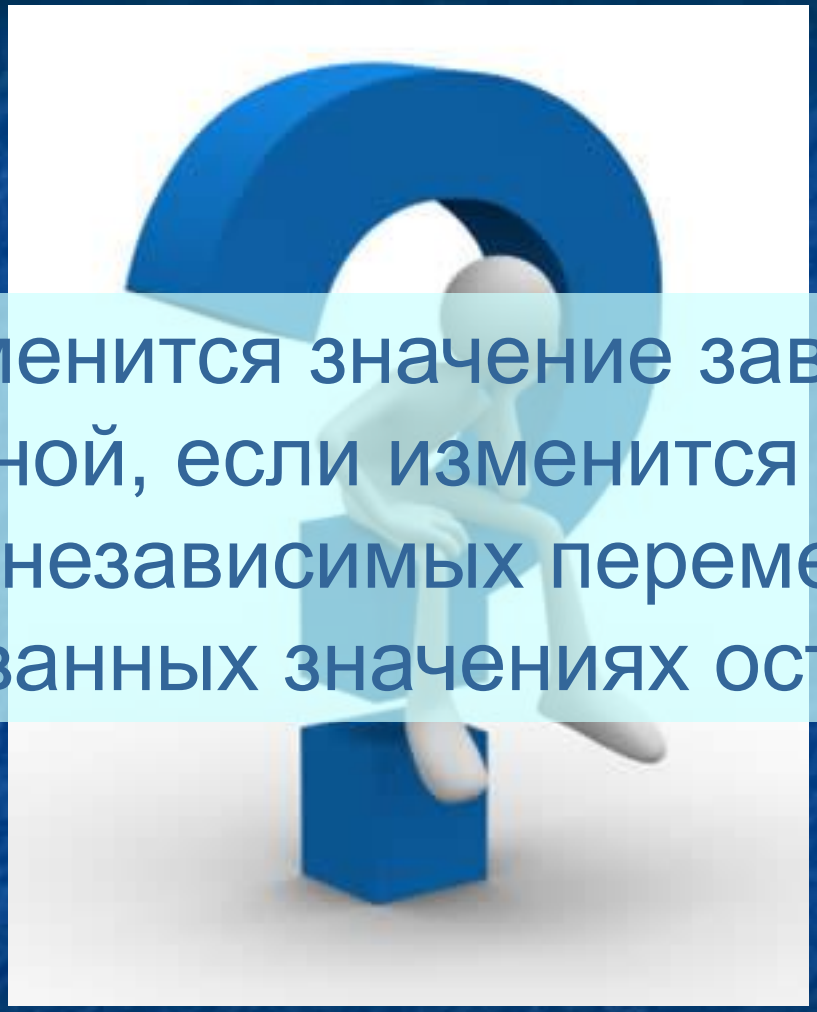
Парная (простая) регрессия в эконометрических расчетах

РЕГРЕССИЯ

Термин *регрессия* (движение назад, возвращение в прежнее состояние) был введен Фрэнсисом Галтоном в конце XIX века при анализе зависимости между ростом родителей и ростом детей. Галтон заметил, что рост детей у очень высоких родителей в среднем меньше, чем средний рост родителей. У очень низких родителей, наоборот, средний рост детей выше. И в том и в другом случае средний рост детей стремится (возвращается) к среднему росту людей в данном регионе. Отсюда и выбор термина, отражающего такую зависимость.

Регрессионный анализ — это процесс определения аналитического выражения функции связи, в котором изменение результативной или зависимой переменной происходит под влиянием факторной, или независимой, переменной.

… техника анализа связи между зависимой переменной и одной или несколькими независимыми переменными.



Как изменится значение зависимой переменной, если изменится значение одной из независимых переменных при фиксированных значениях остальных ?

$$g_t = E[y_t | x_{1,t}, \ldots x_{n,t}] = g(x_{1,t}, \ldots x_{n,t})$$

$$f(y_t | x_{1,t}, \ldots x_{n,t}) = f(y_t - \mu | x_{1,t}, \ldots x_{n,t})$$

$$\mu = \mu(x_{1,t}, \ldots x_{n,t})$$

Зависимая переменная.

Случайная составляющая.

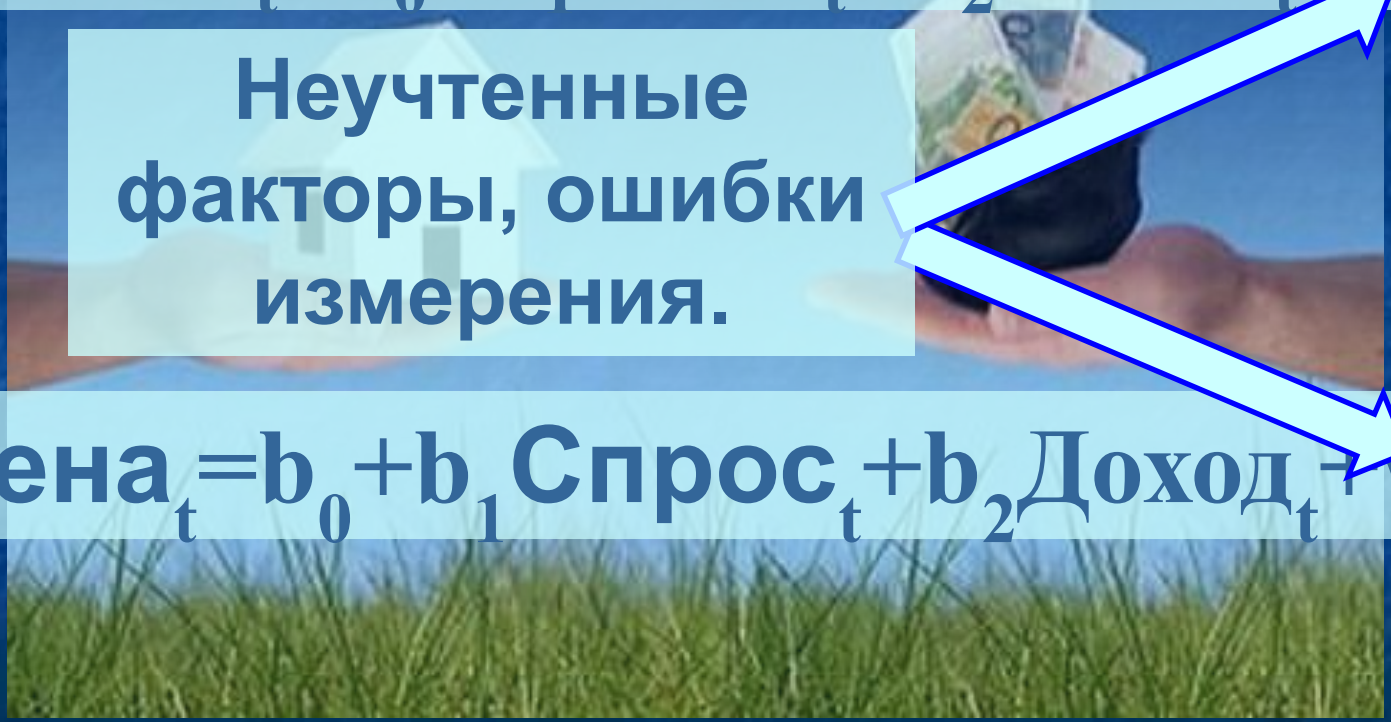
$$y_t = a_1 x_{1,t} \cdots a_n x_{n,t} + v_t$$

Независимые (объясняющие) переменные, регрессоры.

Прямая и обратная функции спроса

$$Спрос_t = a_0 + a_1 Цена_t + a_2 Доход_t + v_t$$

Неучтенные факторы, ошибки измерения.



