Эконометрика

Преподаватель доц., к.э.н. Хайрулин Ильяс Гаяревич
Кафедра математических методов в экономике

Литература

- Магнус Я.Р., Катышев П.К., Пересецкий А.А. - Эконометрика. Начальный курс: учебник.
- Елисеева И.И. – Эконометрика: учебник
- P.Newbold – Statistics for Business & Economics

Эконометрика

- «Эконометрика — это не то же самое, что экономическая статистика. Она не идентична и тому, что мы называем экономической теорией, хотя значительная часть этой теории носит количественный характер. Эконометрика не является синонимом приложений математики к экономике. Как показывает опыт, каждая из трех отправных точек — статистика, экономическая теория и математика — необходимое, но не достаточное условие для понимания количественных соотношений в современной экономической жизни. Это — единство всех трех доставляющих. И это единство образует эконометрику» (Рагнар Фриш, 1933г.)
- **Эконометрика — это наука, которая дает количественное выражение взаимосвязей экономических явлений и процессов**

Место эконометрики в управленческом процессе

Задачи, решаемые эконометрическим методом

- Классификация
- Обоснование связи и выявление значимых факторов
- Проверка гипотез экономической теории
- Прогнозирование
- Анализ

Рис. 1. *Схема анализа*

Этапы эконометрического исследования

1. постановка проблемы
2. получение данных и анализ их качества
3. спецификация модели
4. оценка параметров
5. проверка качества (адекватности) модели
6. интерпретация результатов

Этапы (подробнее)

- качественный анализ связей экономических переменных — выделение зависимых (y) и независимых переменных (x);
- подбор данных;
- спецификация формы связи между y и x;
- оценка параметров модели;
- проверка ряда гипотез о свойствах распределения вероятностей для случайной компоненты (гипотезы о средней, дисперсии и ковариации);
- анализ мультиколлинеарности объясняющих переменных, оценка ее статистической значимости, выявление переменных, ответственных за мультиколлинеарность;
- введение фиктивных переменных;
- выявление автокорреляции, лагов;
- выявление тренда, циклической и случайной компонент;
- проверка остатков на гетероскедастичность;
- анализ структуры связей и построение системы одновременных уравнений;
- проверка условия идентификации;
- оценивание параметров системы одновременных уравнений (двухшаговый и трехшаговый метод наименьших квадратов, метод максимального правдоподобия);
- моделирование на основе системы временных рядов: проблемы стационарности и коинтеграции;
- построение рекурсивных моделей, ARIMA- и VAR- моделей;
- проблемы идентификации и оценивания параметров.

Сбор данных

При моделировании экономических процессов используют следующие типы данных:

- пространственные данные

Пространственными данными является набор сведений по разным объектам, взятым за один и тот же период или момент времени. Например, набор сведений по разным фирмам (объем производства, численность работников, размер основных производственных фондов и пр.).

- временные данные

Временными данными является набор сведений, характеризующий один и тот же объект, но за разные периоды или моменты времени. Например, ежеквартальные данные о средней заработной плате, индексе потребительских цен, числе занятых за последние годы, ежедневный курс доллара США. Отличительной особенностью временных данных является то, что они естественным образом упорядочены по времени.

- панельные данные

Панельными данными является набор сведений по разным объектам, взятый за интервал времени. То есть множество объектов наблюдается в течение определенного времени.

Типы переменных

Типы переменных, участвующих в эконометрической модели:

- эндогенные (зависимые) — значения которых определяются внутри модели, или взаимозависимые (y);
- экзогенные (независимые) – значения которых задаются извне, автономно, в определенной степени они являются управляемыми (планируемыми) (x)
- лаговые — экзогенные или эндогенные переменные эконометрической модели, датированные предыдущими моментами времени и находящиеся в уравнении с текущими переменными. Например: y_t — текущая эндогенная переменная, y_{t-1} — лаговая эндогенная переменная, y_{t-2} — тоже лаговая эндогенная переменная;

Предопределенные переменные (объясняющие переменные). К ним относятся лаговые и текущие экзогенные переменные (x_t , x_{t-1}), а также лаговые эндогенные переменные (y_{t-1}).

Любая эконометрическая модель предназначена для объяснения значений текущих эндогенных переменных (одной или нескольких) в зависимости от значений предопределенных переменных.

Спецификация моделей

Выделяют три основных класса моделей.

I. Регрессионные модели с одним уравнением (факторов может быть один или несколько)

- Линейные
- Нелинейные

II. Модели временных рядов, полученные с помощью следующих методов

- Экспоненциального сглаживания
- Сезонной декомпозиции
- Авторегрессии
- ARIMA и др.

III. Системы одновременных уравнений

Линейность и аддитивность

- Функция нескольких переменных $y=f(x_1,...,x_n)$ **линейна** по всем независимым переменным тогда и только тогда, когда dy/dx_i не включает x_i, то есть когда $d(dy/dx_i)=0$, эффект данного изменения по x_i не зависит от уровня x_i. В противном случае **нелинейна**
- Функция является **аддитивной** по x_i тогда и только тогда, когда dy/x_i не включает x_j , т.е. тогда, когда $d(dy/dx_i)/dx_j=0$, эффект данного изменения по каждой независимой переменной не зависит от уровня другой переменной. В противном случае **мультипликативна**

Примеры оценки линейности функций

Функция $f(x_1, x_2)$	$\frac{df}{dx_1}$	$\frac{df}{dx_2}$	Линейность по x_1	Линейность по x_2	Аддитивность по x_1, x_2
$a_1x^2_1 + a_2x^2_2 + a_3x_1x_2$	$2a_1x_1 + a_3x_2$	$2a_2x_2 + a_3x_1$	нет	нет	нет
x_2/x_1	$-x_2/x^2_1$	$1/x_1$	нет	да	нет
$a_1x^2_1 + a_2x_2$	$2a_1x_1$	a_2	нет	да	да
$a_1x_1x^2_2 + a_2\log x_2$	$a_1x^2_2$	$2a_1x_1x_2 + \frac{a_2}{x_2}$	да	нет	нет
$a_1x_1 + a_2x_2 + a_3x_1x_2$	$a_1 + a_3x_2$	$a_2 + a_3x_1$	да	да	нет
$a_1x_1 + a_2\log x_2$	a_1	a_2/x_2	да	нет	да
$x^{a_1}_1 \cdot x^{a_2}_2$	$a_1x_1{}^{a_1-1}x^{a_2}_2$	$a_2x^{a_1}_1x^{a_2-1}_2$	нет	нет	нет
$a_1x_1 + a_2x_2$	a_1	a_2	да	да	да

Оценка параметров

Этот этап предполагает нахождение неизвестных элементов в модели тем или иным способом.

Наиболее распространенным методом является МНК. МНК применяется к моделям, линейным по параметрам. Если функция регрессии нелинейна по параметрам, необходима её предварительная линеаризация.

Если распределение остатков ненормально, то наилучшим методом их оценки будет не МНК, а ММП.

Также если не выполняются предпосылки МНК, то для нахождения параметров можно использовать ММП.

Проверка качества модели

Это важнейший этап, заключающийся в определении следующего:

- погрешности расчетов
- точности предсказания по модели (доверительный интервал прогноза)
- устойчивости модели к выборке (проверка по тестам Стьюдента и Фишера)

Интерпретация результатов

- Модель должна быть достаточно проста и отражать экономические взаимосвязи. В ином случае параметры не будут интерпретируемы.
- Однако если модель строилась исключительно для прогноза, требования к экономической интерпретации смягчаются.

