(1)}{1-r^2(1)},$$

где a – мат. ожидание (стационарного!) ряда, парные коэффициенты корреляции $r_{01}:=r[x(t),x(t+1)]=r(1)$, $r_{02}:=r[x(t),x(t+2)]=r(2)$, $r_{12}:=r[x(t+1),x(t+2)]=r(1)$.

Эмпирические (выборочные) оценки частных автокорреляционных функций получаются заменой **теоретических** значений автокорреляций $r(\tau)$ их **статистическими** оценками $\hat{r}(\tau)$:

$$\hat{r}_{\text{част}}(2) := \frac{\hat{r}(2) - \hat{r}^2(1)}{1 - \hat{r}^2(1)}.$$

Очевидно, для *нестационарного* временного ряда вычислительные формулы, сводящие частную автокорреляцию к парной, значительно усложняются.

Аналогично, частные автокорреляции высших порядков сводятся к вычислению частных автокорреляций более низкого порядка.

Например, $r_{\text{част}}(3) := r[x(t), x(t+3) | x(t+1) = x(t+2) = a]$, …,

$$r_{\text{част}}(k) := r[x(t), x(t+k) | x(t+1) = x(t+2) = \ldots = x(t+k-1) = a], \ldots$$

Рекуррентные формулы, однако, достаточно сложны.

10.4. Спектральная плотность временного ряда.

Спектральная плотность $p(\omega)$ временного ряда $x(t)$ определяется через его автокорреляционную функцию $r(\tau)$ формулой $p(\omega) := \sum_{\tau=-\infty}^{+\infty} r(\tau) e^{i\tau\omega}$.

От *комплексной экспоненты* можно перейти к тригонометрическим рядам *вещественного аргумента*, если разбить сумму на 3 части (суммируя от $-\infty$ до -1, 0, от 1 до $+\infty$) и заменяя отрицательные значения аргумента τ на $-\tau$ (в силу равенства $r(-\tau) = r(\tau)$).

Получим $p(\omega)=1+2\sum_{\tau=1}^{+\infty} r(\tau)\cos\tau\omega$, где единица получается из нулевого слагаемого $r(0)=\gamma(0)/\gamma(0)=1$.

Формула обращения: $r(\tau)=\frac{1}{\pi}\int_0^{\pi} p(\omega)\cos\tau\omega d\omega$.

10.5. Спектральный анализ.

Рассмотрим временной ряд $x(1)$, $x(2)$, …, $x(N)$, измеренный относительно своего среднего значения. Тогда $Mx(t)=0$.

Интенсивность $I(\omega):=a^2(\omega)+b^2(\omega)$, где

$$a(\omega):=\frac{1}{\sqrt{\pi N}}\sum_{t=1}^{N} x(t)\cos\omega t,\ b(\omega):=\frac{1}{\sqrt{\pi N}}\sum_{t=1}^{N} x(t)\sin\omega t,$$

будет равна $I(\omega)=\frac{1}{\pi N}\left[\left(\sum_{t=1}^{N} x(t)\cos\omega t\right)^2+\left(\sum_{t=1}^{N} x(t)\sin\omega t\right)^2\right]$.

С помощью *основного тригонометрического тождества* и формулы *косинуса разности* получим

$$I(\omega)=\frac{1}{\pi N}\sum_{t=1}^{N} x^2(t)+\frac{2}{\pi N}\sum_{\tau=1}^{N-1}\sum_{t=1}^{N-\tau} x(t)x(t+\tau)\cos\omega\tau.$$

Отсюда $I(\omega)=\frac{\sigma^2}{\pi}\left[1+2\sum_{\tau=1}^{N-1}\frac{N-\tau}{N}\hat{r}(\tau)\cos\omega\tau\right]$, так как при нулевом среднем

$$\frac{\frac{1}{N-\tau}\sum_{t=1}^{N-\tau} x(t)x(t+\tau)}{\frac{1}{N}\sum_{t=1}^{N} x^2(t)}=\frac{\frac{1}{N-\tau}\sum_{t=1}^{N-\tau} x(t)x(t+\tau)}{\sigma^2}=r(\tau).$$

Так как $\frac{N-\tau}{N}\hat{r}(\tau)\underset{N\to+\infty}{\to}\hat{r}(\tau)$, то интенсивность и спектральная плотность связаны равенством $MI(\omega)=\frac{\sigma^2}{\pi}p(\omega)$. Отсюда следует *неотрицательность* спектральной плотности, т.к. среднее значение интенсивности $I(\omega)$ неотрицательно по определению.

11. Выделение неслучайной составляющей временного ряда.

11.1. Гипотеза о неизменности среднего значения.

На этапе *спецификации* регрессионной модели временного ряда $x(t)$ необходимо определить тренд (**неслучайную составляющую**). Для этого проверяется нулевая гипотеза о **постоянстве среднего значения** a временного ряда H_0:$Mx(t)=a$=Const при альтернативной гипотезе H_1:$Mx(t)\neq$Const.

Вслед за этим строится оценка (**аппроксимация**) неизвестной неслучайной составляющей, т.е. решается задача **сглаживания** (**элиминирования**) остатков временного ряда.

11.1.1. Критерий серий, основанный на медиане.

Ранжируем исходный временной ряд $x(1)$, $x(2)$, …, $x(n)$ по возрастанию:

$$x_{(1)}\le x_{(2)}\le\ldots\le x_{(n)}.$$

Такой ряд называют **вариационным**.

Выборочная медиана – это среднее (по местоположению) значение вариационного ряда. Если наблюдения считать равновероятными, то оказаться слева и справа от медианы равновозможно – это соответствует определению медианы СВ.

Определение медианы: $x_{\text{med}} := \begin{cases} x_{\left(\frac{n+1}{2}\right)} \leftarrow n = 2k+1, \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) \leftarrow n = 2k. \end{cases}$.

Заменим члены исходного (неранжированного!) временного ряда знаками «+», если они больше медианы ($x(t)>x_{\text{med}}$) и знаками «–», если они меньше медианы ($x(t)<x_{\text{med}}$). Члены временного ряда, равные медиане при этом не учитываются.

Назовём **серией** последовательность одинаковых знаков. (Серия может состоять и из одного знака!)

Полученная последовательность знаков характеризуется **общим числом серий** $\nu(n)$ и **протяжённостью самой длинной серии** $\tau(n)$.

Очевидно, если исследуемая последовательность обладает постоянным средним значением, то число серий $\nu(n)$ должно быть достаточно *велико* (наблюдаются частые колебания наблюдений вокруг медианного значения), а протяжённость самой длинной серии $\tau(n)$ должна быть довольно *мала* (не может быть долговременного отклонения наблюдений в одну сторону от медианного значения).