$$Цена_t = b_0 + b_1 Спрос_t + b_2 Доход_t + w_t$$

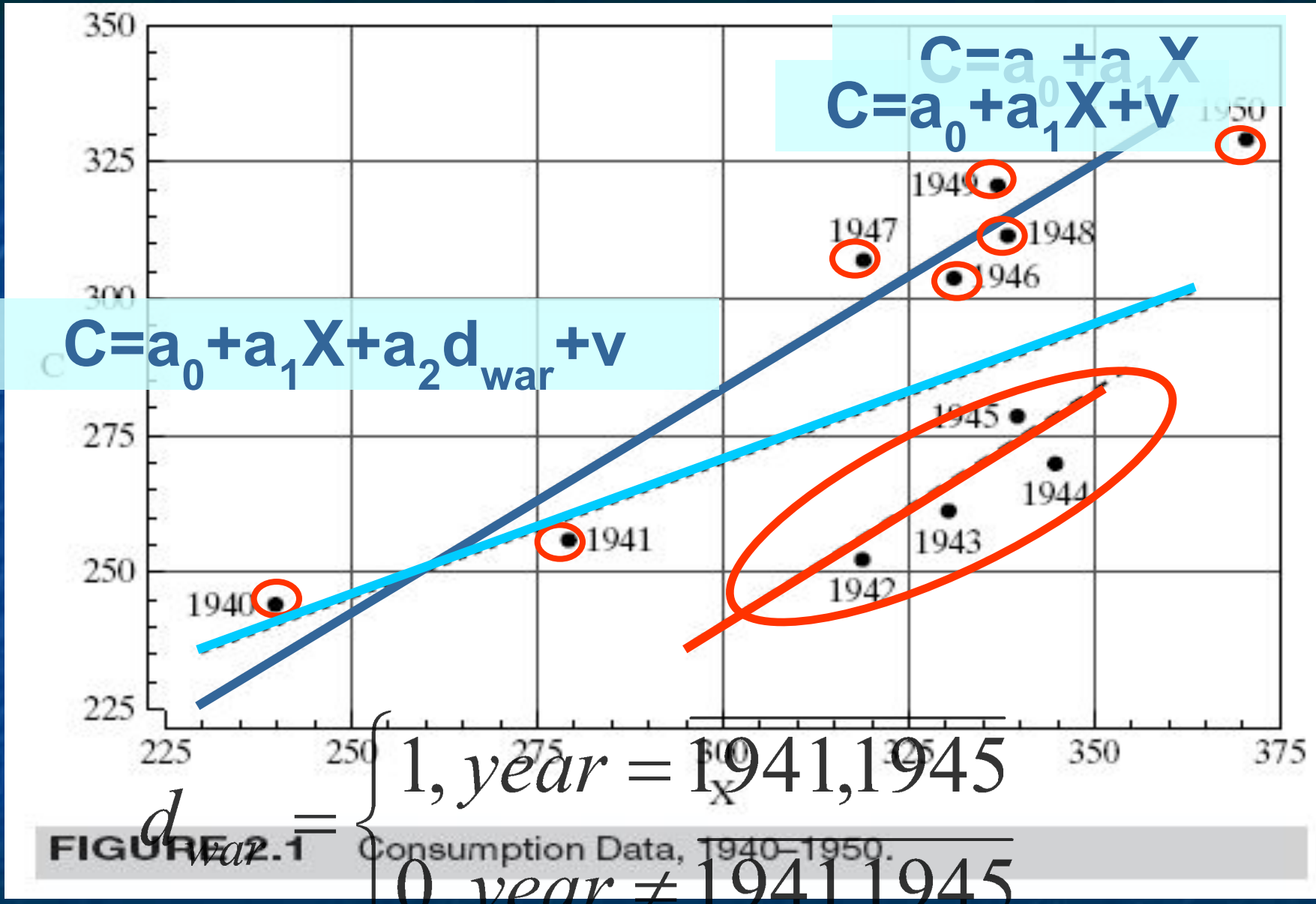


FIGURE 2.1 Consumption Data, 1940–1950.

Линейность регрессионной модели

$$Y=Xa+v,$$
$$Y,v\in R^T,$$
$$X\in M_{T,n},\ a\in R^n$$

$$X=[1,x_1,\ \ldots,\ x_{n-1}]\ \Rightarrow$$

$$y_t=a_0+a_1x_{1,t}+\ldots+a_{n-1}x_{n-1,t}+v_t$$

Является ьёзным

НЕТ !

«Линейность» - относится к способу вхождения параметров и случайной составляющей в модель.

$$y_t = a_0 + a_1 \cos(x_t) + v_t$$

Линейная модель

Нелинейная модель

$$y_t = a_0 + a_0 a_1 \cos(x_t) + v_t$$

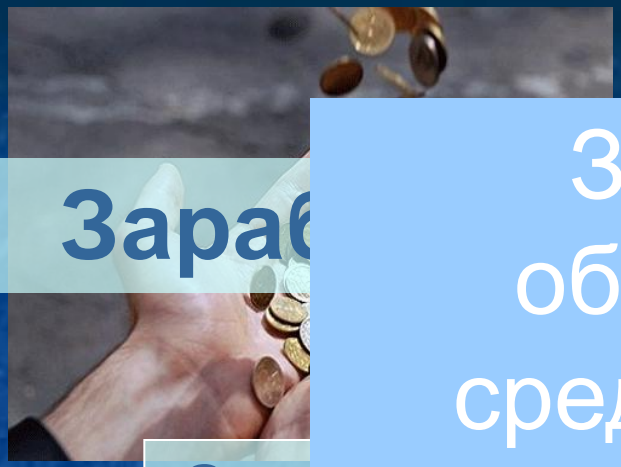
$$y=Ax^{a}e^{v}$$

Линейная модель

$$\ln(y)=\ln(A)+a\ln(x)+v$$

Нелинейная модель

$$y=Ax^{a}+v$$



Зараб

ие$_t$+v$_t$

Зав ие

образования.

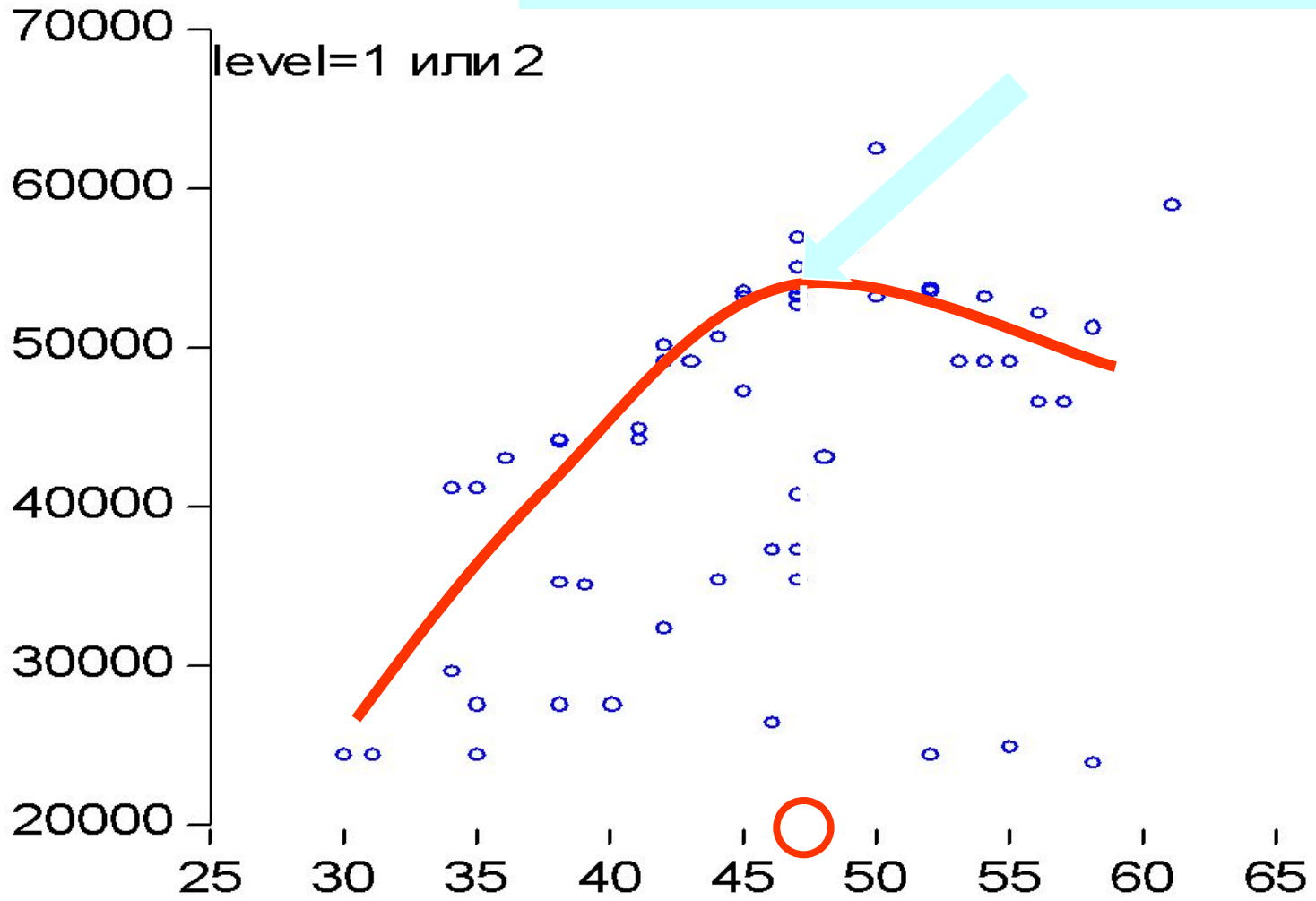
Заработок и образование в среднем растут с возрастом.

$$Заработок_t = a_0 + a_1 Образование_t + + a_2 Возраст_t + v_t$$

Снижение темпа роста доходов

$$Заработок_t = a_0 + a_1 Образование_t + + a_2 Возраст_t + a_3 (Возраст_t)^2 + v_t$$

«Пик карьеры»


70000
60000
50000
40000
30000
20000
level=1 или 2
25
30
35
40
45
50
55
60
65

ЭЛАСТИЧНОСТЬ ФУНКЦИИ

[function elasticity] — предел

На сколько процентов измениться ‘y’, если ‘x’ измениться на 1 % ?