Парная регрессия

Базовые термины и идеи

- Генеральная совокупность (population) (иногда используется калька с англоязычного термина – «популяция») – все множество объектов, в отношении которых формулируется исследовательская гипотеза
- Выборка (sample) – ограниченная по численности группа объектов (респондентов), отбираемая из генеральной совокупности для изучения ее свойств
- Сплошное и выборочное исследование
- Репрезентативность выборки (representativeness of sample) – способность выборки представлять изучаемые явления достаточно полно с точки зрения их изменчивости в генеральной совокупности
- Любое исследование направлено на определение некоторой характеристики или выявление связи между признаками
- Связь может характеризоваться не только величиной (степенью связи) и направлением, но также и надежностью или статистической достоверностью (statistical confidence) - эта характеристика связи показывает, можно ли распространить результаты, полученные на данной *выборке,* на всю генеральную совокупность, из которой взята эта выборка

Парная регрессия

Парная регрессия – это уравнение, описывающее корреляционную связь между парой переменных: зависимой переменной (результатом) y и независимой переменной (фактором) x.

$$y=f(x)$$

Функция может быть как линейной $y=\alpha+\beta x+\varepsilon$, так и нелинейной $y=\alpha+\beta/x+\varepsilon$, $y=\alpha+\beta lnx+\varepsilon$, $y=\alpha\beta^{x}\varepsilon$.

Парная линейная регрессия

Предположим, что для генеральной совокупности связь выражается уравнением

$$y_i=\alpha+\beta x_i+\varepsilon_i,\ i=1,\ldots,N$$

y_i – i-е фактическое значение зависимой переменной y;

x_i – i-е значение независимой переменной x;

α и β – генеральные параметры парной линейной регрессии

N – объем генеральной совокупности

ε_i – теоретические ошибки, в силу наличия еще и других объясняющих факторов, не учтенных в модели или ошибок спецификации

Парная линейная регрессия

- Но у нас есть ограниченное объективными причинами кол-во наблюдений (выборка), для которых мы на практике можем построить уравнение $\hat{y}_i = a + bx_i$

$$y_i = a + bx_i + e_i,\ i = 1,\ldots,n$$

y_i – i-е фактическое значение зависимой переменной y;

$\hat{y}_i$ - i-е значение зависимой переменной, рассчитанное по уравнению прямой

x_i – i-е значение независимой переменной x;

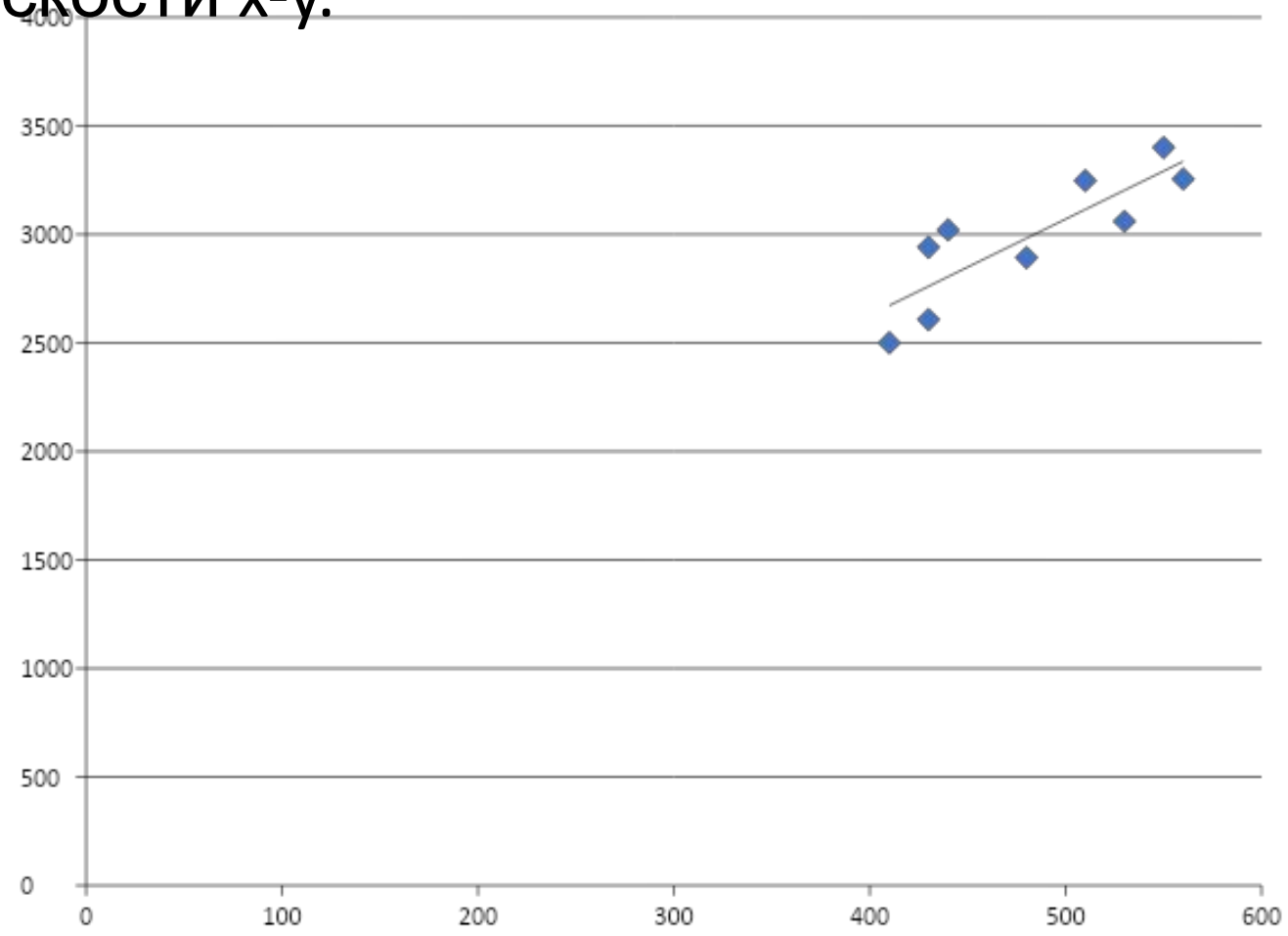
a и b – выборочные параметры парной линейной регрессии

n – объем выборки

e_i – наблюдаемые остатки, в силу наличия еще и других объясняющих факторов, не учтенных в модели или ошибок спецификации

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

Буквально перед нами стоит задача провести прямую линию через множество точек в плоскости x-y.



Построение прямой через множество точек

Два параметра a и b определяют наклон прямой и сдвиг по вертикали. Существует много способов провести прямую через точки на плоскости.

1) можно проводить прямую через две произвольные точки
2) пытаться минимизировать сумму квадратов остатков $\sum_{i=1}^{n}(y_i - a - bx_i)^2$
3) пытаться минимизировать сумму модулей отклонений $\sum_{i=1}^{n}|y_i - a - bx_i|$
4) любая другая мера учета отклонений $\sum_{i=1}^{n} g(y_i - a - bx_i)$, например функция Хубера, которая при малых отклонениях квадратична, а при больших линейна

Построение прямой через множество точек

Каждый метод выбора a и b обладает плюсами и минусами.

Построение по двум точкам неустойчиво к выбору таких точек и может давать противоречивые результаты.

Сумма квадратов отклонений проста в вычислениях, обладает хорошими статистическими свойствами, позволяет выстроить теорию для проверки различных статистических гипотез, но чувствительна к «выбросам»

Сумма модулей отклонений нечувствительна к «выбросам», но сложна в вычислении

Функция Хубера пытается совместить плюсы обеих мер.

