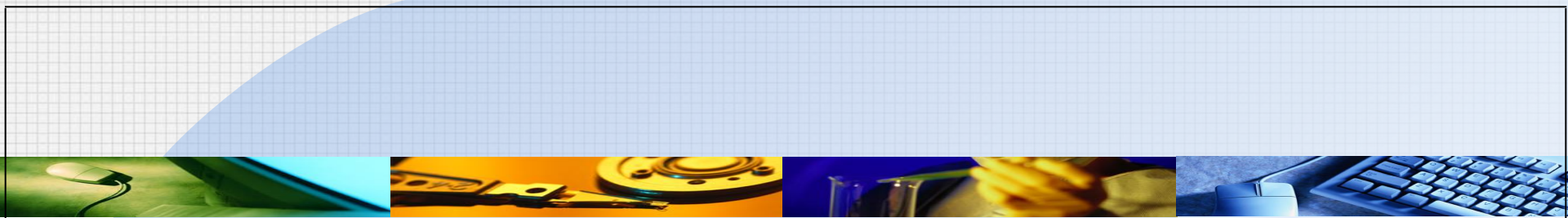


Модель простой линейной регрессии





ОСНОВНЫЕ ПОНЯТИЯ

Определение модели

- Простая линейная регрессия — это модель, описывающая зависимость величины y от одной переменной x в виде $y = a + bx + \varepsilon$
- a, b — коэффициенты
- ε — случайная величина
 - Терминология
- x — объясняющая переменная или существенный фактор или регрессор
- a, b — параметры регрессии
- ε — случайный фактор
- y — результирующий показатель или отклик

Спецификация модели

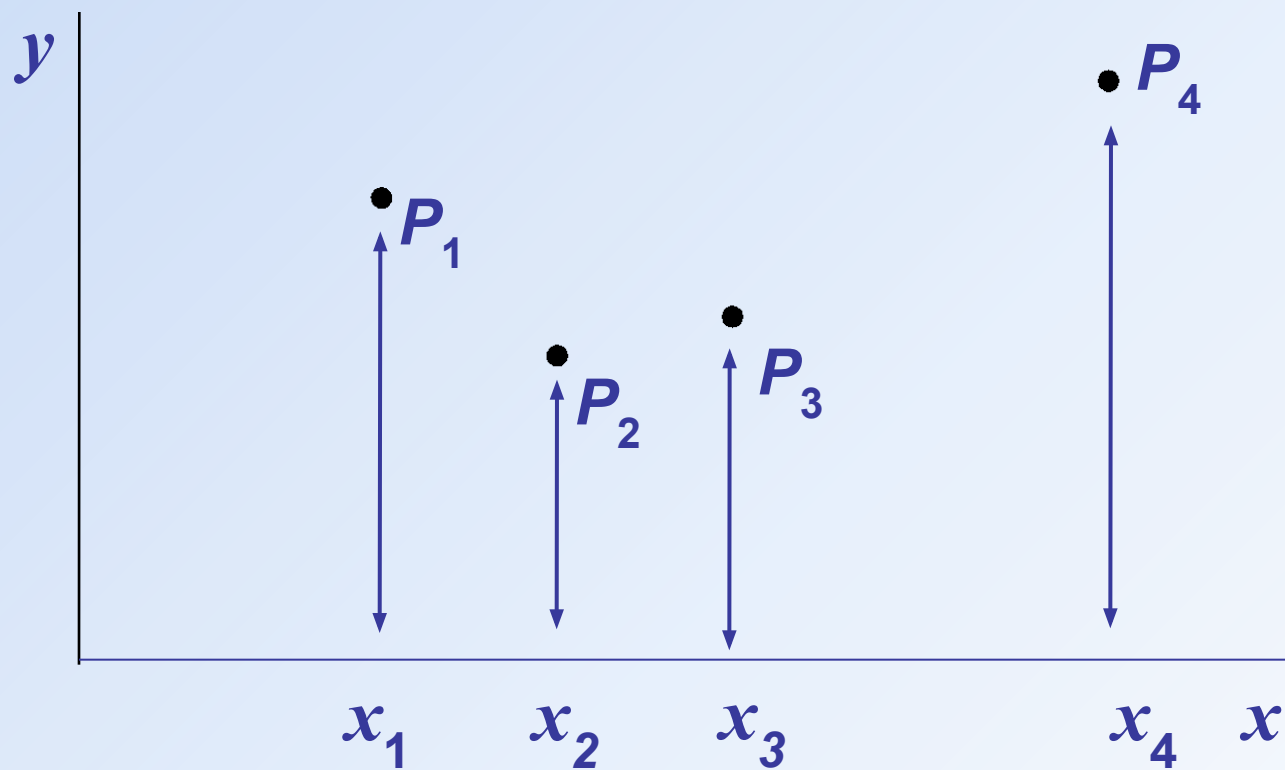
- Система уравнений
$$\begin{cases} y_1 = a + bx_1 + \varepsilon_1, \\ y_2 = a + bx_2 + \varepsilon_2, \\ \vdots \\ y_n = a + bx_n + \varepsilon_n, \end{cases}$$
- – описание моделью выборочных данных
- $(x_1; y_1), (x_2; y_2), \dots, (x_n; y_n)$
- $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ – сериальные ошибки