Так получается **пара критических статистик** ($\nu(n)$, $\tau(n)$).

Нулевая гипотеза H_0 о неизменности среднего значения временного ряда отвергается с вероятностью ошибки $\alpha \in [0{,}05;0{,}0975]$, если не выполняется хотя бы одно из неравенств системы

$$\begin{cases} \nu(n) > \left[\frac{1}{2}\left(n+2-1{,}96\sqrt{n-1}\right)\right], \\ \tau(n) < [1{,}43\ln(n+1)], \end{cases}$$

где […] – целая часть числа.

11.1.2. Критерий восходящих и нисходящих серий.

Как и в предыдущем критерии, вместо исходного временного ряда $x(1)$, $x(2)$, …, $x(n)$ рассмотрим последовательность знаков «+», если $x(t+1)>x(t)$, и «–», если $x(t+1)>x(t)$. (Если последовательно повторяются два или несколько одинаковых значений временного ряда, то во внимание принимается только первое из них!)

Полученная последовательность знаков вновь характеризуется **общим числом серий** $\nu(n)$ и **протяжённостью самой длинной серии** $\tau(n)$.

Очевидно, если исследуемая последовательность обладает постоянным средним значением, то опять же число серий $\nu(n)$ должно быть достаточно *велико*, а протяжённость самой длинной серии $\tau(n)$ должна быть довольно *мала*.

Так получается **пара критических статистик** $(\nu(n), \tau(n))$.

Нулевая гипотеза H_0 о неизменности среднего значения временного ряда отвергается с вероятностью ошибки $\alpha \in [0{,}05; 0{,}0975]$, если не выполняется хотя бы одно из неравенств системы

$$\begin{cases} \nu(n) > \left[\frac{1}{3}(2n-1) - 1{,}96\sqrt{\frac{16n-29}{90}} \right], \\ \tau(n) < \tau_0(n), \end{cases}$$

где […] – целая часть числа, а величина $\tau_0(n)$ задана таблицей

n	n<27	26<n<154	153<n<1171
$\tau_0(n)$	5	6	7

После обнаружения изменчивости среднего значения необходимо её исключить тем или иным образом.

11.2. Аналитический метод выделения неслучайной составляющей.

Тренд ищем в виде регрессии $x(t) = \hat{x}(t, \alpha_0, \alpha_1, \ldots, \alpha_p) + \varepsilon(t)$, $t \in \overline{1,n}$, где

$$\hat{x}(t, \alpha_0, \alpha_1, \ldots, \alpha_p) := \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \ldots + \alpha_p t^p .$$

В частном случае **линейной** регрессии $\hat{x}(t, \alpha_0, \alpha_1) := \alpha_0 + \alpha_1 t$, применяя МНК, получим систему 2-х линейных уравнений с 2-мя неизвестными

$$\begin{cases} \alpha_0 + \alpha_1 \frac{n+1}{2} - \frac{1}{n}\sum_{t=1}^{n} x(t) = 0, \\ \alpha_0 \frac{n+1}{2} + \alpha_1 \frac{(n+1)(2n+1)}{6} - \frac{1}{n}\sum_{t=1}^{n} t x(t) = 0. \end{cases}$$

Если неслучайная составляющая представлена полиномом степени p, аналогично строится СЛАУ p+1 уравнений с p+1 неизвестными α_0, α_1, …, α_p.

Для спецификации порядка полинома p примен***и***м метод последовательных разностей.

11.3. Метод последовательных разностей.

Вслед за Ньютоном определим **конечные (последовательные) разности 1-го порядка**: $\Delta x(t):=x(t)-x(t-1)$.

Тогда **конечные разности 2-го** и **более высоких порядков** определим рекуррентно:

$$\Delta^2 x(t):=\Delta(\Delta x(t))=\Delta(x(t)-x(t-1))=x(t)-x(t-1)-[x(t-1)-x(t-2)]=x(t)-2x(t-1)+x(t-2),$$

$$\Delta^3 x(t):=\Delta(\Delta^2 x(t))=\Delta(x(t)-2x(t-1)+x(t-2))=$$

$$=x(t)-2x(t-1)+x(t-2)-[x(t-1)-2x(t-2)+x(t-3)]=x(t)-3x(t-1)+3x(t-2)-x(t-3),$$

$$\Delta^4 x(t):=\Delta(\Delta^3 x(t))=\Delta(x(t)-3x(t-1)+3x(t-2)-x(t-3))=$$

$$=x(t)-3x(t-1)+3x(t-2)-x(t-3)-[x(t-1)-3x(t-2)+3x(t-3)-x(t-4)]=$$

$$=x(t)-4x(t-1)+6x(t-2)-4x(t-3)+x(t-4),$$

…

$$\Delta^{p+1} x(t):=\Delta(\Delta^p x(t)),\ \ldots$$

Коэффициенты разложений можно получить с помощью **модифицированного треугольника Паскаля**, представляющего коэффициенты **бинома Ньютона** C_p^k соответствующей степени. Модификация заключается в чередовании знаков «+» и «–» в каждой строке, начиная со знака «+». Таким образом,

каждая строка модифицированного треугольника представляет коэффициенты *бинома разности*.

Для линейной модели регрессии $x(t)=\alpha_0+\alpha_1 t+\varepsilon(t)$ получим последовательно

$\Delta x(t)=\alpha_0+\alpha_1 t+\varepsilon(t)-[\alpha_0+\alpha_1(t-1)+\varepsilon(t-1)]=\alpha_1+\varepsilon(t)-\varepsilon(t-1)$,

$\Delta^2 x(t)=\Delta[\alpha_1+\varepsilon(t)-\varepsilon(t-1)]=\alpha_1+\varepsilon(t)-\varepsilon(t-1)-[\alpha_1+\varepsilon(t-1)-\varepsilon(t-2)]=\varepsilon(t)-2\varepsilon(t-1)+\varepsilon(t-2)$.

Таким образом, конечные разности 1-го порядка «стабилизируются», а конечные разности 2-го порядка не содержат неслучайной составляющей.