Метод наименьших квадратов (МНК)

- Рассмотрим задачу наилучшей аппроксимации набора наблюдений x_i, y_i линейной функцией $\widehat{y_i}=a+bx_i$ в смысле минимизации функционала

$$F=\sum_{i=1}^{n}(y_i - a - bx_i)^2$$

- то есть задачу можно сформулировать следующим образом: имея в наличии набор данных x_i, y_i подобрать значения a и b, чтобы функция $F(a,b)$ была минимальна. Эта задача безусловной нелинейной оптимизации решается через нахождение экстремума функции двух переменных. Найдем экстремум функции двух переменных.

Метод наименьших квадратов (МНК)

- Для этого вычислим производные функции F по параметрам a и b

$$\begin{cases} \dfrac{\partial F}{\partial a} = -2\sum_{i=1}^{n}(y_i - a - bx_i) = 0 \\ \dfrac{\partial F}{\partial b} = -2\sum_{i=1}^{n} x_i(y_i - a - bx_i) = 0 \end{cases}$$

Метод наименьших квадратов (МНК)

- Раскроем скобки и получим

$$\begin{cases} an + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

Стандартная форма нормальных уравнений

Метод наименьших квадратов (МНК)

- Решая систему, получим значения a и b

$$\begin{cases} b = \dfrac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \dfrac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} \\ a = \bar{y} - b\bar{x} \end{cases}$$

Так мы нашли неизвестные параметры модели $\hat{y}_i = a + bx_i$

Экономическая интерпретация a и b

- Коэффициент b показывает среднее изменение результативного признака (в единицах измерения y) при изменении величины фактора x на 1 единицу его измерения.
- Коэффициент a показывает среднее значение результативного признака при x=0, если практически x может принимать нулевое значение. В ином случае, коэффициент a не имеет экономической интерпретации.

Коэффициент корреляции

- Наряду с оценками a и b часто сразу оценивают тесноту связи между случайными величинами x и y с помощью линейного коэффициента корреляции r_{xy}

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}} = \frac{\overline{xy} - \bar{x}\bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2}\sqrt{\overline{y^2} - \bar{y}^2}}$$

Величина коэффициента корреляции	0.1 - 0.3	0.3 - 0.5	0.5 – 0.7	0.7 – 0.9	0.9 – 1.0
Характеристика силы связи	слабая	умеренная	заметная	высокая	весьма высокая

средняя сильная

Шкала Чеддока

$r_{xy}>0$ – связь прямая, $r_{xy}<0$ – связь обратная

До этого нас интересовало только качество подгонки прямой к данным. Теперь добавим к постановке задачи некоторые статистические свойства данных. Для одного и того же x_i мы можем наблюдать разные значения y_i. К примеру x-доход семьи, y-расходы на питание. Две семьи с одинаковым доходом могут тратить разное количество денег на питание. Из-за этого у наблюдений будут разные отклонения от расчётных значений, то есть разные ошибки.

Какова природа ошибки ε_i? Откуда берутся отличия фактического значения от расчетного?

1) Наша модель является упрощением действительности и на самом деле есть еще другие параметры, от которых зависит y. Расходы на питание могут также зависеть от региона проживания, количества членов семьи, образа жизни, склонности к потреблению.

2) Трудности в измерении данных (присутствуют ошибки измерения).

Можно считать, что ε_i – случайная величина с некоторой функцией распределения, которой соответствует функция распределения случайной величины y_i.

Основные гипотезы

1. $y_i=\alpha+\beta x_i+\varepsilon_i$, $i=1,\ldots,N$ – спецификация модели
2. x_i - детерминированная величина, где x_i-разные величины
3. $M[\varepsilon_i]=0$, $M[\varepsilon_i]=D[\varepsilon_i]=\sigma^2$ не зависит от x_i или от t
4. $M[\varepsilon_i, \varepsilon_j]=0$ – некоррелированность ошибок для разных наблюдений
5. Ошибки ε_i имеют совместное нормальное распределение $N(0, \sigma^2)$

В этом случае модель называется классической нормальной линейной регрессионной моделью. (Classical Normal Linear Regression Model)

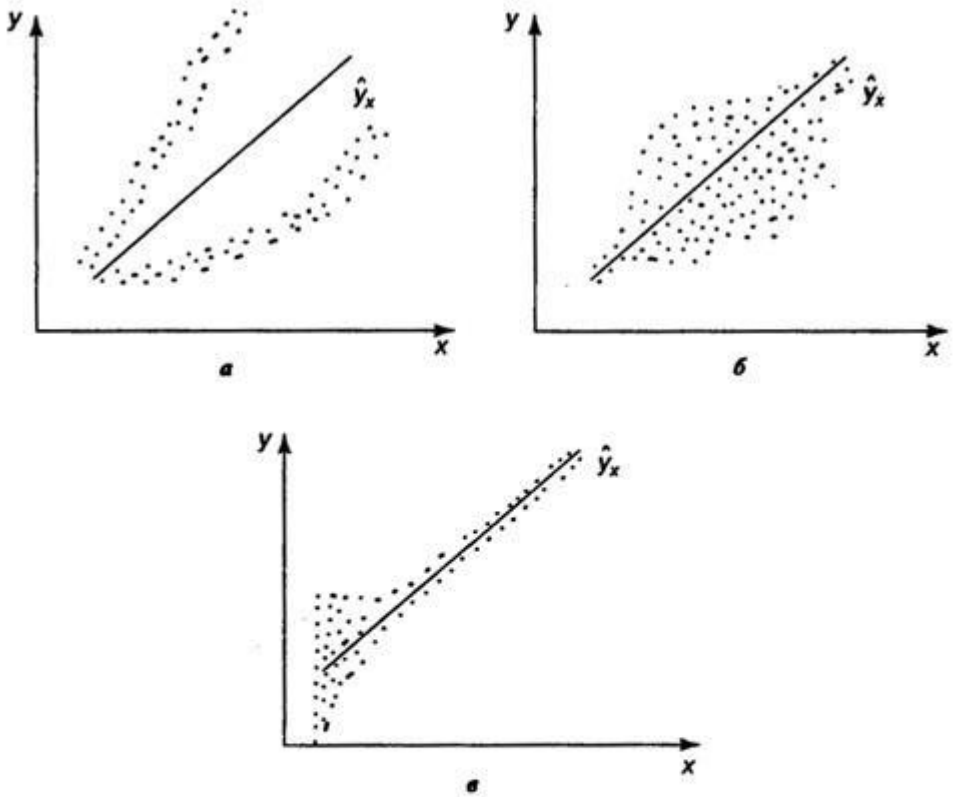
Основные гипотезы

1,2. спецификация модели отражает наше преставление о механизме зависимости y_i от x_i и сам выбор объясняющей переменной x_i . Чтобы установить влияние x_i они должны принимать различные значения.