Теоретическое уравнение модели

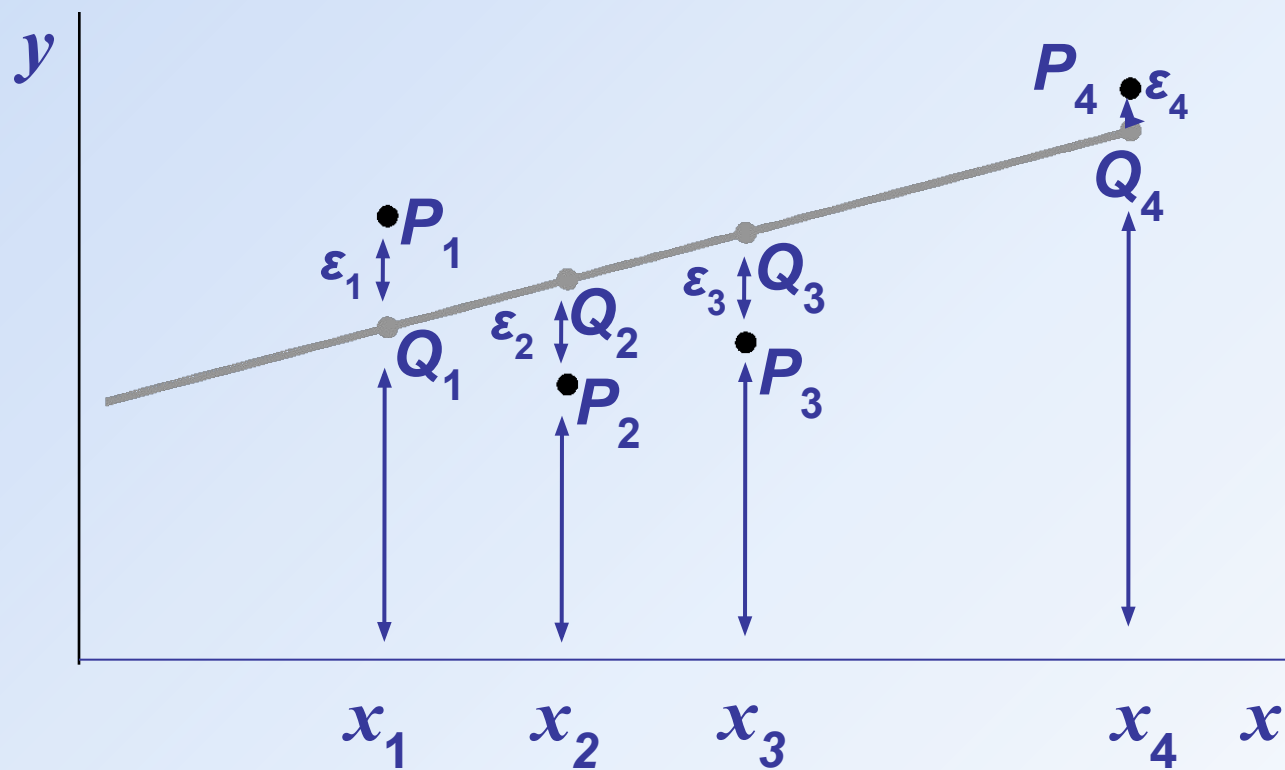


- Сериальная ошибка
- — это разность между имеющимся значением зависимой переменной и соответствующим ему значением, предсказанным по уравнению модели
- Теоретическое уравнение модели
- — такое уравнение, у которого на имеющейся выборке каждая из сериальных ошибок принимает наименьшее значение
- Обозначение $y = a + bx$

Выборка



Теоретическое уравнение



Теоретические ограничения



- У каждой сериальной ошибки математическое ожидание равно нулю
- Дисперсии всех сериальных ошибок одинаковы (*гомоскедастичность возмущений*)
- Сериальные ошибки не коррелируют между собой (*отсутствие автокорреляции возмущений*)
- Объем выборки больше двух
- Выборочные значения существенного фактора не случайны
- Элементы выборки не расположены на одной вертикальной прямой

Теоретические ограничения



- Нормальная регрессия
- *Параметрическая* или *нормальная* или *гауссовская* регрессия —
- все сериальные ошибки имеют нормальное распределение
 - Общий случай
- Сериальные ошибки — одинаково распределенные независимые случайные величины