Аналогично, для квадратической регрессии $x(t)=\alpha_0+\alpha_1 t+\alpha_2 t^2+\varepsilon(t)$ получим

$\Delta x(t)=\alpha_0+\alpha_1 t+\alpha_2 t^2+\varepsilon(t)-[\alpha_0+\alpha_1(t-1)+\alpha_2(t-1)^2+\varepsilon(t-1)]=\alpha_1+\alpha_2(2t-1)+\varepsilon(t)-\varepsilon(t-1)$,

$\Delta^2 x(t)=\Delta[\alpha_1+\alpha_2(2t-1)+\varepsilon(t)-\varepsilon(t-1)]=$

$=\alpha_1+\alpha_2(2t-1)+\varepsilon(t)-\varepsilon(t-1)-[\alpha_1+\alpha_2(2t-3)+\varepsilon(t-1)-\varepsilon(t-2)]=2\alpha_2+\varepsilon(t)-2\varepsilon(t-1)+\varepsilon(t-2)$,

$\Delta^3 x(t)=\Delta[2\alpha_2+\varepsilon(t)-2\varepsilon(t-1)+\varepsilon(t-2)]=$

$=2\alpha_2+\varepsilon(t)-2\varepsilon(t-1)+\varepsilon(t-2)-[2\alpha_2+\varepsilon(t-1)-2\varepsilon(t-2)+\varepsilon(t-3)]=\varepsilon(t)-3\varepsilon(t-1)+3\varepsilon(t-2)+\varepsilon(t-3)$.

Таким образом, конечные разности 2-го порядка «стабилизируются», а конечные разности 3-го порядка не содержат неслучайной составляющей.

Очевидно, подобным свойством обладают все полиномиальные регрессии p-го порядка: их конечные разности p-го порядка «стабилизируются», а $(p+1)$-го порядка – не содержат случайной составляющей.

При этом $M\Delta^{p+1}x(t)=0$ и $D\Delta^{p+1}x(t)=\sigma^2_{x(t)}\left[1+\left(C^1_{p+1}\right)^2+\left(C^2_{p+1}\right)^2+\ldots+\left(C^{p+1}_{p+1}\right)^2\right]=\sigma^2 C^{p+1}_{2p+2}$.

Тогда можно предложить следующий **алгоритм определения порядка сглаживающего полинома**.

Для k=1,2,… последовательно находим конечные разности $\Delta^k x(t)$ и выборочные оценки их дисперсии $\hat{\sigma}^2(k) := \frac{S^2}{C_{2k}^k} = \frac{1}{(n-k)C_{2k}^k}\sum_{t=1}^{n-k}\left(\Delta^k x(t)\right)^2$.

Величина $\hat{\sigma}^2(k)$ будет иметь тенденцию к убыванию с ростом порядка конечных разностей k до p+1. Начиная с k=p+1, она стабилизируется на некотором минимальном уровне. Отсюда определим порядок p сглаживающего полинома.

Приведём таблицу вспомогательных коэффициентов C_{2k}^k.

k	1	2	3	4	5	6	7	8	9	10
C_{2k}^k	2	6	20	70	252	924	3432	12870	48260	184756

Таким образом, получен **верхний предел** порядка полиномов для выделения неслучайной составляющей.

11.4. Метод скользящего среднего.

Его идея заключается в замене исходного временного ряда $x(1)$, $x(2)$, …, $x(n)$ с дисперсией σ^2 **сглаженным рядом** из средних взвешенных соседних $2m$+1 значений

$$\hat{f}(t) := \sum_{k=-m}^{m} w_k x(t+k),\ t \in \overline{m+1, n-m}, \qquad (6)$$

с **весовыми коэффициентами**

$$w_k \in [0,1] : \sum_{k=-m}^{m} w_k = 1$$

и меньшей дисперсией $\sigma^2/(2m+1)$.

При запуске t от m+1 до $n–m$ «маска» для расчёта (6) скользит по оси времени так, что при каждом следующем пересчёте происходит замена только одного слагаемого $x(t–m)$ слагаемым $x(t+m+1)$. Поэтому этот метод назван **методом скользящего среднего**.

12.1. Автокорреляция и авторегрессия.

Автокорреляция – это корреляционная зависимость между последовательными уровнями временного ряда:

$$r_1 := \frac{\sum_{t=2}^{n}(y_t - \bar{y}_1)(y_{t-1} - \bar{y}_2)}{\sqrt{\sum_{t=2}^{n}(y_t - \bar{y}_1)^2}\sqrt{\sum_{t=2}^{n}(y_{t-1} - \bar{y}_2)^2}}$$

– коэффициент автокорреляции уровней ряда **1-го порядка**,

где $\bar{y}_1 := \frac{1}{n-1}\sum_{t=2}^{n} y_t$, $\bar{y}_2 := \frac{1}{n-1}\sum_{t=2}^{n} y_{t-1}$;

$$r_2 := \frac{\sum_{t=3}^{n}(y_t - \bar{y}_3)(y_{t-2} - \bar{y}_4)}{\sqrt{\sum_{t=3}^{n}(y_t - \bar{y}_3)^2}\sqrt{\sum_{t=3}^{n}(y_{t-2} - \bar{y}_4)^2}}$$

– коэффициент автокорреляции уровней ряда **2-го порядка**,

где $\bar{y}_3 := \frac{1}{n-2}\sum_{t=3}^{n} y_t$, $\bar{y}_4 := \frac{1}{n-2}\sum_{t=3}^{n} y_{t-2}$; и т.д.

Последовательность коэффициентов автокорреляции уровней 1-го, 2-го и т.д. порядков называется **автокорреляционной функцией** временного ряда, а график зависимости её значений от величины **лага** (порядка коэффициента корреляции) – **коррелограммой**.

Аналитическим выравниванием временного ряда называют построение аналитической функции для моделирования его **тренда** (**тенденции**).

Виды тренда:

$\tilde{y}_t := a + bt$ – линейный;

$\tilde{y}_t := a + b/t$ – гиперболический;

$\tilde{y}_t := e^{a+bt}$ – экспоненциальный;

$\tilde{y}_t := at^b$ – степенной;

$\tilde{y}_t := a + b_1 t + b_2 t^2 + \ldots + b_k t^k$ – полиномиальный k-го порядка; и.т.п.

12.2. Автокорреляция в остатках.

Автокорреляция в остатках – это корреляционная зависимость между значениями остатков регрессии ε_t за текущий и предыдущие моменты времени. Для её определения используют критерий Дарбина–Уотсона со **статистикой**

$$d := \frac{\sum_{t=2}^{n} (\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^{n} \varepsilon_t^2}, \; 0 \le d \le 4.$$

Коэффициент автокорреляции остатков 1-го прядка

$$r_1^{\varepsilon} := \frac{\sum_{t=2}^{n} \varepsilon_t \varepsilon_{t-1}}{\sum_{t=2}^{n} \varepsilon_t^2}, \; -1 \le r_1^{\varepsilon} \le 1,$$

связан со статистикой Дарбина–Уотсона соотношением $d \approx 2\left(1 - r_1^{\varepsilon}\right)$.

13. Белый шум.