Основные гипотезы

3. $M[\varepsilon_i]=0$, $M[\varepsilon_i]=D[\varepsilon_i]=\sigma^2$ не зависит от x_i или от t

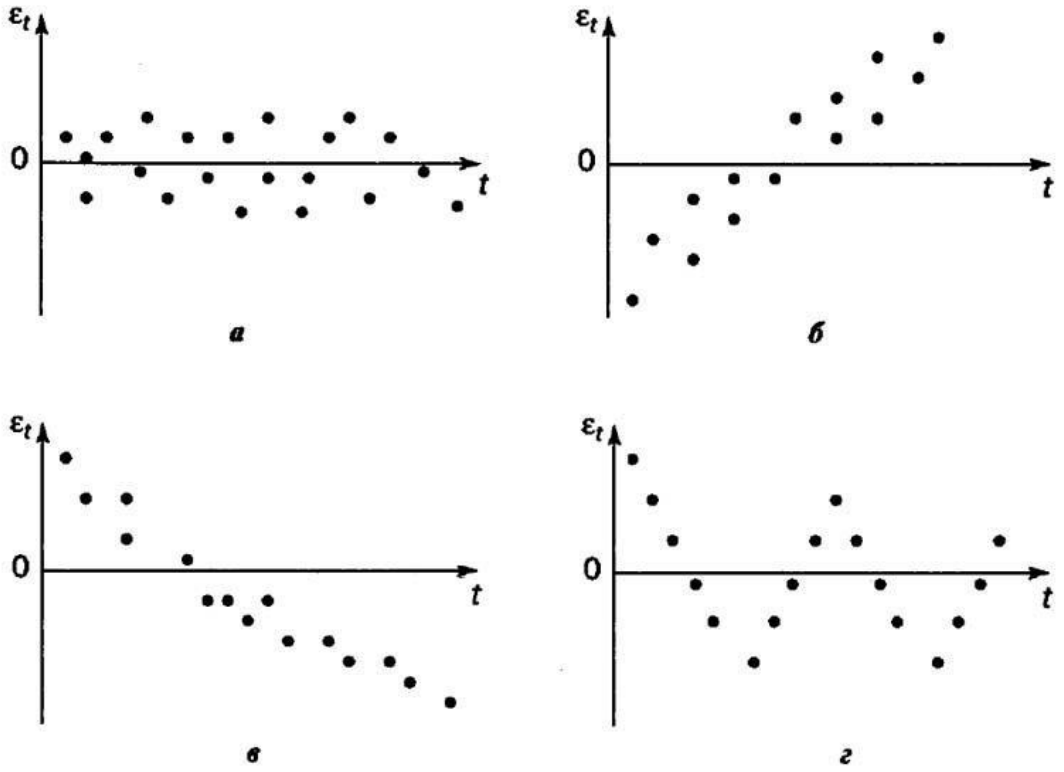
Условие независимости дисперсии ошибки от номера наблюдения или x_i называется гомоскедастичностью. В противоположном случае, наблюдают явление гетероскедастичности



Основные гипотезы

4. $M[\varepsilon_i, \varepsilon_j]=0$ – некоррелированность ошибок для разных наблюдений

В случае, когда это условие не выполняется, говорят об автокорреляции ошибок. Часто такое происходит с временным выборками ил



Теорема Гаусса-Маркова

Задача теперь- статистически оценить три параметра: a,b, σ^2

В предположениях модели:

1. $y_i=\alpha+\beta x_i+\varepsilon_i$, $i=1,\ldots,N$ – спецификация модели
2. x_i - детерминированная величина, где x_i-разные величины
3. $M[\varepsilon_i]=0$, $M[\varepsilon_i]=D[\varepsilon_i]=\sigma^2$ не зависит от x_i или от t
4. $M[\varepsilon_i, \varepsilon_j]=0$ – некоррелированность ошибок для разных наблюдений

оценки a и b, полученные по методу наименьших квадратов(МНК), имеют наименьшую дисперсию (то есть эффективны) в классе всех линейных несмещенных оценок.

Статистические свойства оценок

- Статистические оценки (или просто оценки)— это статистики, которые используются для оценивания неизвестных параметров распределений случайной величины.
- Несмещённая оце́нка в математической статистике — это точечная оценка, математическое ожидание которой равно оцениваемому параметру. $M[b]=\beta$
- Эффективная оценка – это несмещенная оценка, имеющая наименьшую дисперсию из всех возможных несмещенных оценок данного параметра. $D[b]<D[b^*]$

Задача

Пусть X_1, X_2, X_3, X_4 — случайная выборка значений из генеральной нормальной совокупности со средним и дисперсией . Рассмотрим две оценки :

$$\hat{\mu}^{(1)} = \frac{X_1 + 2X_2 + 3X_3 + 4X_4}{10}, \qquad \hat{\mu}^{(2)} = \frac{X_1 + 4X_2 + 4X_3 + X_4}{10}.$$

1. Покажите, что обе оценки несмещенные.
2. Какая из оценок более эффективна?
3. Найдите относительную эффективность двух оценок.
4. Найдите несмещенную оценку, более эффективную, чем каждая из двух оценок .

Дисперсия ошибок и ее оценка

- σ^2 – дисперсия теоретических ошибок (то есть для генеральной совокупности)

$$\sigma^2 = \frac{1}{n} * \sum_{i=1}^{n} e_i^2$$

- S^2 – несмещенная оценка дисперсии ошибок σ^2 – то есть то, что мы можем наблюдать в выборке.

$$S^2 = n/(n-2) * \sigma^2 = \frac{1}{n-2} * \sum_{i=1}^{n} e_i^2$$

Дисперсии параметров a и b

- Величина дисперсии остатков напрямую влияет на дисперсию оценок a и b. (Напоминаю: a и b – случайные величины)
- $D[b]=\frac{S^2}{\sum(x_i-\bar{x})^2}$
- $D[a]=\frac{S^2*\sum x_i^2}{n*\sum(x_i-\bar{x})^2}$

Вывод формул дисперсии остатков в Я.Р. Магнус-Эконометрика:начальный курс

Распределение оценок a и b

- Так как a и b являются линейными функциями от y, то они тоже имеют нормальное распределение

$a \sim N(\alpha, \frac{\sigma^2 * \sum x_i^2}{n * \sum (x_i - \bar{x})^2})$ $b \sim N(\beta, \frac{\sigma^2}{\sum (x_i - \bar{x})^2})$

Значит $(a-\alpha)/\sqrt{\frac{\sigma^2 * \sum x_i^2}{n * \sum (x_i - \bar{x})^2}} \sim N(0,1)$, аналогично b.

Но поскольку мы не знаем σ^2, то заменяем ее на S^2, при этом $t_a = (a-\alpha)/\sqrt{\frac{S^2 * \sum x_i^2}{n * \sum (x_i - \bar{x})^2}} \sim t(n-2)$, и аналогично для **b**

$t_b = (b-\beta)/\sqrt{\frac{S^2}{\sum (x_i - \bar{x})^2}} \sim t(n-2)$

Тест значимости параметров по Стьюденту

- Статистику t_b можно использовать для проверки статистической гипотезы H0: $\beta=\beta_0$ против альтернативной гипотезы H1: $\beta\neq\beta_0$. Наиболее просто выглядит гипотеза H0: $\beta=0$ (в генеральной совокупности связи нет). Тогда