Метод наименьших квадратов

- Задача о поиске теоретического уравнения не разрешима

- Найти a и b такие, что
$$\sum_{i=1}^n (y_i - a - bx_i)^2 \rightarrow \min$$

- Оценки $\hat{a}$ и $\hat{b}$ по методу наименьших квадратов

- Формулы для вычисления

$$b = \frac{\sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y})}{\sum_{k=1}^n (x_k - \bar{x})^2}, \quad \hat{a} = \bar{y} - \hat{b}\bar{x},$$

Эмпирическое уравнение модели



- Эмпирическое уравнение модели –
- *такое уравнение, у которого на имеющейся выборке сумма квадратов сериальных ошибок принимает наименьшее значение*
- Обозначение $y = \hat{a} + bx$

Выровненные значения и остатки



- *Выровненное значение – значение зависимой переменной, предсказанное с помощью эмпирического уравнения модели*
- *Обозначение: выровненное значение с номером i : $\hat{y}_i$*
- *Остаток – это разность между имеющимся значением зависимой переменной и соответствующим ему значением, предсказанным по эмпирическому уравнению*
- *Обозначение: остаток с номером i : e_i*
- *Вычисление: $e_i = y_i - \hat{y}_i$*

Пример

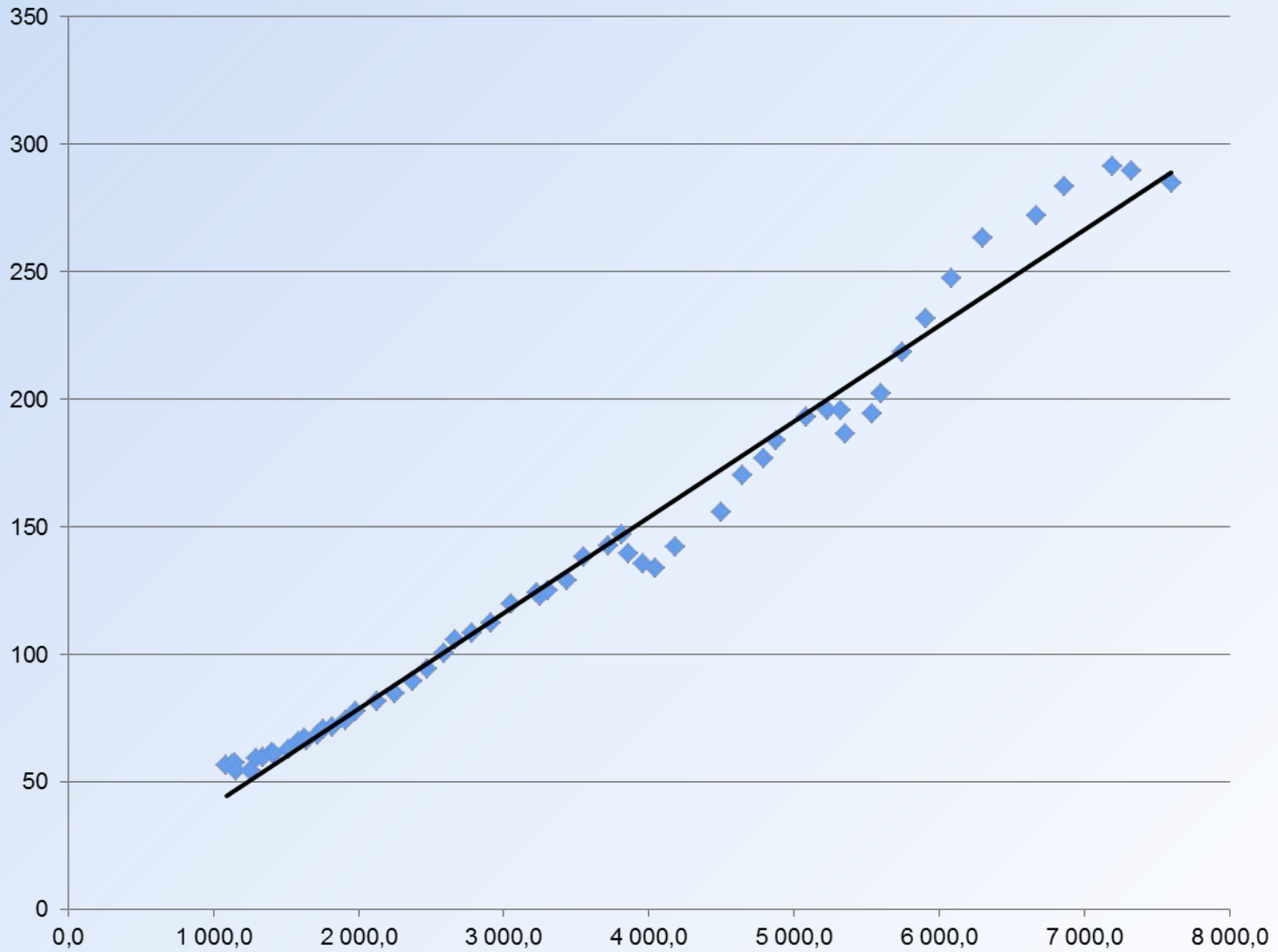


Зависимость расходов на транспорт от дохода (США, 1946-2002 годы)