Определение. Временной ряд $\delta(t)$ называется **белым шумом**, если

1) $M\delta(t)=0=\text{Const}$;

2) $M\left[\delta(t),\delta(t\pm\tau)\right]=\begin{cases}0 \leftarrow \tau=0,\\ \sigma_0^2=\text{Const} \leftarrow \tau\neq 0.\end{cases}$

Таким образом, белый шум стационарен (по крайней мере, *слабо*), т.к. его мат. ожидание и дисперсия не зависят от времени *t*.

14. Линейные временные процессы.

После выделения (элиминации) неслучайной составляющей $\hat{x}(t)$ временного ряда $x(t)=\hat{x}(t)+\varepsilon(t)$ возникает необходимость описания случайного остатка $\varepsilon(t)$. Для этого применяется представление остатка в виде **линейного процесса**.

1. Можно представить остаток $\varepsilon(t)$ линейной комбинацией (конечного или бесконечного числа) настоящего и прошлых (**лаговых**) значений белого шума $\beta(t)$:

$$\varepsilon(t):=\delta(t)+\beta_1\delta(t-1)+\beta_2\delta(t-2)+\ldots=\sum_{k=0}^{+\infty}\beta_k\delta(t-k), \qquad (7)$$

где $\beta_0=1$, $\sum_{k=0}^{+\infty}\beta_k^2<+\infty$. (Множители β_k могут быть отрицательными, в отличие от случая *неотрицательных весовых коэффициентов* для взвешенной суммы.)

2. Эквивалентная постановка задачи в виде классической линейной модели множественной регрессии (быть может, с бесконечным числом регрессоров),

когда в качестве объясняющих переменных рассматриваются лаговые значения исследуемого временного ряда ε(t):

$$\varepsilon(t) := \lambda_1\varepsilon(t-1) + \lambda_2\varepsilon(t-2) + \ldots + \delta(t) = \sum_{k=1}^{+\infty}\lambda_k\varepsilon(t-k) + \delta(t). \qquad (8)$$

Заметим, что *младший* лаг может быть сколь угодно большим, т.к.

$\varepsilon(t-1)=\lambda_1\varepsilon(t-2)+\lambda_2\varepsilon(t-3)+\ldots+\delta(t-1) \Rightarrow$

$$\Rightarrow \varepsilon(t) = \left(\lambda_2 + \lambda_1^2\right)\varepsilon(t-2) + \left(\lambda_3 + \lambda_1\lambda_2\right)\varepsilon(t-3) + \ldots + \delta(t) + \lambda_1\delta(t-1),$$

$\varepsilon(t-2)=\lambda_1\varepsilon(t-3)+\lambda_2\varepsilon(t-4)+\ldots+\delta(t-2) \Rightarrow$

$$\Rightarrow \varepsilon(t) = \lambda_1\left(\lambda_2 + \lambda_1^2\right)\left(\lambda_3 + \lambda_1\lambda_2\right)\varepsilon(t-3) + \ldots + \delta(t) + \lambda_1\delta(t-1) + \left(\lambda^2 + \lambda_1^2\right)\delta(t-2) \Rightarrow \ldots$$

Как видим, формула (8) шаг за шагом превращается в формулу (7), так что нельзя считать случаи представления рядом по запаздываниям белого шума или бесконечной авторегрессией принципиально различными. Кроме того, есть смысл рассмотреть смешанное представление.

3. Общий случай. Это процесс **смешанного типа**

$$\varepsilon(t) := \sum_{k=1}^{p}\lambda_k\varepsilon(t-k) + \delta(t) + \sum_{k=1}^{q}\beta_k\delta(t-k).$$

Здесь число слагаемых в каждой сумме может быть как нулевым, так и бесконечным: $0 \le p \le +\infty$, $0 \le q \le +\infty$.

Рассмотрим частные случаи описанных моделей.

14.1.1. Модель авторегрессии 1-го порядка AP(1). Марковские процессы.

Модель AP(1) (**марковский процесс**, в английской транскрипции AR(1)) – простейший вариант линейного авторегрессионного процесса $\varepsilon(t)=\alpha\varepsilon(t-1)+\delta(t)$,

где α – вещественная постоянная (**коэффициент авторегрессии**), $\delta(t)$ – белый шум с дисперсией σ_0^2.

Для обеспечения *стационарности* необходимо и достаточно выполнение условия $|\alpha|<1$.

Поскольку

$$\varepsilon(t)=\alpha\varepsilon(t-1)+\delta(t)=\alpha[\alpha\varepsilon(t-2)+\delta(t-1)]+\delta(t)=\alpha^2\varepsilon(t-2)+\delta(t)+\alpha\delta(t-1)=$$

$$=\alpha^2[\alpha\varepsilon(t-3)+\delta(t-2)]+\delta(t)+\alpha\delta(t-1)=\alpha^3\varepsilon(t-3)+\delta(t)+\alpha\delta(t-1)+\alpha^2\delta(t-2)=\ldots=$$

$$=\alpha^k\varepsilon(t-k)+\delta(t)+\alpha\delta(t-1)+\alpha^2\delta(t-2)+\ldots+\alpha^{k-1}\delta(t-k+1),$$

то остатки $\delta(t)$, $\delta(t-k)$ в этой модели автокоррелированы, и, следовательно, МНК не даёт эффективных оценок коэффициента авторегрессии α. (Хотя МНК-оценки и являются несмещёнными!)

Из определения модели АР(1) и свойств белого шума получим:

1) $M\varepsilon(t)=M[\alpha\varepsilon(t-1)+\delta(t)]=\alpha M\varepsilon(t)+M\delta(t)=\alpha M\varepsilon(t)\Leftrightarrow(1-\alpha)M\varepsilon(t)=0\Leftrightarrow M\varepsilon(t)\equiv 0$;

2) $D\varepsilon(t)=D[\alpha\varepsilon(t-1)+\delta(t)]=\alpha^2 D\varepsilon(t)+2\alpha\operatorname{cov}(\varepsilon(t-1),\delta(t))+D\delta(t)=\alpha^2 D\varepsilon(t)+\sigma_0^2\Leftrightarrow$

$$\Leftrightarrow(1-\alpha^2)D\varepsilon(t)=\sigma_0^2\Leftrightarrow D\varepsilon(t)=\frac{\sigma_0^2}{1-\alpha^2};$$

3) $\operatorname{cov}(\varepsilon(t),\varepsilon(t\pm k))=M\varepsilon(t)\varepsilon(t\pm k)-\underbrace{M\varepsilon(t)}_{0}\cdot\underbrace{M\varepsilon(t\pm k)}_{0}=$

$$=M\varepsilon(t)\left[\alpha^k\varepsilon(t-k)+\delta(t)+\alpha\delta(t-1)+\ldots+\alpha^{k-1}\delta(t-k+1)\right]=$$

$$=\alpha^k M\varepsilon(t)\varepsilon(t-k)+\underbrace{M\varepsilon(t)\delta(t)}_{0}+\alpha\underbrace{M\varepsilon(t)\delta(t-1)}_{0}+\ldots+\alpha^{k-1}\underbrace{M\varepsilon(t)\delta(t-k+1)}_{0}=\alpha^k D\varepsilon(t)=\frac{\alpha^k\sigma_0^2}{1-\alpha^2};$$

4) $r[\varepsilon(t),\varepsilon(t\pm k)]=\dfrac{\operatorname{cov}[\varepsilon(t),\varepsilon(t\pm k)]}{\sigma_{\varepsilon(t)}\sigma_{\varepsilon(t\pm k)}}=\dfrac{\alpha^k D\varepsilon(t)}{D\varepsilon(t)}=\alpha^k$.