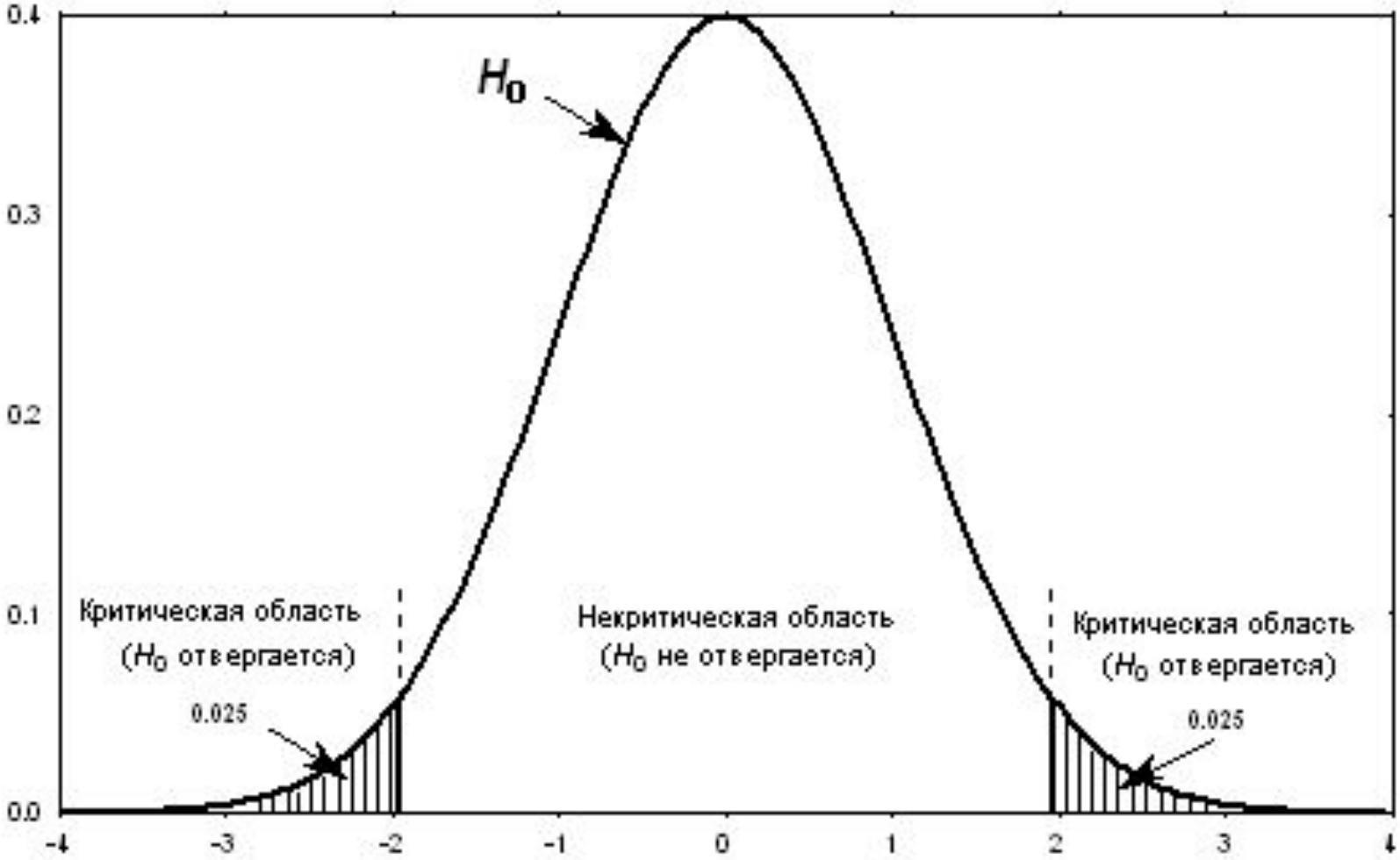
$$t_b=b/s_b \sim t(n-2), \text{ где } s_b=\sqrt{\frac{S^2}{\sum(x_i-\bar{x})^2}}$$

Зададимся, например 2,5% точкой t-распределения с (n-2) степенями свободы $t_{0,025}$, т.е. $P(-t_{0,025} < t_b < t_{0,025})=0,95$

Мы отвергаем гипотезу H_0 (и принимаем H_1) на 5% уровне значимости, если $|t_b|>t_{0,025}$ («редкое» событие с точки зрения гипотезы H_0), в противном случае мы не можем отвергнуть H_0 (и принимаем H_0). Вероятность найти связь там, где ее на самом деле нет ($\beta=0$, а $|t_b|>t_{0,025}$) называется ошибкой первого рода и не превышает уровня значимости.

Аналогично для проверки значимости a используется статистика t_a

Тест значимости параметров по Стьюденту



Число степеней свободы	Доверительные вероятности				
	0,9	0,95	0,98	0,99	0,999
1	6,31375	12,7062	31,821	63,6559	636,578
2	2,91999	4,30266	6,96455	9,924988	31,5998
3	2,35336	3,18245	4,54071	5,840848	12,9244
4	2,13185	2,77645	3,74694	4,60408	8,61008
5	2,01505	2,57058	3,36493	4,032117	6,8685
6	1,94318	2,44691	3,14267	3,707428	5,95872
7	1,89458	2,36462	2,99795	3,499481	5,40807
8	1,85955	2,30601	2,89647	3,355381	5,04137
9	1,83311	2,26216	2,82143	3,249843	4,78089
10	1,81246	2,22814	2,76377	3,169262	4,58676
11	1,79588	2,20099	2,71808	3,105815	4,43688
12	1,78229	2,17881	2,68099	3,054538	4,31784
13	1,77093	2,16037	2,6503	3,012283	4,22093
14	1,76131	2,14479	2,62449	2,976849	4,14031
15	1,75305	2,13145	2,60248	2,946726	4,07279
16	1,74588	2,1199	2,58349	2,920788	4,01487
17	1,73961	2,10982	2,56694	2,898232	3,96511
18	1,73406	2,10092	2,55238	2,878442	3,92174
19	1,72913	2,09302	2,53948	2,860943	3,88332
20	1,72472	2,08596	2,52798	2,845336	3,84956
21	1,72074	2,07961	2,51765	2,831366	3,8193
22	1,71714	2,07388	2,50832	2,818761	3,79223
23	1,71387	2,06865	2,49987	2,807337	3,76764
24	1,71088	2,0639	2,49216	2,796951	3,74537
25	1,70814	2,05954	2,4851	2,787438	3,72514
26	1,70562	2,05553	2,47863	2,778725	3,70666
27	1,70329	2,05183	2,47266	2,770685	3,68949
28	1,70113	2,04841	2,46714	2,763263	3,67392
29	1,69913	2,04523	2,46202	2,756387	3,65952
30	1,69726	2,04227	2,45726	2,749985	3,64598
60	1,67065	2,0003	2,39012	2,660272	3,46015
120	1,65765	1,97993	2,35783	2,617417	3,37342

Доверительные интервалы параметров α и β

Разрешив неравенство $P\{|(b-\beta)/S_b|<t_{0,025}\}=0,95$ относительно **β** получим

$$P\{b-t_{0,025}*S_b<\beta<b+t_{0,025}*S_b\}=0,95$$

То есть в интервал $[b-t_{0,025}*S_b ; b+t_{0,025}*S_b]$ истинный параметр β попадет с вероятностью 95%.

Аналогично составляется доверительный интервал для α.

Качество модели. Дисперсионный анализ.

- Рассмотрим вариацию (разброс) значений y_i вокруг своего среднего значения $\sum(y_i - \bar{y})^2$

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

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

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

Качество модели. Дисперсионный анализ.

- Можно доказать, что третье слагаемое равно нулю. $\sum(y_i - \hat{y_i})(\hat{y_i} - \bar{y}) = 0$
- Тогда

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

TSS(total) ESS(error) RSS(regression)

Многомерная теорема Пифагора

Качество модели. Дисперсионный анализ.

Средняя ошибка аппроксимации

$$\overline{A} = \frac{1}{n}\sum\left|\frac{y-\hat{y}}{y}\right| \cdot 100\ \% \ .$$

Коэффициент детерминации

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

R^2 принимает значения от 0 до 1 и показывает долю дисперсии результативного признака, объясненную регрессией. Если $\sum_{i=1}^{n} e_i^2$ велика, то модель не объясняет вариацию (изменчивость) результата, и R^2 близко к 0.