- Transp – совокупные расходы на транспорт в США за год (в миллиардах долларов в ценах 2000 года)
- DPI – совокупный личный располагаемый доход в США за год (в миллиардах долларов в ценах 2000 года)

Пример

Transportation



Пример

Зависимость расходов на транспорт от дохода (США, 1946-2002 годы)

	<i>Коэффициенты</i>
Y-пересечение	3,878780296
DPI	0,037518081

- Уравнение модели $\text{Transp} = 3,8788 + 0,0375 \text{ DPI}$
- Transp – расходы на транспорт
- DPI – личный располагаемый доход

Интерпретация уравнения модели



Зависимость расходов на транспорт от дохода
(США, 1946-2002 годы)

$$\text{Transp} = 3,8788 + 0,0375 \text{DPI}$$

- *Коэффициент при DPI:*

если доход увеличивается на 1 млрд. долларов, то расходы на транспорт возрастают на 37,5 млн. долларов

- *Свободный член:*

формально показывает, что нулевом доходе расходы на транспорт будут равны 3,8788 млрд. долларов

Интерпретация уравнения модели



- *Коэффициент при объясняющей переменной:*
показывает, на сколько единиц примерно изменяется
зависимая переменная при увеличении
независимой переменной на единицу
- Свободный член равен величине зависимой*
переменной при нулевом значении существенного
фактора



ТЕОРЕМА О СУММЕ КВАДРАТОВ

Суммы квадратов

- Остатки: $e_1 = y_1 - \hat{y}_1, e_2 = y_2 - \hat{y}_2, \dots, e_n = y_n - \hat{y}_n$
- Любoй анализ качества модели – это анализ остатков
- Полная сумма квадратов (*total sum of squares*):

$$TSS = \sum_{k=1}^n (y_k - \bar{y})^2$$

- Регрессионная сумма квадратов (*regression sum of squares*):
- Сумма квадратов ошибок (*error sum of squares*)

$$ESS = \sum_{k=1}^n (y_k - \hat{y}_k)^2$$

Теорема о сумме квадратов



- *Если в модели простой регрессии выполняются все теоретические предположения, то верно равенство:*

$$TSS = RSS + ESS$$

Пример

Зависимость расходов на транспорт от дохода (США, 1946-2002 годы)

Дисперсионный анализ		
	df	SS
Регрессия	1	284507,5155
Остаток	55	5135,225939
Итого	56	289642,7414

Сумма RSS

Сумма TSS

Сумма ESS

Значимость модели



- Модель является значимой, если в теоретическом уравнении модели коэффициент при существенном факторе не равен нулю

Проверка значимости модели

■ Тест Фишера

- Основная гипотеза – модель незначимая
- Альтернативная – модель значимая
- Наблюдаемое значение:

$$F = \frac{(n - 2) \cdot RSS}{ESS}$$

- Критическое значение: квантиль уровня $1 - \alpha$ распределения Фишера с 1 и $n - 2$ степенями свободы
- Выводы: если наблюдаемое больше критического, то модель значимая (с возможной $100\alpha\%$ -й ошибкой)

если наблюдаемое меньше критического, то гипотеза о незначимости модели не отвергается

Проверка
при
заданном
уровне
значимости
 α

Пример

Зависимость расходов на транспорт от дохода
(США, 1946-2002 годы)

Наблюдаемое
значение

F	Значимость F
3047,171349	7,45928E-50

Проверка при
уровне
значимости
 $\alpha = 0,05$

Критическое
значение

4,016195493

p -значение
меньше 0,05

Модель значимая
(с возможной 5%-й ошибкой)

Коэффициент детерминации



- Коэффициент детерминации:

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

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

- Выводы о качестве модели
- Коэффициент меньше примерно 0,2:
- модель плохо описывает имеющиеся данные
- Коэффициент больше примерно 0,7: модель линейной регрессии дает хорошее описание
- Коэффициент от 0,2 до 0,7: нельзя сделать вывод о качестве модели

Пример



Зависимость расходов на транспорт от дохода (США, 1946-2002 годы)

Множественный R	0,991095597	
R-квадрат	0,982270483	
Нормированный R-квадрат	0,981948128	
Стандартная ошибка	9,662698606	
Наблюдения	57	