Коррелограмма марковского процесса $r(\tau):=r[\varepsilon(t),\varepsilon(t\pm\tau)]=\alpha^\tau$.

Следовательно, при выполнении условия стационарности $|\alpha|<1$ теснота корреляционной связи между членами ряда $\varepsilon(t)$ убывает по мере их удаления друг от друга.

Из равенства $\alpha=r(1)$ получаем **выборочную оценку коэффициента авторегрессии** $\hat{\alpha} := \hat{r}(1)$ через выборочную оценку парного коэффициента автокорреляции.

Для **частной автокорреляционной функции марковского процесса** получим:

$$r_{\text{част}}(k) := r\left[\varepsilon(t), \varepsilon(t+k) \mid \varepsilon(t+1) = \varepsilon(t+2) = \ldots = \varepsilon(t+k-1) = M\varepsilon(t) \equiv 0\right] =$$

$$= \begin{cases} \dfrac{r(2)-r^2(1)}{1-r^2(1)} = \dfrac{\alpha^2-\alpha^2}{1-\alpha^2} = 0 \leftarrow k=2, \\ 0 \leftarrow k \geq 3, \end{cases} = 0 \quad \forall k \geq 2.$$

Это свойство частной автокорреляции можно использовать для *спецификации* модели авторегрессии: если частные корреляции $\hat{r}_{\text{част}}(\tau)$ невязок $\varepsilon(t) := x(t) - \hat{x}(t)$ незначимо отличаются от 0 при τ=2,3,…, то оправданно применение модели АР(1) для описания случайных остатков.

Спектральная плотность марковского процесса

$$p(\tilde{\omega}) = \frac{2\sigma_0^2}{1-2\alpha\cos 2\pi\tilde{\omega}+\alpha^2}, \; 0 \leq \tilde{\omega} \leq 1/2, \; \tilde{\omega} := \omega/2\pi.$$

14.1.2. Идентификация параметров модели AP(1).

После спецификации модели $\varepsilon(t)=\alpha\varepsilon(t-1)+\delta(t)$ авторегрессии AP(1) необходимо статистически оценить её параметры α и σ_0^2.

Напомним, что $\hat{\varepsilon}(t) := x(t) - \hat{x}(t)$ – **невязка неслучайной составляющей** $\hat{x}(t)$ временного ряда $x(t)$. Тогда выборочная дисперсия остатков находится по формуле выборочной дисперсии (смещённой ил несмещённой).

14.2. Модель авторегрессии 2-го порядка AP(2). Процессы Юла.

Модель AP(2) (**процесс Юла**, в английской транскрипции AR(2)) – частный случай линейного авторегрессионного процесса $\varepsilon(t)=\alpha_1\varepsilon(t-1)+\alpha_2\varepsilon(t-2)+\delta(t)$, где α_1, α_2 – вещественные постоянные (**коэффициенты авторегрессии**), $\delta(t)$ – белый шум с дисперсией σ_0^2.

Последовательно подставляя в правую часть вместо $\varepsilon(t-1)$, $\varepsilon(t-2)$ их выражения через $\delta(t)$, $\delta(t-1)$, $\delta(t-2)$, …, получим, что как и в AP(1)-модели, $\varepsilon(t)$ зависит только от прошлых лаговых значений белого шума, но не от будущих значений $\delta(t+1)$, $\delta(t+2)$, …

Мат. ожидание процесса Юла

$$a := M\varepsilon(t) = \alpha_1 \underbrace{M\varepsilon(t-1)}_{a} + \alpha_2 \underbrace{M\varepsilon(t-2)}_{a} + \underbrace{M\delta(t)}_{0} = (\alpha_1+\alpha_2)a \Leftrightarrow a \equiv 0$$

при условии *стационарности* процесса $\varepsilon(t)$.

Обозначим для краткости **автоковариации**:

$$\gamma(0) := D\varepsilon(t) = \text{Const} > 0;$$

$$\gamma(1) := \operatorname{cov}(\varepsilon(t), \varepsilon(t-1)) = M\varepsilon(t)\varepsilon(t-1) - \underbrace{M\varepsilon(t)}_{0} \cdot \underbrace{M\varepsilon(t-1)}_{0} =$$

$$= M\varepsilon(t-1)[\alpha_1\varepsilon(t-1) + \alpha_2\varepsilon(t-2) + \delta(t)] = \alpha_1 \underbrace{M\varepsilon^2(t-1)}_{\gamma(0)} + \alpha_2 \underbrace{M\varepsilon(t-1)\varepsilon(t-2)}_{\gamma(1)} + \underbrace{M\varepsilon(t-1)\delta(t)}_{0} =$$

$$= \alpha_1\gamma(0) + \alpha_2\gamma(1);$$

$$\gamma(2) := \operatorname{cov}(\varepsilon(t), \varepsilon(t-2)) = M\varepsilon(t)\varepsilon(t-2) - \underbrace{M\varepsilon(t)}_{0} \cdot \underbrace{M\varepsilon(t-2)}_{0} =$$

$$= M\varepsilon(t-2)[\alpha_1\varepsilon(t-1) + \alpha_2\varepsilon(t-2) + \delta(t)] = \alpha_1 \underbrace{M\varepsilon(t-1)\varepsilon(t-2)}_{\gamma(1)} + \alpha_2 \underbrace{M\varepsilon^2(t-2)}_{\gamma(0)} + \underbrace{M\varepsilon(t-2)\delta(t)}_{0} =$$

$$= \alpha_1\gamma(1) + \alpha_2\gamma(0).$$

Отсюда получим СЛАУ 2-х однородных уравнений с 3-мя неизвестными $\gamma(0)$, $\gamma(1)$, $\gamma(2)$: $\begin{cases} \gamma(1) - \alpha_1\gamma(0) - \alpha_2\gamma(1) = 0, \\ \gamma(2) - \alpha_1\gamma(1) - \alpha_2\gamma(0) = 0. \end{cases}$