Проверка значимости уравнения по критерию Фишера

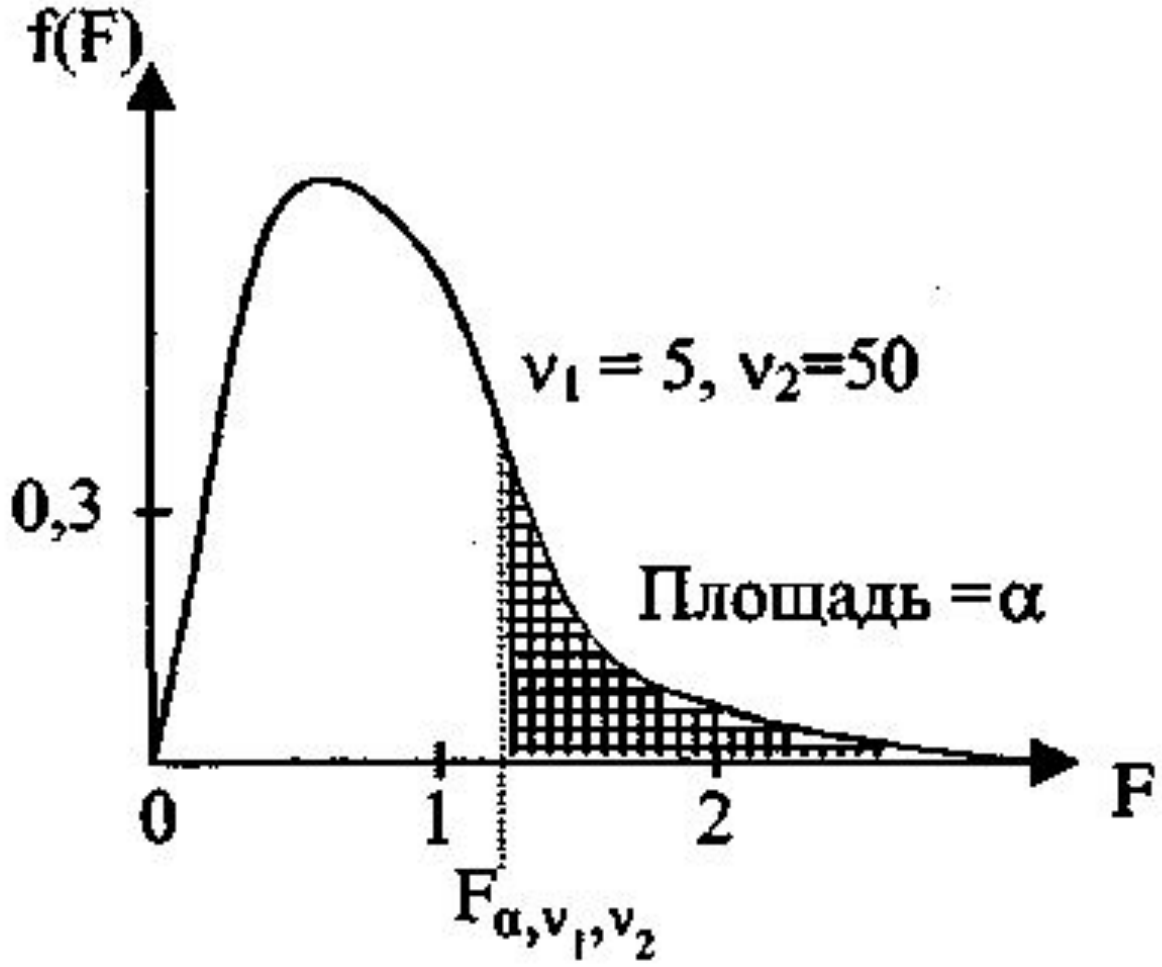
- Под незначимостью модели понимается в общем виде одновременное равенство коэффициентов перед всеми факторами x, или, что то же самое, равенство $\widehat{y_i}$=a=$\bar{y}$, и график функции регрессии параллелен оси абсцисс. Тогда $R^2 = \frac{\sum(\widehat{y_i}-\bar{y})^2}{\sum(y_i-\bar{y})^2} = 0.$

Проверим данную гипотезу H_0: $R^2 = 0$ против H_1: $R^2 \neq 0$.

Для этого составляется статистика Фишера

$\boldsymbol{F = \frac{R^2}{1-R^2} * \frac{n-m-1}{m}}$~F(m,n-m-1), где m-количество факторов в модели, которая сравнивается с табличным значением распределения Фишера F_α(m,n-m-1) для выбранного уровня значимости α.

Проверка значимости уравнения по критерию Фишера



k_2 \ k_1	1	2	3	4	5	6	8	12	24	∞
1	161,45	199,50	215,72	224,57	230,17	233,97	238,89	243,91	249,04	254,32
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,84	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,90	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,91	2,74	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,79	2,61	2,40
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,69	2,50	2,30
13	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,60	2,42	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,53	2,35	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,48	2,29	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,24	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,19	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,31	2,11	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,08	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,25	2,05	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,23	2,03	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,38	2,20	2,00	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,18	1,98	1,73
25	4,24	3,38	2,99	2,76	2,60	2,49	2,34	2,16	1,96	1,71
26	4,22	3,37	2,98	2,74	2,59	2,47	2,32	2,15	1,95	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,30	2,13	1,93	1,67
28	4,20	3,34	2,95	2,71	2,56	2,44	2,29	2,12	1,91	1,65
29	4,18	3,33	2,93	2,70	2,54	2,43	2,28	2,10	1,90	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,89	1,62
35	4,12	3,26	2,87	2,64	2,48	2,37	2,22	2,04	1,83	1,57
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,00	1,79	1,51
45	4,06	3,21	2,81	2,58	2,42	2,31	2,15	1,97	1,76	1,48
50	4,03	3,18	2,79	2,56	2,40	2,29	2,13	1,95	1,74	1,44
60	4,00	3,15	2,76	2,52	2,37	2,25	2,10	1,92	1,70	1,39
70	3,98	3,13	2,74	2,50	2,35	2,23	2,07	1,89	1,67	1,35
80	3,96	3,11	2,72	2,49	2,33	2,21	2,06	1,88	1,65	1,31
90	3,95	3,10	2,71	2,47	2,32	2,20	2,04	1,86	1,64	1,28
100	3,94	3,09	2,70	2,46	2,30	2,19	2,03	1,85	1,63	1,26
125	3,92	3,07	2,68	2,44	2,29	2,17	2,01	1,83	1,60	1,21
150	3,90	3,06	2,66	2,43	2,27	2,16	2,00	1,82	1,59	1,18
200	3,89	3,04	2,65	2,42	2,26	2,14	1,98	1,80	1,57	1,14
300	3,87	3,03	2,64	2,41	2,25	2,13	1,97	1,79	1,55	1,10
400	3,86	3,02	2,63	2,40	2,24	2,12	1,96	1,78	1,54	1,07
500	3,86	3,01	2,62	2,39	2,23	2,11	1,96	1,77	1,54	1,06
1000	3,85	3,00	2,61	2,38	2,22	2,10	1,95	1,76	1,53	1,03
∞	3,84	2,99	2,60	2,37	2,21	2,09	1,94	1,75	1,52	1,00

Качество модели

F-критерий

$$F_{\text{факт}} = \frac{\Sigma(\hat{y} - \bar{y})^2 / m}{\Sigma(y - \hat{y})^2 / (n - m - 1)} = \frac{R^2}{1 - R^2} \cdot \frac{n - m - 1}{m}.$$

T-критерий

$$t_b = \frac{b}{m_b}; \quad t_a = \frac{a}{m_a}; \quad t_r = \frac{r}{m_r}.$$

$$m_b = \sqrt{\frac{\sum(y - \hat{y}_x)^2 / (n-2)}{\sum(x - \bar{x})^2}} = \sqrt{\frac{S_{\text{ост}}^2}{\sum(x - \bar{x})^2}} = \frac{S_{\text{ост}}}{\sigma_x \sqrt{n}};$$