относительному приращению независимой переменной x Δx/x когда Δx и Δy→ 0. «Экономико-математический словарь»

$$el_x(y) \approx (\Delta y/y)/(\Delta x/x)$$

$$el_x(y) \approx [d(\ln(y))/[d(\ln(x))]$$

$$el_{x1}(y)$$

$$\ln(y_t)=a_0+a_1\ln(x_{1,t})+\ldots$$
$$\ldots+a_{n-1}\ln(x_{n-1,t})+v_t$$

Логолинейная модель

СПЕЦИФИКАЦИЯ МОДЕЛИ

- Любое исследование в эконометрике начинается с формулировки вида модели, исходя из установленной связи между переменными
- Если с помощью коэффициентов парной корреляции установлена значимая устойчивая связь между переменными, то её можно использовать для построения модели парной регрессии

Основная модель регрессии — это модель парной, или однофакторной, регрессии, которая называется полиномом первой степени. Модель парной регрессии применяется для характеристики процессов, равномерно развивающихся во времени.

Парная регрессия представляет собой модель, где среднее значение зависимой переменной у рассматривается как функция одной независимой переменной

Общий вид модели парной регрессии зависимости переменной y от переменной x:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

где y_i — результативные переменные, $i = 1, 2, \ldots n$;
x_i — факторные переменные;
β_0, β_1 — неизвестные параметры модели парной регрессии;
ε_i — случайная ошибка регрессионной модели.

$$\hat{y} = f(x)$$

СФЕРА ПРИМЕНЕНИЯ МОДЕЛИ

- Парная регрессия достаточна, если имеется ярко выраженный доминирующий фактор, который и используется в качестве независимой переменной, поскольку остальные факторы считаются неизменными

ПРАВИЛЬНОСТЬ ПРИМЕНЕНИЯ

- правильность применения корреляционного и регрессионного анализа при изучении взаимосвязей переменных подтверждается наличием нормального распределения совокупности, по изучаемым переменным, то есть её однородности

ПРАВИЛЬНОСТЬ ПРИМЕНЕНИЯ

- Подтверждается попаданием теоретических значений y(x) в пределы между минимальным и максимальным значением результативного признака y

ОШИБКА ПОСТРОЕНИЯ МОДЕЛИ (ε)

ВИД составляющей ОШИБКИ	СПОСОБ МИНИМИЗАЦИИ ОШИБКИ
Ошибки спецификации	Изменение формы модели (вида уравнения)
Ошибки выборки	Увеличение объема исходных данных
Ошибки измерения	Рост качества (достоверности) данных

Для спецификации модели используются

- Линейные функции, например,

$$f(x) = b_0 + b_1 x$$

- Нелинейные функции, например,

$$f(x) = b_0 x^{b1}$$

- Нелинейные функции можно преобразовать, прологарифмировать значения переменных и работать дальше с линейными функциями

ВЫБОР ВИДА ФУНКЦИИ

Осуществляется

- Графическим методом (метод визуальной оценки)
- Аналитическим методом
- Экспериментальным методом

Графический метод

обычно осуществляется по графическому изображению реальных статистических данных в виде точек в декартовой системе координат, которое называется *корреляционным полем (диаграммой рассеивания)* (рис. 4.1).

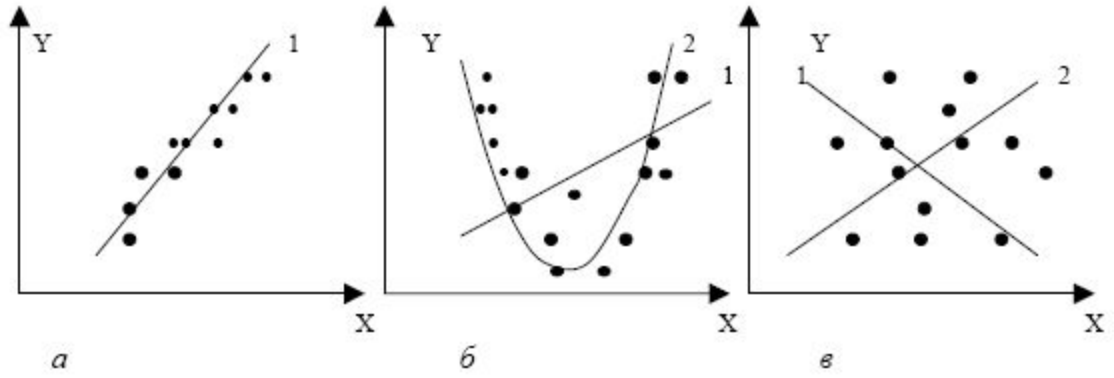


Рис. 4.1

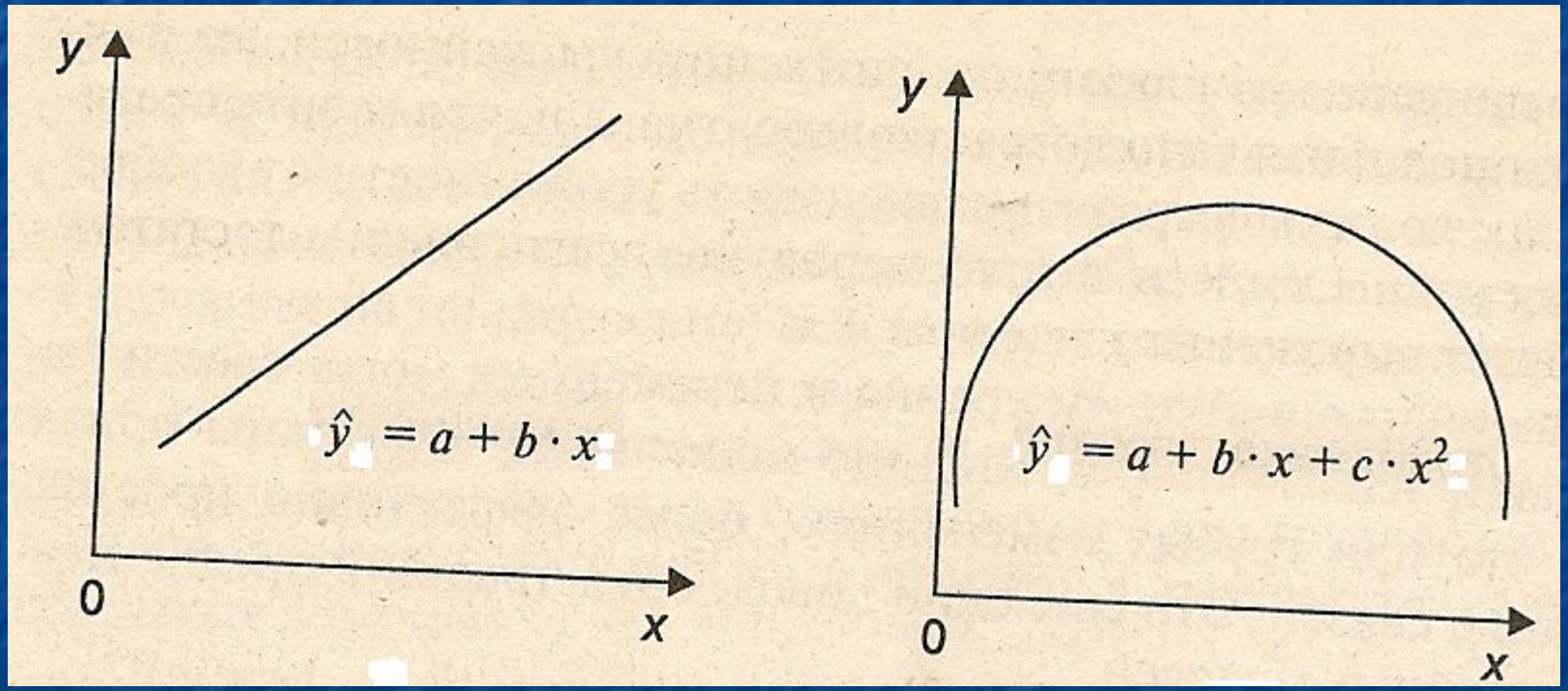
На рис 4.1 представлены три ситуации.

На графике 4.1, *а* взаимосвязь между X и Y близка к линейной, и прямая 1 достаточно хорошо соответствует эмпирическим точкам. Поэтому в данном случае в качестве зависимости между X и Y целесообразно выбрать линейную функцию $Y = b_0 + b_1 X$.