Модель
качественная



СТАНДАРТНЫЕ ОШИБКИ

Стандартная ошибка модели



- Стандартная ошибка модели
 - несмещенная оценка среднего квадратического отклонения сериальных ошибок
- *Формула вычисления:*

$$s = \sqrt{\frac{1}{n-2} ESS}$$

n – объем выборки

ESS – сумма квадратов сериальных ошибок

Пример

Зависимость расходов на транспорт от дохода (США, 1946-2002 годы)

<i>Регрессионная статистика</i>		
Множественный R	0,991095597	
R-квадрат	0,982270483	
Нормированный R-квадрат	0,981948128	
Стандартная ошибка	9,662698606	
Наблюдения	57	

*Стандартная
ошибка модели*

Стандартные ошибки параметров



- Стандартная ошибка параметра a
 - несмещенная оценка среднего квадратического отклонения случайной величины $\hat{a}$
- *Формула вычисления:*

$$s_a = s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{k=1}^n (x_k - \bar{x})^2}}$$

s – стандартная ошибка модели

n – объем выборки

Стандартные ошибки параметров



- Стандартная ошибка параметра b
– несмещенная оценка среднего квадратического отклонения случайной величины b

- *Формула вычисления:*

$$s_b = s \sqrt{\frac{1}{\sum_{k=1}^n (x_k - \bar{x})^2}}$$

s – стандартная ошибка модели

Пример

Зависимость расходов на транспорт от дохода (США, 1946-2002 годы)

	<i>Коэффициенты</i>	<i>Стандартная ошибка</i>
Y-пересечение	3,878780296	2,716479676
DPI	0,037518081	0,000679661

*Стандартная ошибка
свободного члена*

*Стандартная ошибка
параметра при DPI*

Интервальные оценки



- *Интервальная оценка параметра:*
- показывает с вероятностью $1 - \alpha$, в каком интервале содержится истинное значение параметра
- Вероятность $1 - \alpha$ — *надежность*
- Интервал обычно вычисляется с помощью точечной оценки параметра

Интервальные оценки

- *Интервальная оценка свободного члена:*

нижняя граница интервала $\hat{a} - t_{1-\alpha, n-2} \cdot s_a$

верхняя граница интервала $\hat{a} + t_{1-\alpha, n-2} \cdot s_a$

$\hat{a}$ — точечная оценка свободного члена

s_a — стандартная ошибка свободного члена

$t_{1-\alpha, n-2}$ — двусторонняя квантиль уровня $1 - \alpha$
распределения Стьюдента с $n - 2$ степенями
свободы

Интервальные оценки

- Интервальная оценка углового коэффициента:

нижняя граница интервала $\hat{b} - t_{1-\alpha, n-2} \cdot s_b$

верхняя граница интервала $\hat{b} + t_{1-\alpha, n-2} \cdot s_b$

$\hat{b}$ – точечная оценка углового коэффициента

s_b – стандартная ошибка углового коэффициента

$t_{1-\alpha, n-2}$ – двусторонняя квантиль уровня $1 - \alpha$
распределения Стьюдента с $n - 2$ степенями
свободы

Пример

Зависимость расходов на транспорт от дохода (США, 1946-2002 годы)

*Интервальная оценка
свободного члена*

	<i>Коэффициенты</i>	<i>Нижние 95%</i>	<i>Верхние 95%</i>
Y-пересечение	3,878780296	-1,565166628	9,32272722
DPI	0,037518081	0,03615601	0,038880151

*Интервальная оценка
параметра DPI*



ЗНАЧИМОСТЬ ПАРАМЕТРОВ МОДЕЛИ

Определения



- Параметр при существенном факторе x называется значимым, если его истинное значение не равно нулю
- Значимость параметра при x означает: модель учитывает влияние данного фактора на зависимую переменную
- Параметр при существенном факторе x называется статистически незначимым, если его значимость не установлена
- Статистическая незначимость параметра при x означает: возможно, модель не учитывает влияние данного фактора на зависимую переменную

Значимость модели и параметров



- В модели простой линейной регрессии значимость параметра при существенном факторе равносильна значимости модели!