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

$$m_{r_{xy}} = \sqrt{\frac{1 - r_{xy}^2}{n-2}}.$$

Качество модели

Доверительные интервалы параметров

$$\Delta_a = t_{\text{табл}} m_a, \qquad \Delta_b = t_{\text{табл}} m_b.$$

$$\gamma_a = a \pm \Delta_a; \qquad \gamma_{a_{\min}} = a - \Delta_a; \qquad \gamma_{a_{\max}} = a + \Delta_a;$$

$$\gamma_b = b \pm \Delta_b; \qquad \gamma_{b_{\min}} = b - \Delta_b; \qquad \gamma_{b_{\max}} = b + \Delta_b.$$

Качество модели

Доверительный интервал прогноза

$$m_{\hat{y}_p} = \sigma_{ост} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{\sum (x - \overline{x})^2}},$$

где $\sigma_{ост} = \sqrt{\frac{\Sigma(y - \hat{y})^2}{n - m - 1}}$;

$$\gamma_{\hat{y}_p} = \hat{y}_p \pm \Delta_{\hat{y}_p};$$

где $\Delta_{\hat{y}_p} = t_{табл} \cdot m_{\hat{y}_p}$

Задача

- По семи территориям приуральского района известны значения двух показателей за один год

Район	Расходы на покупку продовольственных товаров в общих расходах, %, *y*	Среднедневная заработная плата одного работающего, руб., *x*
Удмуртская респ.	68,8	45,1
Свердловская обл.	61,2	59,0
Башкортостан	59,9	57,2
Челябинская обл.	56,7	61,8
Пермская обл.	55,0	58,8
Курганская обл.	54,3	47,2
Оренбургская обл.	49,3	55,2

	y	x	yx	x^2	y^2
1	68,8	45,1	3102,88	2034,01	4733,44
2	61,2	59,0	3610,80	3481,00	3745,44
3	59,9	57,2	3426,28	3271,84	3588,01
4	56,7	61,8	3504,06	3819,24	3214,89
5	55,0	58,8	3234,00	3457,44	3025,00
6	54,3	47,2	2562,96	2227,84	2948,49
7	49,3	55,2	2721,36	3047,04	2430,49
Итого	405,2	384,3	22162,34	21338,41	23685,76
Среднее значение	57,89	54,90	3166,05	3048,34	3383,68

Линейная модель на основе МНК

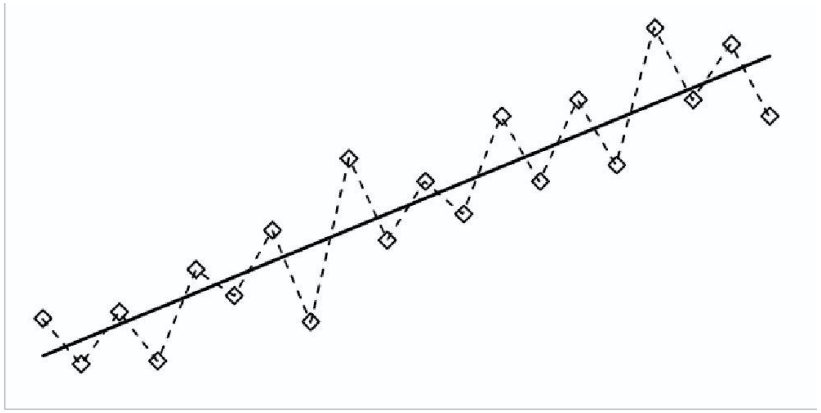
$$\begin{cases} n \cdot a + b \sum x = \sum y, \\ a \sum x + b \sum x^2 = \sum y \cdot x. \end{cases}$$

$$b = \frac{\overline{y \cdot x} - \overline{y} \cdot \overline{x}}{\sigma_x^2} = \frac{3166,05 - 57,89 \cdot 54,9}{5,86^2} \approx -0,35,$$

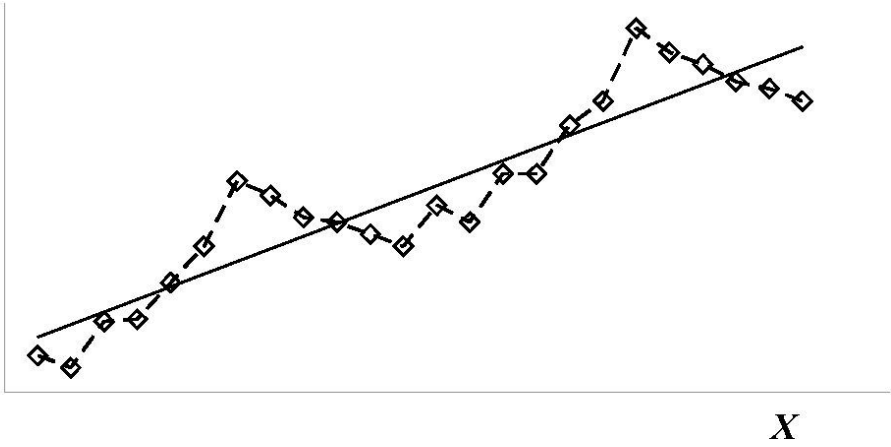
$$a = \overline{y} - b \cdot \overline{x} = 57,89 + 0,35 \cdot 54,9 \approx 76,88.$$

	y	x	yx	x^2	y^2	$\hat{y}_x$	$y - \hat{y}_x$	A_i
1	68,8	45,1	3102,88	2034,01	4733,44	61,3	7,5	10,9
2	61,2	59,0	3610,80	3481,00	3745,44	56,5	4,7	7,7
3	59,9	57,2	3426,28	3271,84	3588,01	57,1	2,8	4,7
4	56,7	61,8	3504,06	3819,24	3214,89	55,5	1,2	2,1
5	55,0	58,8	3234,00	3457,44	3025,00	56,5	–1,5	2,7
6	54,3	47,2	2562,96	2227,84	2948,49	60,5	–6,2	11,4
7	49,3	55,2	2721,36	3047,04	2430,49	57,8	–8,5	17,2
Итого	405,2	384,3	22162,34	21338,41	23685,76	405,2	0,0	56,7
Среднее значение	57,89	54,90	3166,05	3048,34	3383,68	x	x	8,1
σ	5,74	5,86	x	x	x	x	x	x
σ^2	32,92	34,34	x	x	x	x	x	x