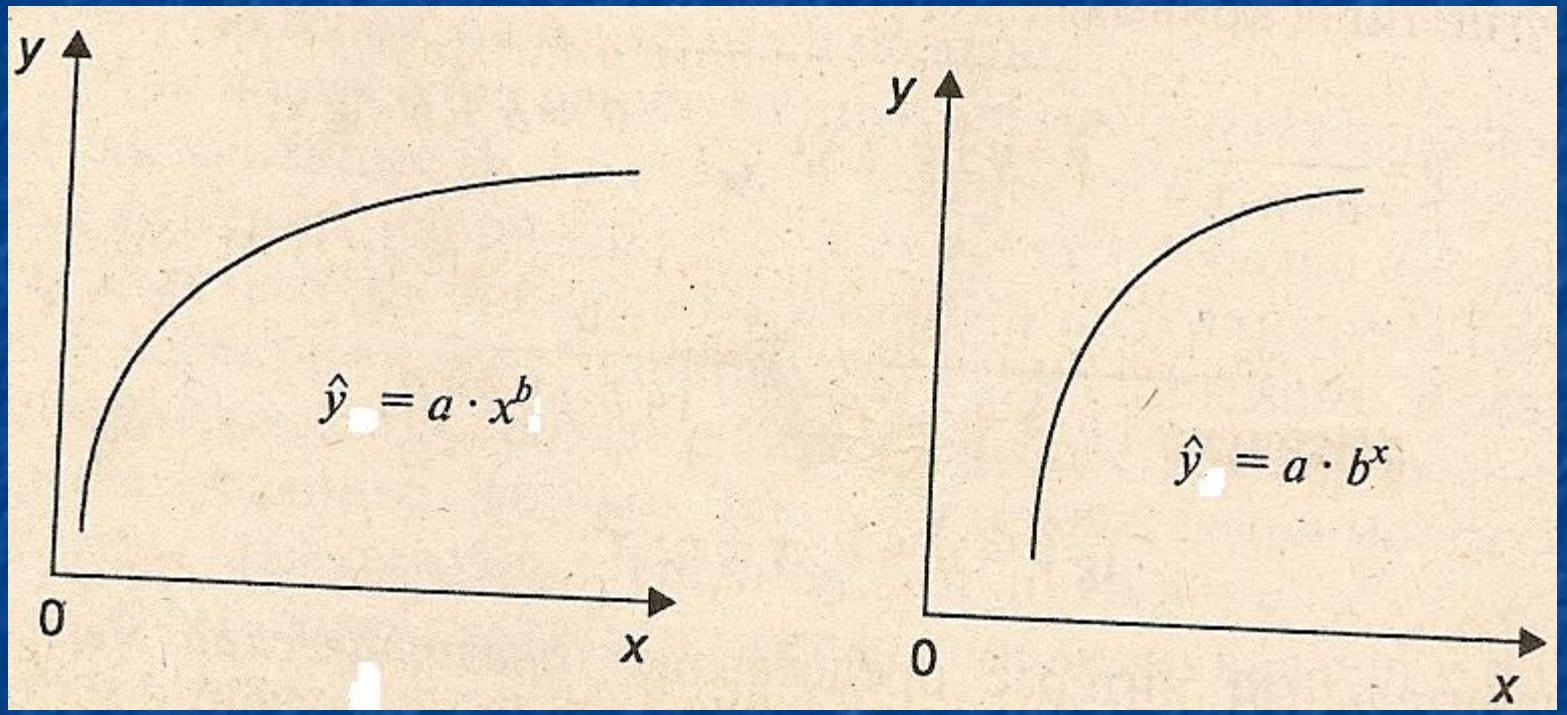
На графике 4.1, *б* реальная взаимосвязь между X и Y, скорее всего, описывается квадратичной функцией $Y = aX^2 + bX + c$ (линия 2), и какую бы мы не провели прямую (например, линия 1), отклонения точек наблюдений от нее будут существенными и неслучайными.

На графике 4.1, *в* явная взаимосвязь между X и Y отсутствует.

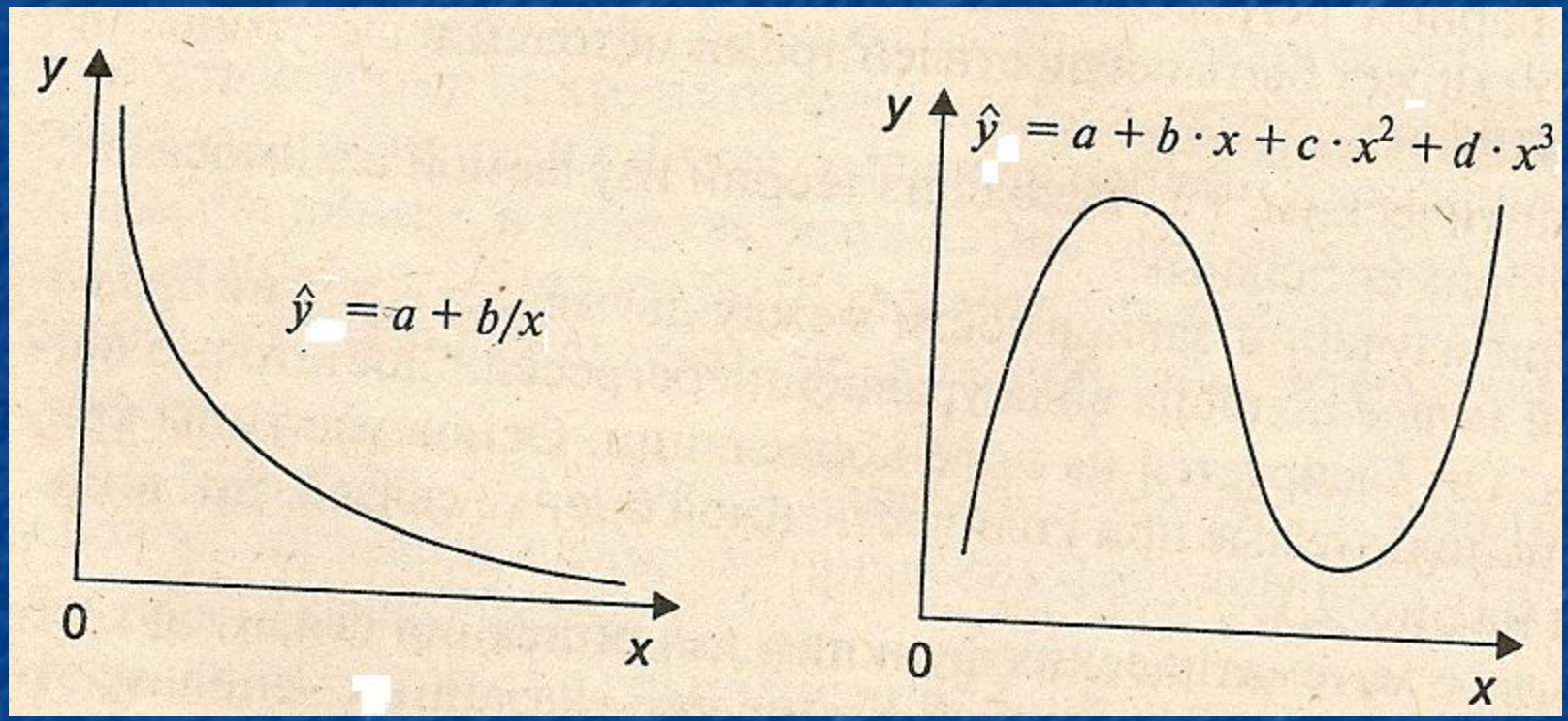
ОСНОВНЫЕ ТИПЫ КРИВЫХ



ОСНОВНЫЕ ТИПЫ КРИВЫХ



ОСНОВНЫЕ ТИПЫ КРИВЫХ



ОСНОВНЫЕ ТИПЫ КРИВЫХ

$$\hat{y} = \frac{1}{a + b \cdot x}; \qquad \hat{y} = a + b \cdot x + c \cdot \frac{1}{x}; \qquad \hat{y} = a + b \cdot \lg x;$$

$$\hat{y} = \frac{1}{a + b \cdot x + c \cdot x^2}; \qquad \hat{y} = \frac{a}{1 + b \cdot e^{-c \cdot x}};$$

$$\lg \hat{y} = a + b \cdot x + c \cdot x^2.$$

Аналитический метод

- Основан на изучении качественной природы связи исследуемых признаков

То есть, форма связи известна, например, зависимость величины налога, от уровня налоговой ставки

Экспериментальный метод

- Используется при применении компьютерных статистических прикладных пакетов
- Основывается на сравнении величины остаточной дисперсии, рассчитанной для разных типов кривых, и выборе кривой, где её величина минимальна

ПРАКТИКА ПОКАЗЫВАЕТ

- Число наблюдений должно в 6-7 раз превышать число рассчитываемых параметров при переменной x.
- Усложнение типа кривой требует увеличение числа наблюдений.
- Искать линейную регрессию, имея менее 7 наблюдений не имеет смысла.