Проверка значимости параметра

■ Тест Стьюдента

- Основная гипотеза – параметр b незначимый

- Альтернативная – параметр b значимый

- Наблюдаемое значение:
$$\hat{t} = \frac{\hat{b}}{s_b}$$

- Критическое значение: квантиль уровня $1 - \alpha$ распределения Стьюдента с $n - 2$ степенями свободы

- Выводы: если наблюдаемое больше критического, то параметр значимый (с возможной $100\alpha\%$ -й ошибкой)
если наблюдаемое меньше критического, то гипотеза о незначимости параметра не отвергается (статистическая незначимость параметра)

Проверка
при
заданном
уровне
значимости
 α

Пример

Зависимость расходов на транспорт от дохода (США, 1946-2002 годы)

Проверка при
уровне значимости
 $\alpha = 0,05$

	Коэффициенты	t-статистика	P-Значение
Y-пересечение	3,878780296	1,42787017	0,158983049
DPI	0,037518081	55,20118974	7,45928E-50





ПРОГНОЗИРОВАНИЕ

Виды прогнозирования



- *Безусловное прогнозирование (предсказание):*
значение существенного фактора, соответствующее прогнозируемому значению, *известно*
- *Условное прогнозирование:*
значение существенного фактора, соответствующее прогнозируемому значению, *не известно*

Точечный прогноз

- *Точечный прогноз:*

значение зависимой переменной, вычисленное с помощью эмпирического уравнения модели

- *Вычисление:* $\hat{y}_0 = \hat{a} + bx_0$

x_0 — значение соответствующего существенного фактора

Стандартная ошибка



- *Стандартная ошибка точечного прогноза:*
несмещенная оценка стандартного отклонения
случайной величины $(\hat{a} - a) + (\hat{b} - b)x_0$

- *Вычисление:*

$$s_0 = s \cdot \sqrt{\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{k=1}^n (x_k - \bar{x})^2} \right)}$$

s – стандартная ошибка точечного прогноза

x_0 – значение соответствующего существенного фактора

Интервальный прогноз



- *Интервальная прогноз:*
- показывает с вероятностью $1 - \alpha$, в каком интервале содержится истинное значение зависимой переменной
- Вероятность $1 - \alpha$ — *надежность*

Интервальный прогноз

- *Вычисление:*

нижняя граница интервала $\hat{y}_0 - t_{1-\alpha, n-2} \cdot s_0$

верхняя граница интервала $\hat{y}_0 + t_{1-\alpha, n-2} \cdot s_0$

$\hat{y}_0$ – точечный прогноз

s_0 – стандартная ошибка прогноза

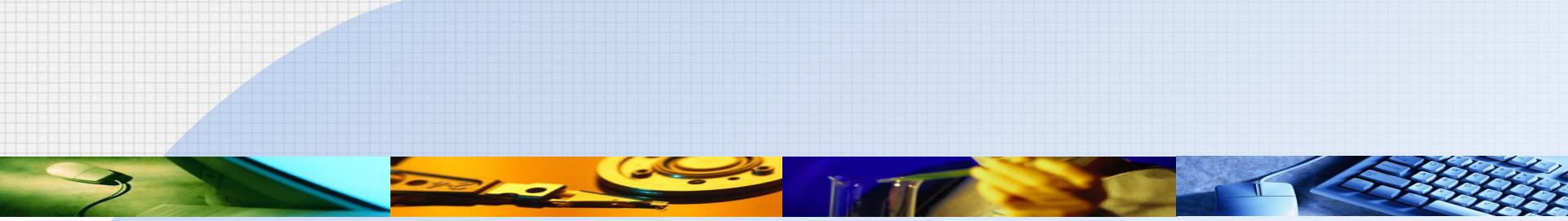
$t_{1-\alpha, n-2}$ – двусторонняя квантиль уровня $1 - \alpha$
распределения Стьюдента с $n - 2$ степенями
свободы

Пример

Зависимость расходов на транспорт от дохода (США, 1946-2002 годы)

x_0 (2003 год)	7787,4
Точечный прогноз y_0	296,0471
Стандартная ошибка прогноза	10,16842

Интервальный прогноз	
Нижняя 95% граница	Нижняя 95% граница
275,669122	316,4250403



НЕЛИНЕЙНАЯ РЕГРЕССИЯ

Нелинейные модели

- Два вида регрессий:

нелинейные относительно объясняющих переменных, но линейные по оцениваемым параметрам

Все после замены становятся линейными

$$y = a + b \ln x$$

$$y = a + b / x$$

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

нелинейные по оцениваемым параметрам

$$y = e^{a+bx}$$

$$y = ax^b$$

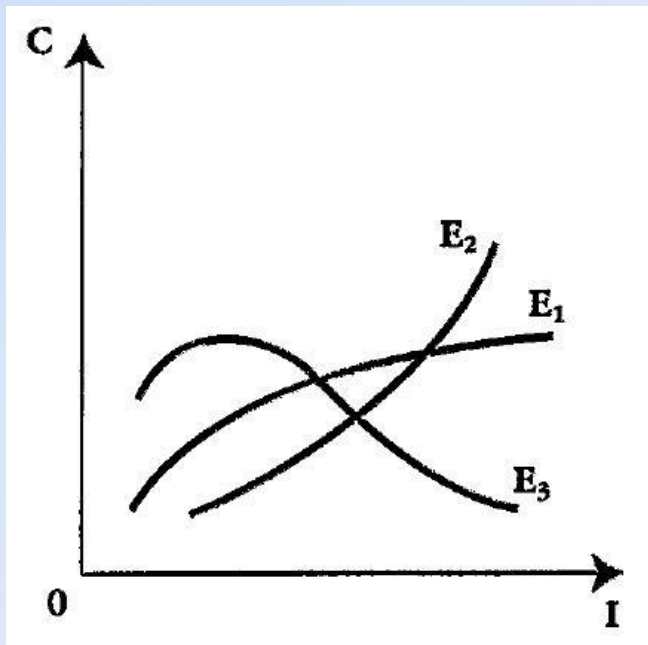
Некоторые сводятся к линейным после логарифмирования