Поделив эти уравнения на $\gamma(0)>0$, получим СЛАУ 2-х неоднородных уравнений с 2-мя неизвестными **коэффициентами парной автокорреляции** $r(1)$, $r(2)$, если считать заданными **коэффициенты авторегрессии** α_1, α_2; или же – с 2-мя неизвестными коэффициентами α_1, α_2, если считать заданными коэффициенты парной автокорреляции $r(1)$, $r(2)$:

$$\begin{cases} r(1) - \alpha_1 - \alpha_2 r(1) = 0, \\ r(2) - \alpha_1 r(1) - \alpha_2 = 0, \end{cases} \Leftrightarrow \begin{cases} r(1) = \dfrac{\alpha_1}{1-\alpha_2}, \\ r(2) = \alpha_2 + \dfrac{\alpha_1^2}{1-\alpha_2}, \\ \alpha_1 = \dfrac{r(1)[1-r(2)]}{1-r^2(1)}, \\ \alpha_2 = \dfrac{r(2)-r^2(1)}{1-r^2(1)}. \end{cases}$$

Найдём

$$\gamma(0) := D\varepsilon(t) = M\varepsilon^2(t) - \underbrace{M^2\varepsilon(t)}_{0} = M\varepsilon(t)\left[\alpha_1\varepsilon(t-1) + \alpha_2\varepsilon(t-2) + \delta(t)\right] =$$

$$= \alpha_1 \underbrace{M\varepsilon(t)\varepsilon(t-1)}_{\gamma(1)} + \alpha_2 \underbrace{M\varepsilon(t)\varepsilon(t-2)}_{\gamma(2)} + M\left[\alpha_1\varepsilon(t-1) + \alpha_2\varepsilon(t-2) + \delta(t)\right]\delta(t) =$$

$$= \alpha_1\gamma(1) + \alpha_2\gamma(2) + \alpha_1 \underbrace{M\varepsilon(t-1)\delta(t)}_{0} + \alpha_2 \underbrace{M\varepsilon(t-2)\delta(t)}_{0} + \underbrace{M\delta^2(t)}_{\sigma_0^2} = \alpha_1\gamma(1) + \alpha_2\gamma(2) + \sigma_0^2 .$$

Отсюда $\sigma_0^2 = \gamma(0) - \alpha_1\gamma(1) - \alpha_2\gamma(2) = \gamma(0)\left[1 - \alpha_1 r(1) - \alpha_2 r(2)\right] = \sigma_\varepsilon^2 \dfrac{1+\alpha_2}{1-\alpha_2}\left[(1-\alpha_2)^2 - \alpha_1^2\right].$

Визуально реализации моделей АР(1) и АР(2) не различимы.

Различение этих спецификаций можно провести по среднему расстоянию $d(\alpha_1,\alpha_2)$ между **пиками** ряда (формула Кендэла):

$$d(\alpha_1,\alpha_2) = \frac{2\pi}{\arccos 0{,}5(\alpha_1 + \alpha_2 - 1)}.$$

Однако, такая диагностика **не очень чувствительна**:

1) для белого шума $\alpha_1=\alpha_2=0 \Rightarrow d(0,0) = 2\pi/\arccos(-0{,}5) = 2\pi/(2\pi/3) = 3;$

2) для марковских процессов с учётом условия стационарности

$$\alpha_2=0,\ -1<\alpha_1<1 \Rightarrow d(\alpha_1,0) = 2\pi/\arccos 0{,}5(\alpha_1 - 1) \in (3;4);$$

3) для процессов Юла с учётом условий стационарности $d \in (4;6)$.

Более обоснованные выводы можно сделать на основе анализа автокорреляционной функции, спектра и т.д.

Автокорреляционную функцию процесса Юла для $\tau=1$ и $\tau=2$ определили выше: $r(1) = \dfrac{\alpha_1}{1-\alpha_2}$, $r(2) = \alpha_2 + \dfrac{\alpha_1^2}{1-\alpha_2}$.

Частная автокорреляционная функция

$$r_{\text{част}}(\tau) := r\left[\varepsilon(t), \varepsilon(t+\tau) \mid \varepsilon(t+1) = \varepsilon(t+2) = \ldots = \varepsilon(t+\tau-1) = M\varepsilon(t) \equiv 0\right]$$

модели АР(2) обладает **характеристическим свойством** $r_{\text{част}}(\tau)=0$ при $\tau=3,4,\ldots$ (в отличие от моделей АР(1), у которых частная автокорреляция становится нулевой, начиная с лага $\tau=2$).

При этом

$$r_{\text{част}}(2)=\frac{r(2)-r^2(1)}{1-r^2(1)}=\frac{\alpha_2+\frac{\alpha_1^2}{1-\alpha_2}-\left(\frac{\alpha_1}{1-\alpha_2}\right)^2}{1-\left(\frac{\alpha_1}{1-\alpha_2}\right)^2}=\frac{\alpha_2(1-\alpha_2)^2+\alpha_1^2(1-\alpha_2)-\alpha_1^2}{(1-\alpha_2)^2-\alpha_1^2}=$$

$$=\frac{\alpha_2(1-2\alpha_2+\alpha^2)+\alpha_1^2-\alpha_1^2\alpha_2-\alpha_1^2}{1-2\alpha_2+\alpha_2^2-\alpha_1^2}=\frac{\alpha_2(1-2\alpha_2+\alpha_2^2-\alpha_1^2)}{1-2\alpha_2+\alpha_2^2-\alpha_1^2}=\alpha_2\neq 0.$$

Спектральная плотность процесса Юла

$$p(\tilde{\omega})=\frac{2\sigma_0^2}{1-2\alpha_1(1-\alpha_2)\cos 2\pi\tilde{\omega}-2\alpha_2\cos 4\pi\tilde{\omega}+\alpha_1^2+\alpha_2^2},\ 0\le\tilde{\omega}\le 1/2,\ \tilde{\omega}:=\omega/2\pi.$$

При этом

$$\lim_{\alpha_2\to 0}\frac{2\sigma_0^2}{1-2\alpha_1(1-\alpha_2)\cos 2\pi\tilde{\omega}-2\alpha_2\cos 4\pi\tilde{\omega}+\alpha_1^2+\alpha_2^2}=\frac{2\sigma_0^2}{1-2\alpha_1\cos 2\pi\tilde{\omega}+\alpha_1^2},$$

т.е. спектральная плотность марковского процесса является *предельным случаем* спектральной плотности процесса Юла.

14.3.1. Модели АР(*k*).