ОЦЕНИВАНИЕ НЕИЗВЕСТНЫХ КОЭФФИЦИЕНТОВ МОДЕЛИ ПАРНОЙ РЕГРЕССИИ

В модели парной регрессии результативной переменной y от факторной переменной x неизвестными являются коэффициенты $\beta_0 \ldots \beta_n$. Существуют определенные **методы оценки** неизвестных коэффициентов модели парной регрессии.

- Метод наименьших квадратов (МНК)
- Метод наименьших разностей
- Метод функционала

МНК

1. **Метод наименьших квадратов** (МНК), при котором рассчитывается сумма квадратов отклонений наблюдаемых значений результативной переменной y от теоретических значений $\tilde{y}$ (рассчитанных на основании функции регрессии $f(x)$):

$$F = \sum_{i=1}^{n} (y_i - f(x_i, \beta))^2 \text{ или } F = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2.$$

Для определения оптимальных значений неизвестных коэффициентов $\beta_0 \ldots \beta_n$ функционал F минимизируется по данным параметрам:

$$F = \sum_{i=1}^{n} (y_i - f(x_i, \beta))^2 \to \min,$$

т.е. рассчитываются такие коэффициенты $\beta_0 \ldots \beta_n$, при которых сумма квадратов отклонений наблюдаемых значений результативной переменной y от теоретических значений $\tilde{y}$ была бы минимальной.

Достоинства МНК — сведение всех вычислительных процедур к простому вычислению неизвестных коэффициентов; доступность математических выводов.

Недостаток МНК — чувствительность оценок к резким выбросам, встречающимся в исходных данных. МНК является наиболее распрост-

раненным методом оценки неизвестных коэффициентов модели парной регрессии.

2. **Метод**, при котором рассчитывается сумма модулей отклонений наблюдаемых значений результативной переменной y от теоретических значений :

$$F = \sum_{i=1}^{n} \left| y_i - f(x_i, \beta) \right| \text{ или } F = \sum_{i=1}^{n} \left| y_i - \tilde{y}_i \right|.$$

Для определения оптимальных значений неизвестных коэффициентов $\beta_0 \ldots \beta_n$ функционал F минимизируется по данным параметрам:

$$F = \sum_{i=1}^{n} \left| y_i - f(x_i, \beta) \right| \to \min,$$

т.е. рассчитываются такие коэффициенты $\beta_0 \ldots \beta_n$, при которых сумма модулей отклонений наблюдаемых значений результативной переменной y от теоретических значений $\tilde{y}$ была бы минимальной.

Достоинство данного метода — нечувствительность оценок к резким выбросам.

Недостатки данного метода:

1) сложность вычислительной процедуры;
2) возможность соответствия различным значениям оцениваемых коэффициентов $\beta_0 \ldots \beta_n$ одинаковых сумм модулей отклонений.

3. **Метод**, при котором рассчитывается функционал вида:

$$F = \sum_{i=1}^{n} g(y_i - f(x_i, \beta)) \text{ или } F = \sum_{i=1}^{n} g(y_i - \tilde{y}_i),$$

где g — мера или вес, с которой отклонение $(y_i - f(x_i, \beta))$ входит в данный функционал.

КЛАССИЧЕСКИЙ МЕТОД НАИМЕНЬШИХ КВАДРАТОВ ДЛЯ МОДЕЛИ ПАРНОЙ РЕГРЕССИИ

Предположим, что между результативной переменной x и факторной переменной y существует линейная связь, которая описывается равенством:

$$y_i = \beta_0 + \beta_1 x_i. \quad (1)$$

Суть метода наименьших квадратов состоит в том, что нужно рассчитать такие значения коэффициентов $\tilde{\beta}_0$ и $\tilde{\beta}_1$, которые минимизировали бы сумму квадратов отклонений наблюдаемых значений результативной переменной y от теоретических значений $\tilde{y}$, т.е. доставляли минимум функции (1):

$$F = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 = \sum_{i=1}^{n} (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i) \to \min. \quad (2)$$

Значения результативной *(y)* и факторной *(x)* переменных известны из наблюдений. Следовательно, при минимизации функции (1) неизвестными являются только значения коэффициентов модели регрессии β_0 и β_1.

Для определения минимума функции двух переменных рассчитываются частные производные этой функции по каждому из оцениваемых параметров и приравниваются к нулю.

Полученная система уравнений называется стационарной системой уравнений для функции (1).

В результате преобразования стационарной системы уравнений получим систему двух нормальных линейных уравнений:

$$\begin{cases} \tilde{\beta}_1 \sum_{i=1}^{n} x_i^2 + \tilde{\beta}_0 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i \times y_i \\ \tilde{\beta}_1 \sum_{i=1}^{n} x_i + \tilde{\beta}_0 \times n = \sum_{i=1}^{n} y_i. \end{cases}$$

Решением системы нормальных уравнений являются оценки неизвестных коэффициентов модели парной регрессии:

$$\tilde{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i \times y_i - \sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\overline{xy} - \bar{x} \times \bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\mathrm{cov}(x,y)}{G^2(x)},$$

$$\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1 \bar{x},$$

где $\bar{y}$ — среднее значение результативной переменной;

$\bar{x}$ — среднее значение факторной переменной;

$\overline{xy}$ — среднее арифметическое значение произведения результативной и факторной переменных;

$G^2(x)$ — дисперсия факторной переменной;

$\mathrm{cov}(x, y)$ — ковариация между результативной и факторной переменными.

Для проверки правильности оценки коэффициентов модели регрессии может быть проведено сравнение сумм $\sum y = \sum \tilde{y}$ (при этом допустимо небольшое расхождение из-за округления расчетов).

ПРОВЕРКА ГИПОТЕЗ В МОДЕЛИ ПАРНОЙ РЕГРЕССИИ

ПРОВЕРКА ГИПОТЕЗЫ О ЗНАЧИМОСТИ ПАРНОГО КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ

Значимость парного коэффициента корреляции между факторной переменной x и результативной переменной y означает его значимое отличие от нуля.

Основной гипотезой, выдвигаемой при проверке значимости коэффициента корреляции, является гипотеза H_0 о незначимости полученного коэффициента: $H_0 \div r_{yx} = 0$. **Обратной** (или **альтернативной**) является гипотеза H_1 о значимости парного коэффициента корреляции: $H_0 \div r_{yx} \neq 0$.

Выдвинутые гипотезы проверяются с помощью t-статистики или t-критерия **Стьюдента** в том случае, если объем выборки достаточно велик ($n \geq 30$) и коэффициент корреляции по модулю значительно меньше единицы $0{,}45 \leq |r_{yx}| \leq 0{,}75$. Наблюдаемое значение t-критерия $t_{набл}$ сравнивают со значением t-критерия, определяемым по таблице распределения Стьюдента, или с критическим значением $t_{крит}$.

Критическое значение t-критерия:

$$t_{крит}(\alpha; n-h),$$

где α — уровень значимости;

h — число оцениваемых по выборке коэффициентов;

$(n-h)$ — число степеней свободы, определяется по таблице распределений t-критерия Стьюдента.

Наблюдаемое значение t-критерия Стьюдента для проверки гипотезы $H_0 \div r_{yx} = 0$ в случае линейной модели парной регрессии:

$$t_{набл} = \frac{r_{yx}}{\sqrt{1-r_{yx}^2}} \times (n-2),$$

где r_{yx} — выборочный парный коэффициент корреляции между переменными x и y.

Если $|t_{набл}| > t_{крит}$, т.е. модуль наблюдаемого значения t-критерия больше критического значения t-критерия, то с вероятностью $(1-\alpha)$ основная гипотеза о незначимости парного линейного коэффициента корреляции отвергается. Между переменными x и y существует корреляционная связь, которую можно оценить с помощью построения модели парной регрессии.

Если $|t_{набл}| \leq t_{крит}$, т.е. модуль наблюдаемого значения t-критерия меньше или равен критическому значению t-критерия, то с вероятностью α основная гипотеза о незначимости коэффициента корреляции принимается.

Основная гипотеза H_0 может быть проверена (помимо t-критерия) с помощью ***z*-статистики Фишера** в том случае, если модуль парного коэффициента корреляции близок к единице.

Проверка основной гипотезы $H_0 \div r_{yx} = 0$ отождествляется с проверкой гипотезы о незначимости величины z — $H_0 \div z = 0$:

$$t_{набл} = \frac{z}{\omega(z)},$$

где $\omega(z)$ — стандартная ошибка величины z.

Критическое значение $t_{крит}$ определяется по таблице нормального распределения (z-распределения) с доверительной вероятностью $(1-\alpha)$.

ПРОВЕРКА ГИПОТЕЗЫ О ЗНАЧИМОСТИ КОЭФФИЦИЕНТОВ МОДЕЛИ ПАРНОЙ РЕГРЕССИИ

Проверка гипотезы о значимости коэффициентов модели парной регрессии является весьма важным этапом перед практическим использованием построенной модели регрессии. Значимость коэффициентов означает их значимое отличие от нуля.