Пример

■ Кривые Энгеля

Эрнст Энгель (1821-1896)
немецкий экономист и статистик

показывает зависимость между объёмом потребления товаров или услуг и доходом потребителя при неизменных ценах и



E₁ — кривая для нормальных товаров

E₂ — кривая для предметов роскоши

E₃ — кривая для низкокачественных товаров

Основные нелинейные модели

- Гиперболическая

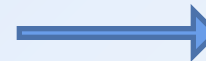
$$y = a + b / x$$

- Параболическая

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

- Экспоненциальная

$$y = e^{a+bx}$$



$$\ln y = a + bx$$

- Степенная

$$y = ax^b$$



$$\ln y = \ln a + b \ln x$$

После
замены
становятся
линейными

Полулогарифмическая
регрессия

Логарифмическая
регрессия



ВЫБОР ЛУЧШЕЙ МОДЕЛИ

Оценка качества модели



Инструменты

- Точечная диаграмма (*расположение точек вдоль линии тренда*)
- Статистика Фишера (*значимость модели по тесту Фишера*)
- Коэффициент детерминации (*оценка качества модели по его величине*)
- Средняя относительная погрешность (*оценка качества модели по её величине*)

$$\delta_{\text{ср}} = \frac{100\%}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{y_i}$$

Оценка качества модели



Характеристики подходящей модели

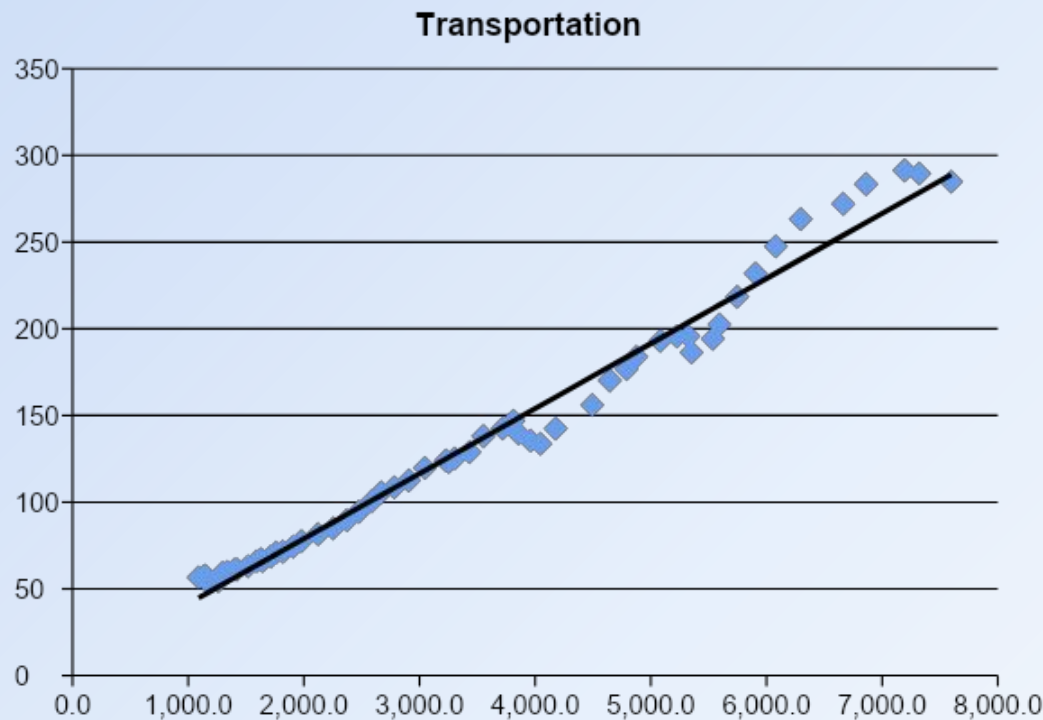
- На диаграмме точки расположены, в основном, вдоль линии тренда
- Модель значимая
- Коэффициент детерминации не меньше заданного уровня (обычно 0,65-0,7)
- Средняя относительная погрешность не меньше заданного уровня (обычно 10% - 25%)