Автокорреляция остатков



Отрицательная автокорреляция



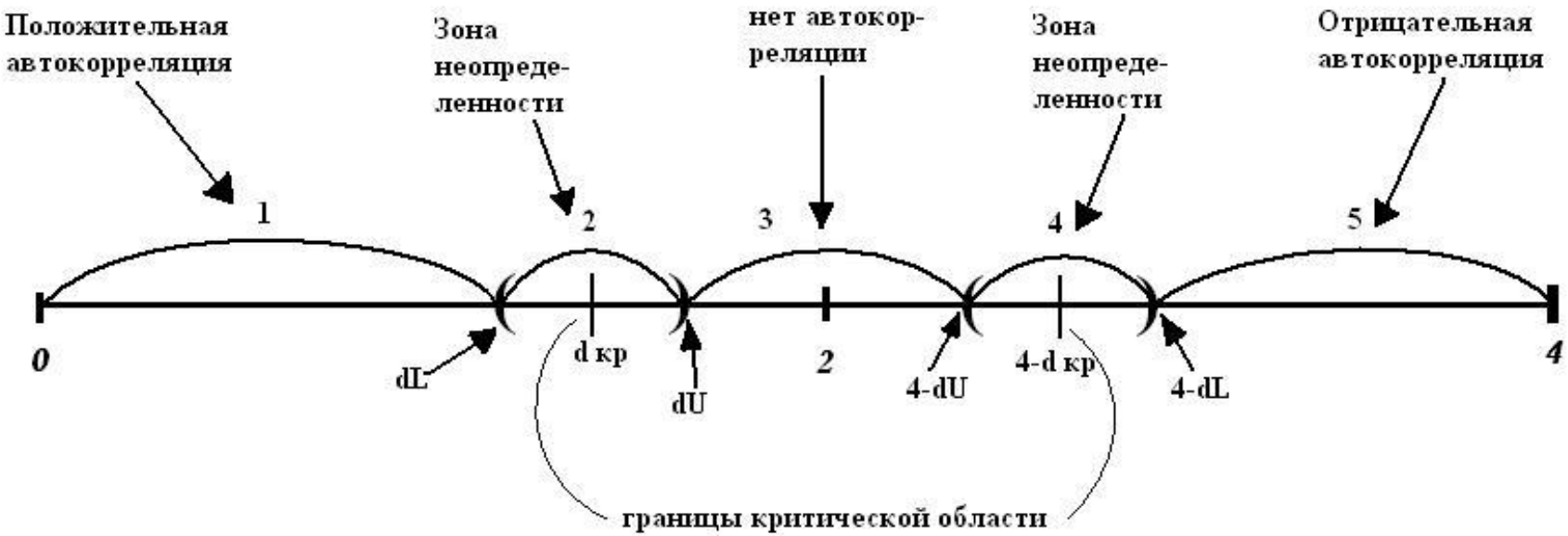
Положительная автокорреляция

Критерий Дарбина-Уотсона (тест автокорреляции остатков)

$$DW = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} = \frac{\sum_{t=2}^{T} e_t^2 + \sum_{t=2}^{T} e_{t-1}^2 - 2\sum_{t=2}^{T} e_t e_{t-1}}{\sum_{t=1}^{T} e_t^2} =$$

$$= 2 - 2\frac{\sum_{t=2}^{T} e_t e_{t-1}}{\sum_{t=1}^{T} e_t^2} \approx 2(1 - \rho_1),$$

Критерий Дарбина-Уотсона (тест автокорреляции остатков)



n	$k^1 = 1$		$k^1 = 2$		$k^1 = 3$		$k^1 = 4$		$k^1 = 5$	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
6	0,61	1,40	–	–	–	–				
7	0,70	1,36	0,47	1,90	–	–				
8	0,76	1,33	0,56	1,78	0,37	2,29				
9	0,82	1,32	0,63	1,70	0,46	2,13				
10	0,88	1,32	0,70	1,64	0,53	2,02				
11	0,93	1,32	0,66	1,60	0,60	1,93				
12	0,97	1,33	0,81	1,58	0,66	1,86				
13	1,01	1,34	0,86	1,56	0,72	1,82				
14	1.05	1.35	0.91	1.55	0.77	1.78				
16	1,10	1,37	0,98	1,54	0,86	1,73	0,74	1,93	0,62	2,15
17	1,13	1,38	1,02	1,54	0,90	1,71	0,78	1,90	0,67	2,10
18	1,16	1,39	1,05	1,53	0,93	1,69	0,82	1,87	0,71	2,06
19	1,18	1,40	1,08	1,53	0,97	1,68	0,86	1,85	0,75	2,02
20	1,20	1,41	1,10	1,54	1,00	1,68	0,90	1,83	0,79	1,99
21	1,22	1,42	1,13	1,54	1,03	1,67	0,93	1,81	0,83	1,96
22	1,24	1,43	1,15	1,54	1,05	1,66	0,96	1,80	0,86	1,94
23	1,26	1,44	1,17	1,54	1,08	1,66	0,99	1,79	0,90	1,92
24	1,27	1,45	1,19	1,55	1,10	1,66	1,01	1,78	0,93	1,90
25	1,29	1,45	1,21	1,55	1,12	1,66	1,04	1,77	0,95	1,89
26	1,30	1,46	1,22	1,55	1,14	1,65	1,06	1,76	0,98	1,88
27	1,32	1,47	1,24	1,56	1,16	1,65	1,08	1,76	1,01	1,86
28	1,33	1,48	1,26	1,56	1,18	1,65	1,10	1,75	1,03	1,85
29	1,34	1,48	1,27	1,56	1,20	1,65	1,12	1,74	1,05	1,84
30	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83

Критерий Дарбина-Уотсона (неправильный расчет)

	Среднедушевая ЗП, тыс.руб. X	et	et^2	et-1	(et-et-1)^2
1	45,1	7,524	56,614		
2	59,0	4,733	22,397	7,524	7,793
3	57,2	2,810	7,896	4,733	3,697
4	61,8	1,201	1,443	2,810	2,588
5	58,8	-1,537	2,361	1,201	7,495
6	47,2	-6,249	39,054	-1,537	22,210
7	55,2	-8,482	71,943	-6,249	4,984
Сумма	384,3	0,000	201,708	8,482	48,768

DW=48,768/201,708=0,24

Критерий Дарбина-Уотсона (верный расчет)

	Среднедушевая ЗП, тыс.руб. X	et	et^2	et-1	(et-et-1)^2
1	45,1	7,52	56,61		
2	47,2	-6,25	39,05	7,52	189,711
3	55,2	-8,48	71,94	-6,25	4,984
4	57,2	2,81	7,90	-8,48	127,506
5	58,8	-1,54	2,36	2,81	18,892
6	59,0	4,73	22,40	-1,54	39,303
7	61,8	1,20	1,44	4,73	12,471
Сумма	384,3	0,000	201,708	-1,201	392,867

DW=392,867/201,708=1,94

Задача

номер предприятия	выпуск, тыс. руб., y	Стоимость ОФ, тыс.руб. X	Yt^2	Xt^2	Xt*Yt	Yt модель	|Yt-Yt модель| /Yt	et	et^2	Yt модель-Yср	Yt-Yср	X-Xср
1	2608	430	6801664	184900	1121440	2760,55	0,06	-152,55	23270,50	-231,12	-383,67	-52,22
2	2500	410	6250000	168100	1025000	2672,03	0,07	-172,03	29595,25	-319,63	-491,67	-72,22
3	3060	530	9363600	280900	1621800	3203,12	0,05	-143,12	20482,42	211,45	68,33	47,78
4	3255	560	10595025	313600	1822800	3335,89	0,02	-80,89	6542,84	344,22	263,33	77,78
5	2893	480	8369449	230400	1388640	2981,83	0,03	-88,83	7891,08	-9,83	-98,67	-2,22
6	2941	430	8649481	184900	1264630	2760,55	0,06	180,45	32563,38	-231,12	-50,67	-52,22
7	3020	440	9120400	193600	1328800	2804,80	0,07	215,20	46309,43	-186,86	28,33	-42,22
8	3248	510	10549504	260100	1656480	3114,60	0,04	133,40	17794,81	122,94	256,33	27,78
9	3400	550	11560000	302500	1870000	3291,63	0,03	108,37	11743,87	299,96	408,33	67,78
Сумма	26925,00	4340,00	81259123,00	2119000,00	13099590,00	26925,00	0,44	0,00	196193,61	0,00	0,00	0,00
Среднее	2991,67	482,22	9028791,44	235444,44	1455510,00	2991,67	0,05	0,00	21799,29	0,00	0,00	0,00