Выдвинутые гипотезы проверяются с помощью t-статистики или t-критерия **Стьюдента**. При этом наблюдаемое значение t-критерия $t_{набл}$ сравнивают со значением t-критерия, определяемым по таблице распределения Стьюдента, или с критическим значением $t_{крит}$.

Критическое значение t-критерия $t_{крит}(\alpha; n-k)$ зависит от уровня значимости и числа степеней свободы.

Уровень значимости α определяется как $\alpha = 1 - \gamma$, где величина γ называется доверительной вероятностью попадания оцениваемого параметра в доверительный интервал. Доверительную вероятность необходимо брать близкую к единице (0,95, 0,99).

Число степеней свободы определяется как разность между объемом выборки (n) и числом оцениваемых параметров по данной выборке (h). Для модели парной линейной регрессии число степеней свободы равно $(n-2)$, так как по выборке оцениваются только два параметра β_0 и β_1.

Наблюдаемое значение t-критерия Стьюдента для проверки гипотезы $H_0 \div \beta_0 = 0$:

$$t_{набл} = \frac{\tilde{\beta}_0}{\omega(\beta_0)},$$

где $\tilde{\beta}_0$ — оценка коэффициента модели регрессии β_0;

$\omega(\beta_0)$ — величина стандартной ошибки коэффициента модели регрессии β_0.

Наблюдаемое значение t-критерия Стьюдента для проверки гипотезы $H_0 \div \beta_1 = 0$:

$$t_{набл} = \frac{\tilde{\beta}_1}{\omega(\beta_1)},$$

где $\tilde{\beta}_1$ — оценка коэффициента модели регрессии β_1;

$\omega(\beta_1)$ — величина стандартной ошибки коэффициента модели регрессии β_1.

Если $|t_{набл}| > t_{крит}$, т.е. модуль наблюдаемого значения t-критерия больше критического значения t-критерия, то с вероятностью $(1-\alpha)$ основная гипотеза о незначимости коэффициентов модели регрессии отвергается (коэффициенты модели регрессии значимо отличаются от нуля).

Если $|t_{набл}| \le t_{крит}$, т.е. модуль наблюдаемого значения t-критерия меньше или равен критическому значению t-критерия, то с вероятностью α основная гипотеза о незначимости коэффициентов модели регрессии принимается (коэффициенты модели регрессии почти не отличаются от нуля или равны нулю).

УРАВНЕНИЕ ПАРНОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

$$\hat{Y} = b_0 + b_1 x$$

где по МНК

$$b_1 = r_{xy} \frac{\delta_y}{\delta_x} \quad b_0 = \bar{y} - b_1 \bar{x}$$

Значимость уравнения подтверждается коэффициентом детерминации, который в этом случае: $R^2 = r_{xy}^2$, чем ближе к 1, тем лучше качество уравнения регрессии

Критерий значимости Фишера, n – число наблюдений, m – число параметров в модели регрессии, m=p+1(для парной оно равно 2):

$$F = \frac{r^2 (n-m)}{(1-r^2)(m-1)} > F_\alpha (k_1 = m-1; k_2 = n-m)$$

ПРИМЕР

Между объемом продукции и прямыми материальными затратами на её производство установлена линейная зависимость на основе $r_{xy}=0,866$, $n=7$. Необходимо обосновать, что уравнение парной линейной регрессии значимо.

$R^2=r^2=0,866^2=0,75$ – на 75% вариация прямых материальных затрат объясняется вариацией объема продукции. В случае парной линейной регрессии $m=2$.

$F=(0,75(7-2))/((1-0,75)(2-1))=15>F0,05(1;5)=6,6$

Если построить уравнение, оно значимо с вероятностью 95%.

Доверительный интервал для линии регрессии в случае парной регрессии

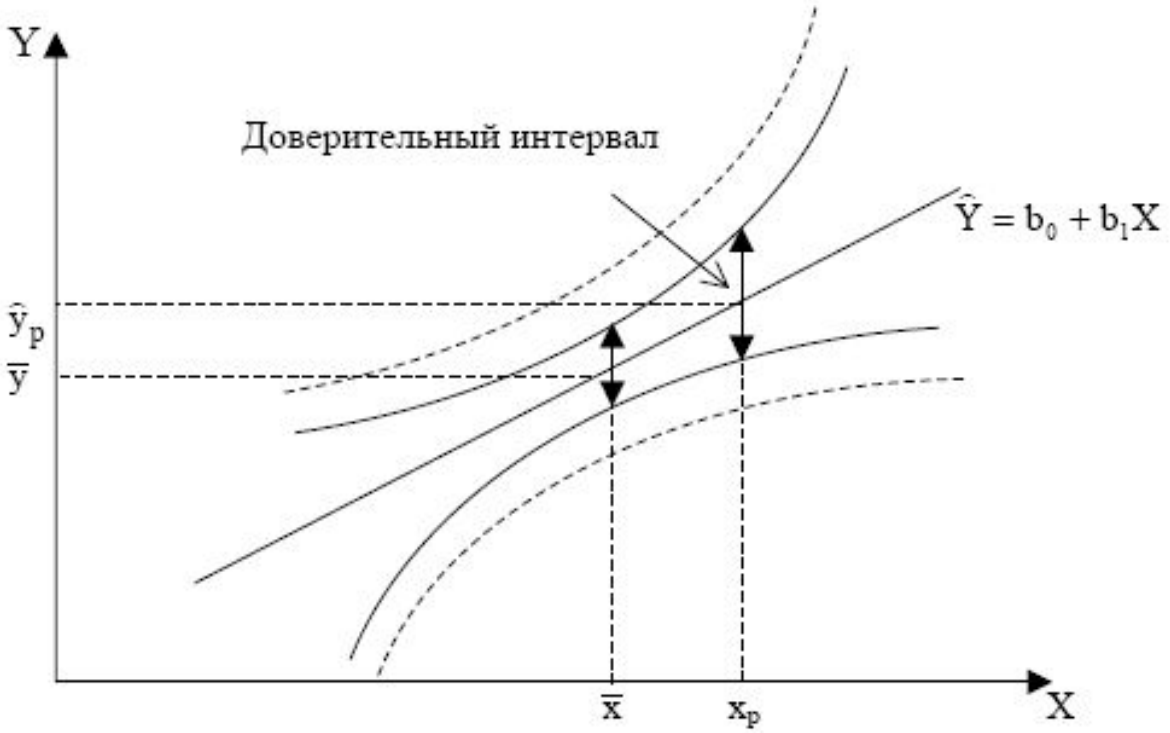
$$\hat{Y}(x_k) \pm t_{\alpha/2}(n-2) \times \left[S \sqrt{\frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}} \right], \text{ где}$$

$$S = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - p - 1}}, S^2 - \text{остаточная дисперсия},$$

хараетеризует необъясненную часть вариации y;

x_k *– значение фактора, для которого строят доверительный интервал*

ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ В МОДЕЛИ



интервал
определяет границы, за пределами которых могут оказаться не более 100α % точек наблюдений при $X = x_p$. Заметим, что данный интервал шире доверительного интервала для условного математического ожидания (на рис. границы этого интервала отмечены пунктирной линией).

Проводя анализ построенных интервалов, несложно заметить, что наиболее узкими они будут при $X_p = \bar{x}$. По мере удаления X_p от среднего значения доверительные интервалы расширяются
Поэтому необходимо достаточно осторожно экстраполировать полученные результаты на прогнозные области. С другой стороны, с ростом числа наблюдений n эти интервалы сужаются к линии регрессии при $n \to \infty$.