Аналогично моделям авторегрессии АР(1) и АР(2) определяются **авторегрессионные модели** АР(k) более высоких порядков

$$\varepsilon(t)=\alpha_1\varepsilon(t-1)+\alpha_2\varepsilon(t-2)+\ldots+\alpha_k\varepsilon(t-k)+\delta(t).$$

При выполнении соответствующих условий стационарности **характеристическое свойство частной автокорреляционной функции**:

$$r_{\text{част}}(\tau)=0 \text{ при } \tau=k+1,k+2, \ldots$$

Спектральная функция АР(k) выглядит гораздо сложнее.

14.3.2. Условия стационарности модели АР(k).

Авторегрессионной модели АР(k)

$$\varepsilon(t)=\alpha_1\varepsilon(t-1)+\alpha_2\varepsilon(t-2)+\ldots+\alpha_k\varepsilon(t-k)+\delta(t)$$

ставится в соответствие **характеристическое уравнение**

$$1-\alpha_1 z-\alpha_2 z^2-\ldots-\alpha_k z^k=0$$

над **полем комплексных чисел** C.

Его корни называются **характеристическими**.

По Великой теореме алгебры (теореме Гаусса), дополненной теоремой Безу, характеристических корней над полем C с учётом кратности k штук: $z_1, z_2, \ldots, z_k \in C$.

Необходимое и достаточное условие стационарности линейной авторегрессионной модели: все характеристические корни лежат вне единичного круга комплексной плоскости, т.е. $|z|>1$.

В частности, для АР(1)-модели данное условие принимает вид:

$$1-\alpha z=0 \Leftrightarrow z_1=1/\alpha \Rightarrow |z_1|=1/|\alpha|>1 \Leftrightarrow |\alpha|<1.$$

При этом в эконометрических исследованиях $\alpha \in \Re$, так что *единственный* характеристический корень АР(1)-модели вещественен: $z_1 \in \Re$.

Для АР(2)-модели условие стационарности становится более громоздким:

$$1-\alpha_1 z-\alpha_2 z^2=0 \Leftrightarrow \alpha_2 z^2+\alpha_1 z-1=0 \Rightarrow z_{1,2}=\frac{-\alpha_1 \pm \sqrt{\alpha_1^2+4\alpha_2}}{2\alpha_2} \Rightarrow \left|\frac{-\alpha_1 \pm \sqrt{\alpha_1^2+4\alpha_2}}{2\alpha_2}\right|>1.$$

Эквивалентным является требование **ограниченности коэффициентов автокорреляции**

$$r(1), r(2) \in [-1;1] \Leftrightarrow \begin{cases} -1 \le \dfrac{\alpha_1}{1-\alpha_2} \le 1, \\ -1 \le \alpha_2 + \dfrac{\alpha_1^2}{1-\alpha_2} \le 1, \end{cases} \text{ или } \begin{cases} -1 \le r(1) \le 1, \\ -1 \le r(2) \le 1, \\ r(2) \ge 2r^2(1)-1. \end{cases}$$

14.4.1. Модели с распределёнными лагами.

Эконометрические модели, содержащие не только текущие, но и лаговые значения регрессоров, называются моделями с **распределённым лагом**. Если максимальная величина лага конечна, то модель с распределённым лагом имеет общий вид $y_t = a + b_0 x_t + b_1 x_{t-1} + \ldots + b_k x_{t-k} + \varepsilon_t$.

Коэффициент регрессии b_0 (**краткосрочный мультипликатор**) при переменной x_t характеризует среднее абсолютное изменение y_t при изменении x_t на 1 единицу своего измерения в некоторый фиксированный момент времени t без учёта воздействия лаговых значений регрессора x.

В момент t+1 воздействие регрессора x_t на результат y_t составит b_0+b_1 условных единиц, в момент времени t+2 составит b_0+b_1+b_2 условных единиц и т.д.

Эти суммы называют **промежуточными мультипликаторами**. Для максимального лага $t+k$ воздействие регрессора на результат определяется **долгосрочным мультипликатором** $b=b_0+b_1+\ldots+b_k$.

Величины $\beta_j := b_j/b$, $0 \le j \le k$, называются **относительными коэффициентами** модели с распределёнными лагами. Если все коэффициенты b_j одного знака, то

$$\forall j \in \overline{1,k}\ \ 0<\beta_j<1 \text{ и } \sum_{j=0}^{k} \beta_j = 1.$$

Формулой средней арифметической взвешенной определяется **средний лаг** $\bar{k} := \sum_{j=0}^{k} j \cdot \beta_j = \sum_{j=1}^{k} j \cdot \beta_j$, описывающий средний период, в течение которого будет происходить изменение результата под воздействием изменения регрессора в момент t.

Медианный лаг – это период, в течение которого с момента времени t будет реализована половина общего воздействия фактора на результат:

$$\sum_{j=0}^{k_{Me}-1} \beta_j \approx 0{,}5,$$

где k_{Me} – медианный лаг.

14.4.2. Метод Койка.

В **распределении Койка** предполагается, что коэффициенты при лаговых значениях регрессора убывают в геометрической прогрессии:

$$b_j=b_0\lambda^j,\ j=0,1,2,\ldots,\ 0<\lambda<1.$$

Тогда уравнение регрессии примет вид $y_t=a+b_0x_t+b_0\lambda x_{t-1}+b_0\lambda^2 x_{t-2}+\ldots+\varepsilon_t$.

14.4.3. Метод Алмон

В **методе Алмон** предполагаются полиномиальные распределения весов текущих и лаговых значений регрессоров $b_j=c_0+c_1j+c_2j^2+\ldots+c_kj^k$. Тогда уравнение регрессии примет вид

$y_t=a+c_0z_0+c_1z+c_2z^2+\ldots+c_kz^k+\varepsilon_t$, где $z_i := \sum_{j=1}^{k} j^i \cdot x_{t-j}$, $i=1,2,\ldots,k$; $j=1,2,\ldots,p$.

Схема расчёта параметров модели с распределённым лагом методом Алмон:

1) устанавливается максимальная величина лага l;

2) устанавливается степень k полинома для описания структуры лага;

3) рассчитываются значения переменных $z_0, z_1,\ldots, z_k$;

4) определяются параметры уравнения линейной регрессии y_t от z_i;

5) рассчитываются параметры исходной модели с распределённым лагом.

Модели с лаговыми значениями зависимой переменной в качестве регрессоров называют **авторегрессионными**. Например, в модели $y_t=a+b_0x_t+c_1y_{t-1}+\varepsilon_t$ коэффициент b_0 характеризует краткосрочное изменение y_t под воздействием изменения регрессора x_t на 1 единицу.