Могут использоваться модели с меньшим коэффициентом

Могут использоваться модели с большей погрешностью

Пример

Зависимость расходов на транспорт от дохода (США, 1946-2002 годы)



Точки
расположены
вдоль
линейного
тренда

Пример

Зависимость расходов на транспорт от дохода (США, 1946-2002 годы)

Модель
значимая

- Статистика Фишера

F	Значимость F
3047,171349	7,45928E-50

- Коэффициент детерминации

R-квадрат	0,982270483
-----------	-------------

- Средняя относительная погрешность

Средняя относительная погрешность	5,26%
---	-------

Модель подходящая

Выбор модели



Два этапа

- *Первый этап: выбор подходящих моделей*

Обычно используются: линейная, гиперболическая, параболическая, экспоненциальная, степенная модели

Для моделей с зависимой переменной, отличной от исходной, предсказанные значения, остатки, коэффициенты детерминации и среднюю относительную погрешность необходимо вычислять отдельно!

Выбор модели



Два этапа

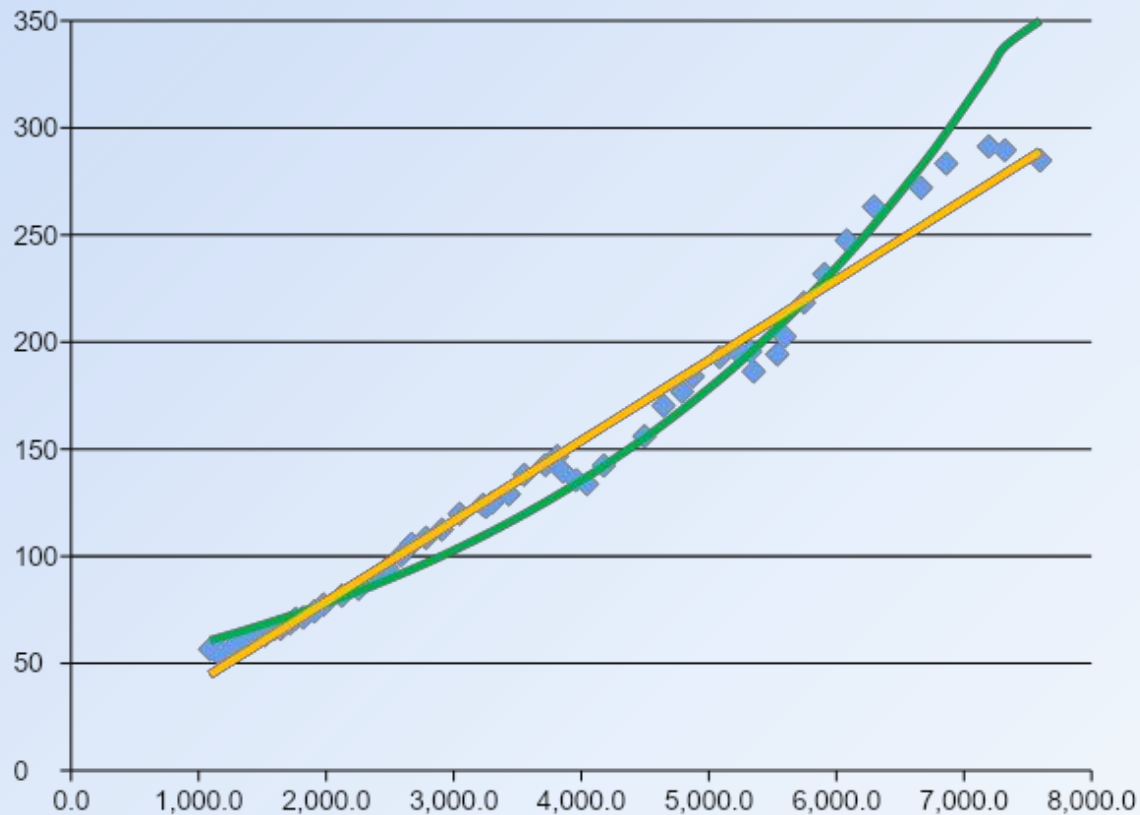
- *Второй этап: выбор лучшей модели*

Для сравнения подходящих моделей используются такие же инструменты, как на первом этапе

Пример

Зависимость расходов на транспорт от дохода (США, 1946-2002 годы)

Transportation



Все модели
подходящие

Пример

Зависимость расходов на транспорт от дохода (США, 1946-2002 годы)

	Основные характеристики			
	Линейная	Параболическая	Экспоненциальная	Степенная
Коэффициент детерминации	0,9823	0,9897	0,9736	0,9824
Средняя относительная погрешность	5,26%	3,52%	7,01%	5,32%
Значимость параметра при DPI	да	да (при DPI и DPI^2)	да	да

Лучшая модель – линейная