ИНТЕРВАЛЬНЫЙ ПРОГНОЗ НА ОСНОВЕ УРАВНЕНИЯ ПАРНОЙ РЕГРЕССИИ

$$\hat{Y}(x_k) \pm t_{\alpha/2}(n-2) S \left[\sqrt{1 + \frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}} \right], \text{ где}$$

$$S = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - p - 1}}, S^2 - \text{остаточная дисперсия},$$

характеризует необъясненную часть вариации y;

x_k – значение фактора, для которого строят доверительный интервал

Применение функции «Тенденция»

i	y	x1	$\hat{y}$	ε
1	5	8	5,377	-0,37705
2	10	11	8,426	1,57377
3	10	12	9,443	0,557377
4	7	9	6,393	0,606557
5	5	8	5,377	-0,37705
6	6	8	5,377	0,622951
7	6	9	6,393	-0,39344
8	5	9	6,393	-1,39344
9	6	8	5,377	0,622951
10	8	12	9,443	-1,44262
Итого	**68**	94	**68**	4,44E-15

Применение функции «Линейн»

	a1	a0		
	1,016393	-2,7541		
	0,207362	1,975937	Ошибки a1 и a0	
R^2	0,750195	1,024295		
F	24,025	8	n-m	
	25,20656	8,393443		
	RSS	ESS		

Применение инструмента Regression

Regression Statistics	
Multiple R	0,866138072
R Square	0,75019516
Adjusted R Square	0,718969555
Standard Error	1,024295039
Observations	10

ANOVA	*df*	*SS*	*MS*	*F*
Regression	1	25,20655738	25,20655738	24,025
Residual	8	8,393442623	1,049180328	
Total	9	33,6		

			Вероятность ошибки	
	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	**-2,754098361**	1,975936941	-1,393818954	0,20087
X Variable	**1,016393443**	0,20736247	4,901530373	0,00119
		tтабл	2,306	

Смысл коэффициентов регрессии в уравнении У(х)= b_0 + b_1 Х

- b_0 – отражает усредненной влияние всех неучтенных факторов
- b_1 – означает среднее изменение величины у, в зависимости от изменения значений переменной х, если остальные факторы, влияющие на у и не связанные с х, неизменны

Поэтому если константа, включенная в модель делает уравнение значимым, когда оно незначимо без нее, то эта модель неверна

Знак при коэффициенте регрессии показывает:

- Для коэффициента в, если $b_1 < 0$, то связь прямая, если $b_1 > 0$, то связь обратная
- Для коэффициента регрессии b_0 , если $b_0 > 0$, то изменение результата происходит медленнее, чем изменение фактора, то есть $V_x > V_y$

Расчетные формулы

1. Оценки коэффициентов однофакторной регрессионной модели:

$$b_1 = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2}, \qquad b_0 = \overline{y} - b_1\overline{x},$$

где

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, \qquad \overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i, \qquad \overline{xy} = \frac{1}{N}\sum_{i=1}^{N} x_i y_i, \qquad \overline{x^2} = \frac{1}{N}\sum_{i=1}^{N} x_i^2,$$

x - независимая переменная, y - зависимая переменная, N - число элементов выборочной совокупности.

2. Коэффициент корреляции:

$$r_{xy} = b_1\frac{\sigma_x}{\sigma_y} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\sigma_x\sigma_y},$$

где σ_x, σ_y - среднеквадратические ошибки, вычисляемые по формулам

$$\sigma_x = \sqrt{\frac{1}{n}\sum x_i^2 - \overline{x}^2}, \quad \sigma_y = \sqrt{\frac{1}{n}\sum y_i^2 - \overline{y}^2}.$$

3. Коэффициент детерминации:

$$D = r^2.$$

4. Дисперсионное отношение Фишера (*F*-критерий):

$$F_{расч} = \frac{\sum(\hat{y} - \bar{y})^2 / m}{\sum(y - \hat{y})^2 / (n - m - 1)} = \frac{r_{xy}^2}{1 - r_{xy}^2}(n - 2),$$

где $\hat{y}$ – расчетное значение зависимой переменной ($\hat{y} = b_0 + b_1 x$), n – число элементов выборочной совокупности, m – число факторов.

5. Стандартные ошибки параметров линейной регрессии:

$$s_{b_1} = \sqrt{\frac{\sum(y - \hat{y})^2 / (n - 2)}{\sum(x - \bar{x})^2}} = \sqrt{\frac{S_{ост}^2}{\sum(x - \bar{x})^2}} = \frac{S_{ост}}{\sigma_x \sqrt{n}},$$

$$s_{b_0} = \sqrt{\frac{\sum x^2}{n\sum(x - \bar{x})^2} \cdot \frac{\sum(y - \hat{y})^2}{(n - 2)}} = \sqrt{S_{ост}^2 \frac{\sum x^2}{n^2 \sigma_x^2}} = S_{ост} \frac{\sqrt{\sum x^2}}{n\sigma_x},$$

где $S_{ост}^2$ – остаточная дисперсия, рассчитываемая по формуле

$$S_{ост}^2 = \frac{\sum(y - \hat{y})^2}{n - m - 1}.$$

6. t-статистики Стьюдента:

$$t_{b_0} = \frac{b_0}{s_{b_0}}, \quad t_{b_1} = \frac{b_1}{s_{b_1}}.$$

При оценке значимости коэффициента линейной регрессии на начальном этапе можно использовать следующее *"грубое" правило*, позволяющее не прибегать к таблицам.

Если стандартная ошибка коэффициента больше его модуля ($|t| < 1$), то коэффициент не может быть признан значимым, т. к. доверительная вероятность здесь при двусторонней альтернативной гипотезе составит менее чем 0.7.

Если $1 < |t| < 2$, то найденная оценка может рассматриваться как относительно (слабо) значимая. Доверительная вероятность в этом случае лежит между значениями 0.7 и 0.95.

Если $2 < |t| < 3$, то это свидетельствует о значимой линейной связи между X и Y. В этом случае доверительная вероятность колеблется от 0.95 до 0.99.

Наконец, если $|t| > 3$, то это почти гарантия наличия линейной связи.

Доверительные интервалы для коэффициентов уравнения регрессии

7. Доверительные интервалы:

$$b_0 - \Delta_{b_0} \le b_0 \le b_0 + \Delta_{b_0}, \quad b_1 - \Delta_{b_1} \le b_1 \le b_1 + \Delta_{b_1},$$

где Δ_{b_0}, Δ_{b_1} – предельные ошибки, рассчитываемые по формулам

$$\Delta_{b_0} = t_{табл} S_{b_0}, \quad \Delta_{b_1} = t_{табл} S_{b_1},$$

$t_{табл}$ – табличное значение t-статистики.

8. Индекс корреляции: $\sqrt{1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}}$.

9. Усредненное значение коэффициента эластичности: $b_1 \cdot \frac{\bar{x}}{\bar{y}}$

Решение типовых задач

Задание По данным табл. 1. построить линейное уравнение регрессии, отражающее зависимость стоимости квартиры от ее жилой площади.

Таблица 1.

№ п.п.	Стоимость (долл.)	Жилая площадь (кв. м.)		№ п.п	Стоимость (долл.)	Жилая площадь (кв. м.)
1.	5000	30,2		9.	5740	33
2.	5200	32		10.	5570	31
3.	5350	32		11.	5530	30
4.	5880	37		12.	6020	34
5.	5430	30		13.	7010	38
6.	5430	30		14.	6420	31
7.	5430	30		15.	7150	39
8.	5350	29		16.	7190	39,5

Для построенного уравнения вычислить

1) коэффициент корреляции;
2) коэффициент детерминации;
3) дисперсионное отношение Фишера;
4) стандартные ошибки коэффициентов регрессии;
5) *t*-статистики Стьюдента;
6) доверительные границы коэффициентов регрессии.

Дать содержательную интерпретацию коэффициента регрессии построенной модели.

Решение с помощью табличного процессора Excel.

1. Ввод исходных данных.

2. Подготовка данных и оформление их в виде табл. *2.* для расчета оценок коэффициентов регрессии.

Таблица 2.

№ п.п.	y	x	x^2	xy	y^2
1.	5000	30,2	912,04	151000	25000000
2.	5200	32	1024	166400	27040000
3.	5350	32	1024	171200	28622500
4.	5880	37	1369	217560	34574400
5.	5430	30	900	162900	29484900
6.	5430	30	900	162900	29484900
7.	5430	30	900	162900	29484900
8.	5350	29	841	155150	28622500
9.	5740	33	1089	189420	32947600
10.	5570	31	961	172670	31024900
11.	5530	30	900	165900	30580900
12.	6020	34	1156	204680	36240400
13.	7010	38	1444	266380	49140100
14.	6420	31	961	199020	41216400
15.	7150	39	1521	278850	51122500
16.	7190	39,5	1560,3	284005	51696100
Среднее значение	5856,25	32,86	1091,39	194433,44	34767688,50

3. Расчет коэффициентов регрессии:

$$b_1 = \frac{194433{,}44 - 32{,}86 \cdot 5856{,}25}{1091{,}39 - 32{,}86^2} = 170{,}239;$$

$$b_0 = 5856{,}25 - 170{,}239 \cdot 32{,}86 = 262{,}847.$$

Построенная модель может быть записана в следующем виде:

$$y = 262{,}847 + 170{,}239x.$$

Коэффициент регрессии b_1 этой модели показывает, что в среднем увеличение полезной площади на 1 кв. м. приводит к увеличению ее стоимости на 170,24 долл.

4. Расчет коэффициента корреляции и детерминации

$$\sigma_x = \sqrt{1091{,}39 - 32{,}86^2} = 3{,}444; \qquad \sigma_y \quad \sqrt{3476768{,}50 - 5856{,}25^2} = 687{,}040;$$

$$r = 170{,}239 \cdot \frac{3{,}444}{687{,}040} = 0{,}853; \qquad D = 0{,}853^2 \cdot 100\% = 72{,}818\%.$$

Коэффициент корреляции достаточно высокий, что свидетельствует о существенной зависимости стоимости квартир от полезной площади. Коэффициент детерминации показывает, что величина стоимости квартиры объясняется величиной полезной площади только на 72,82 %.