Долгосрочный мультипликатор равен сумме краткосрочного и промежуточных мультипликаторов: $b := b_0 + b_0c_1 + b_0c_1^2 + b_0c_1^3 + \ldots = b_0\left(1 + c_1 + c_1^2 + c_1^3 + \ldots\right) = \frac{b_0}{1-c_1}$.

14.5. Условия применимости МНК для идентификации авторегрессий. Проверка гетероскедастичности остатков.

При малом объёме выборки для оценки гетероскедастичности используют **метод Голдфелда–Квандта**. Его алгоритм:

1) ранжировать n наблюдений по возрастанию переменной x;

2) исключить C центральных наблюдений: $\frac{n-C}{2} > p$ – **число идентифицируемых параметров**;

3) разделить оставшиеся (n–C) наблюдений на 2 группы (соответственно, с *малыми* и *большими* значениями регрессора x) и определить для каждой группы своё уравнение регрессии;

4) определить отношение $R := S_1/S_2$ остаточных сумм квадратов для первой (S_1) и второй (S_2) групп;

5) при выполнении нулевой гипотезы H_0 о гомоскедастичности R удовлетворяет F-критерию с $\frac{n-C-2p}{2}$ степенями свободы; чем больше R превышает табличное (критическое) значение F-критерия, тем больше нарушена предпосылка о равенстве дисперсий остатков.

15.1. Системы эконометрических уравнений; их классификация.

Классификация систем эконометрических уравнений:

1) **система независимых уравнений** – каждая зависимая переменная (*y*) рассматривается как функция одного набора регрессоров (*x*):

$$\begin{cases} y_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1m}x_m + \varepsilon_1, \\ y_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2m}x_m + \varepsilon_2, \\ \ldots \\ y_n = a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nm}x_m + \varepsilon_n. \end{cases}$$

Для её решения используется МНК.

2) **система рекурсивных уравнений** – зависимая переменная (*y*) одного уравнения является регрессором в следующем уравнении:

$$\begin{cases} y_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1m}x_m + \varepsilon_1, \\ y_2 = b_{21}y_1 + a_{21}x_1 + a_{22}x_2 + \ldots + a_{2m}x_m + \varepsilon_2, \\ y_3 = b_{31}y_1 + b_{32}y_2 + a_{21}x_1 + a_{22}x_2 + \ldots + a_{2m}x_m + \varepsilon_2, \\ \ldots \\ y_n = b_{n1}y_1 + b_{n2}y_2 + \ldots + b_{n,n-1}y_{n-1} + a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nm}x_m + \varepsilon_n. \end{cases}$$

Для её решения используется МНК, который применяется к уравнениям системы по очереди, начиная с первого.

3) **система взаимосвязанных (одномоментных) уравнений** – зависимые переменные (*y*) в одних уравнениях входят в левую часть, а в других – в правую:

$$\begin{cases} y_1 = b_{12}y_2 + b_{13}y_3 + \ldots + b_{1n}y_n + a_{11}x_1 + a_{12}x_2 + \ldots + a_{1m}x_m + \varepsilon_1, \\ y_2 = b_{21}y_1 + b_{23}y_3 + \ldots + b_{2n}y_n + a_{21}x_1 + a_{22}x_2 + \ldots + a_{2m}x_m + \varepsilon_2, \\ \ldots \\ y_n = b_{n1}y_1 + b_{n2}y_2 + \ldots + b_{n,n-1}y_{n-1} + a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nm}x_m + \varepsilon_n. \end{cases}$$

Её называют ещё **структурной формой модели**, а её коэффициенты – **структурными**. МНК для её решения не применим, т.к. он даёт несостоятельные оценки идентифицируемых параметров.

Эндогенные переменные – взаимозависимые переменные (y), определённые внутри модели.

Экзогенные переменные – независимые переменные (x), определённые вне модели.

Предопределённые переменные – экзогенные и лаговые (за предыдущие моменты времени) эндогенные переменные системы.

Система линейных зависимостей всех эндогенных переменных от всех предопределённых называется **приведённой формой модели**:

$$\begin{cases} \tilde{y}_1 = \delta_{11}x_1 + \delta_{12}x_2 + \ldots + \delta_{1m}x_m, \\ \tilde{y}_2 = \delta_{21}x_1 + \delta_{22}x_2 + \ldots + \delta_{2m}x_m, \\ \ldots \\ \tilde{y}_n = \delta_{n1}x_1 + \delta_{n2}x_2 + \ldots + \delta_{nm}x_m. \end{cases}$$

Её коэффициенты называются **приведёнными.**

15.2. Идентифицируемость и идентификация уравнений системы.

Счётное правило (необходимое условие идентифицируемости):

$D+1=H$ – уравнение идентифицируемо;

$D+1<H$ – уравнение неидентифицируемо;

$D+1>H$ – уравнение сверхидентифицируемо;

где H – число эндогенных переменных в уравнении;

D – число предопределённых переменных системы, отсутствующих в уравнении.

Достаточное условие идентифицируемости – определитель матрицы, составленной из коэффициентов при переменных, отсутствующих в исследуемом уравнении, не равен нулю, а ранг этой матрицы не менее числа эндогенных переменных системы без единицы.

15.3. Косвенный и двухшаговый МНК; их сравнение.

Для идентификации параметров идентифицируемого уравнения применяется **косвенный метод наименьших квадратов** (КМНК), для сверхидентифицируемого – **двухшаговый метод наименьших квадратов** (ДМНК).

Алгоритм КМНК:

1) составить приведённую форму модели и определить параметры каждого уравнения обычным МНК;

2) подставить регрессионные уравнения в уравнение структурной формы и путём алгебраических преобразований выразить оценки структурных коэффициентов через приведённые.

Алгоритм ДМНК:

1) составить приведённую форму модели и определить параметры каждого уравнения обычным МНК;

2) заменить эндогенные переменные в правой части идентифицируемого уравнения регрессионными выражениями, полученными в приведённой форме модели;

3) обычным МНК определить параметры структурного уравнения, взяв в качестве исходных данных фактические значения предопределённых переменных и расчётные значения эндогенных переменных, стоящих в правой части исследуемого структурного уравнения.

Литература.

1. Тихомиров Н.П., Дорохина Е.Ю. Эконометрика: Учебник. – М.: Экзамен, 2003. – 512 с.

2. Эконометрика: Учебник / Под ред. И.И.Елисеевой. – М.: Проспект, 2009. – 288 с.

3. Эконометрика: Учебник / Под ред. И.И.Елисеевой. – 2-е изд. – М.: Финансы и статистика, 2007. – 576 с.

4. Практикум по эконометрике: Учеб. пособие / Под ред. И.И.Елисеевой. – 2-е изд. – М.: Финансы и статистика, 2006. – 344 с.