5. Расчет дисперсионного отношения Фишера

$$F_{расч} = \frac{0{,}853^2}{(1 - 0{,}853^2)}\ 14 = 37{,}504.$$

Сравнение расчетного значения F-критерия с табличным $F_{1;\,14} = 4{,}60$ для 95%-ного уровня значимости позволяет сделать вывод об адекватности построенной модели.

6. Расчет стандартных ошибок по формулам в которых используется средняя квадратическая ошибка $S_{ост}$, вычисленная в соответствии с данными табл. 1.2.3.

$$s_{b_0} = \frac{382{,}933 \cdot \sqrt{17462{,}29}}{3{,}444 \cdot 16} = 918{,}356; \quad s_{b_1} = \frac{382{,}933}{3{,}444 \cdot \sqrt{16}} = 27{,}798.$$

7. Расчет доверительных границ для коэффициентов уравнения регрессии

$$\Delta_{b_0} = 2{,}1448 \cdot 918{,}356 = 1969{,}691;$$

$$\Delta_{b_1} = 2{,}1448 \cdot 27{,}798 = 59{,}622;$$

$$262{,}847 - 1969{,}691 \le b_0 \le 262{,}847 + 1969{,}691;$$

$$-1706{,}691 \le b_0 \le 2232{,}538;$$

$$170{,}239 - 59{,}622 \le b_1 \le 170{,}239 + 59{,}622;$$

$$110{,}616 \le b_1 \le 229{,}861.$$

ЭФФЕКТИВНОСТЬ ОЦЕНОК МНК

- Оценки коэффициентов модели регрессии, полученные классическим МНК , являются наилучшими, то есть несмещенными, состоятельными и эффективными, если выполняются предпосылки теоремы Гаусса-Маркова

Предпосылки МНК (условия Гаусса–Маркова)

1 . Математическое ожидание случайного отклонения ε_i равно нулю: $M(\varepsilon_i) = 0$ для всех наблюдений.

Данное условие означает, что случайное отклонение в среднем не оказывает влияния на зависимую переменную. В каждом конкретном наблюдении случайный член может быть либо положительным, либо отрицательным, но он не должен иметь систематического смещения.

2 . Дисперсия случайных отклонений ε_i постоянна:

$$D(\varepsilon_i) = D(\varepsilon_j) = \sigma^2 \text{ для любых наблюдений } i \text{ и } j.$$

Данное условие подразумевает, что несмотря на то, что при каждом конкретном наблюдении случайное отклонение может быть либо большим, либо меньшим, не должно быть некой априорной причины, вызывающей большую ошибку (отклонение).

Выполнимость данной предпосылки называется *гомоскедастичностью (постоянством дисперсии отклонений).* Невыполнимость данной предпосылки называется *гетероскедастичностью (непостоянством дисперсий отклонений).*

3 . Случайные отклонения ε_i и ε_j являются независимыми друг от друга для $i \neq j$.

Выполнимость данной предпосылки предполагает, что отсутствует систематическая связь между любыми случайными отклонениями. Другими словами, величина и определенный знак любого случайного отклонения не должны быть причинами величины и знака любого другого отклонения.

если данное условие выполняется, то говорят об отсутствии *автокорреляции.*

4 . Случайное отклонение должно быть независимо от объясняющих переменных.

Обычно это условие выполняется автоматически при условии, что объясняющие переменные не являются случайными в данной модели.

Следует отметить, что выполнимость данной предпосылки не столь критична для эконометрических моделей.

5 . Модель является линейной относительно параметров.

Теорема Гаусса–Маркова. Если предпосылки *1 – 5* выполнены, то оценки, полученные по МНК, обладают следующими свойствами:

1. Оценки являются несмещенными, т. е. $M(b_0) = в_0, M(b_1) = в_1$. Это вытекает из того, что $M(e_i) = 0$ и говорит об отсутствии систематической ошибки в определении положения линии регрессии.
2. Оценки состоятельны, т. к. дисперсия оценок параметров при возрастании числа n наблюдений стремится к нулю: $D(b_0) \xrightarrow[n\to\infty]{} 0$, $D(b_1) \xrightarrow[n\to\infty]{} 0$. Другими словами, при увеличении объема выборки надежность оценок увеличивается (b_0 наверняка близко к β_0, b_1 – близко к β_1).
3. Оценки эффективны, т. е. они имеют наименьшую дисперсию по сравнению с любыми другими оценками данных параметров, линейными относительно величин y_i.

В англоязычной литературе такие оценки называются *BLUE (Best Linear Unbiased Estimators) – наилучшие линейные несмещенные оценки.*

Если предпосылки *2* и *3* нарушены, т. е. дисперсия отклонений непостоянна и (или) значения e_i, e_j связаны друг с другом, то свойства несмещенности и состоятельности сохраняются, но свойство эффективности – нет.

Наряду с выполнимостью указанных предпосылок при построении классических линейных регрессионных моделей делаются еще некоторые предположения :

- объясняющие переменные не являются случайными величинами;
- случайные отклонения имеют нормальное распределение;
- число наблюдений существенно больше числа объясняющих переменных;
- отсутствуют ошибки спецификации;
- отсутствует мультиколлинеарность.

Основные предпосылки модели парной линейной регрессии

$$Y=b_0+b_1x+\varepsilon$$

- Связь между Y и x является линейной;
- X может использоваться для прогноза Y;
- Остатки ε имеют нормальное распределение;
- Дисперсия ошибок постоянна;
- Отсутствуют ошибки спецификации;
- Ошибки являются независимыми случайными величинами.

НЕЛИНЕЙНАЯ РЕГРЕССИЯ

- Если между экономическими явлениями существуют нелинейные соотношения, то они выражаются с помощью нелинейных функций
- Различают два класса нелинейных регрессий :

1. Нелинейные по объясняющим переменным, но линейные по оцениваемым параметрам
2. Нелинейные по оцениваемым параметрам

НЕЛИНЕЙНАЯ РЕГРЕССИЯ ПО ОБЪЯСНЯЮЩИМ ПЕРЕМЕННЫМ (ошибка аддитивна)

- Полиномы

(чаще 2-ой степени)

$$\hat{y} = a + b \cdot x + c \cdot x^2$$

- Равносторонняя гипербола

(например,
кривая Филлипса, зависимость процента прироста заработной платы от уровня безработицы;
Кривая Энгеля , зависимость доли расходов на непродовольственные товары от дохода)

$$\hat{y} = a + b \cdot x + c \cdot x^2 + d \cdot x^3$$

$$\hat{y} = a + b/x$$

$$\hat{y} = a - b/x$$

НЕЛИНЕЙНАЯ РЕГРЕССИЯ ПО ПАРАМЕТРАМ (ошибка неаддитивна)

- Степенная $y = a x^{b} \varepsilon$
- Показательная $y = a b^{x} \varepsilon$
- Экспоненциальная $y = e^{a+bx} \varepsilon$

НЕЛИНЕЙНАЯ РЕГРЕССИЯ ПО ОБЪЯСНЯЮЩИМ ПЕРЕМЕННЫМ

- Применяется метод замены

($x=x_1$; $x^2=x_2$ и т.д.)

- Параметры определяются, как в линейной регрессии по МНК

НЕЛИНЕЙНАЯ РЕГРЕССИЯ ПО ПАРАМЕТРАМ

- Применяем логарифмирование
- Если после применения логарифмирования, получаем линейную зависимость, то регрессия называется внутренне линейной, если нет, то внутренне нелинейной

ПРОВЕРКА ПРАВИЛЬНОСТИ ПРИМЕНЕНИЯ ЛИНЕЙНОЙ РЕГРЕССИИ

$$\left|R^2 - r^2\right| < 0{,}1$$

- Где R^2 – индекс (коэффициент) детерминации, полученный по модели нелинейной регрессии
- Где r^2 – квадрат линейного коэффициента корреляции

ПРОВЕРКА ПРАВИЛЬНОСТИ ПРИМЕНЕНИЯ ЛИНЕЙНОЙ РЕГРЕССИИ

- Если не выполняется неравенство, то проверка сложнее на основе t-статистики

$$t = \frac{R^2 - r^2}{2\sqrt{\frac{(R^2 - r^2) - (R^2 - r^2)^2 (2 - (R^2 + r^2))}{n}}}$$

- Если $t > t_{табл}$, то различия между рассматриваемыми показателями существенны и замена нелинейной регрессии уравнением линейной функции невозможна

СРЕДНЯЯ ОШИБКА АПРОКСИМАЦИИ

- Для проверки качества уравнения регрессии применяется средняя ошибка аппроксимации
- Если она в пределах 5-7%, модель хорошо подобрана к исходным данным

$$A = \frac{1}{n}\sum \left| \frac{(y - \hat{y})}{y} \right| 100\%$$