ЛИНЕЙНАЯ ПАРНАЯ РЕГРЕССИЯ

Лекция №2

ПОСТАНОВКА ЗАДАЧИ

Пусть объясняющая переменная X и объясняемая переменная Y связаны соотношением:

$$Y=mX+b+\varepsilon, \qquad (*)$$

Эта модель является регрессионной, если $M_x Y=mx+b$, т. е. если $M\varepsilon=0$

где m *и* b - детерминированные величины, ε - случайное возмущение.

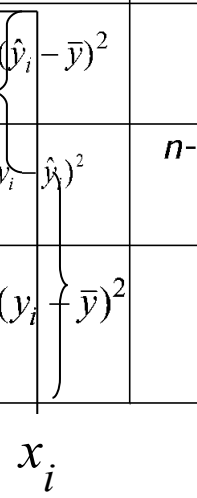
Получены наблюдения: (x_1,y_1), (x_2,y_2), …, (x_n,y_n).
Требуется по наблюдениям найти в некотором смысле наилучшие оценки значений m *и* b. Тогда оценивание Y по известному x можно производить по формуле:

Далее дается два подхода к определению таких оценок и формулируются условия, при которых эти подходы дают одинаковый результат.

МЕТОД НАИМЕНЬШИХ КВАДРАТОВ (МНК)

Обозначим:

Компоненты дисперсии	Сумма квадратов	Число степеней свободы	Средние квадраты
Регрессионная	$Q_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	p	$s_R^2 = \frac{Q_R}{p}$
Остаточная	$Q_e = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$n-(p+1)$	$s^2 = \frac{Q_e}{n-(p+1)}$
Общая	$Q = \sum_{i=1}^{n} (y_i - \bar{y})^2$	$n-1$	



Q_e - *остаточная сумма*

поле корреляции

МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

Параметры регрессии определяются из условия минимума остаточной суммы:

МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

Необходимое условие экстремума:

Откуда получаем *нормальную систему уравнений:*

МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

Обозначим:

Тогда из (1) получим:

Решая систему (2), найдем:

(3)

- выборочный коэффициент регрессии

МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

Заметим, что

s_x -выборочное среднее квадратичное отклонение x

- выборочная дисперсия X

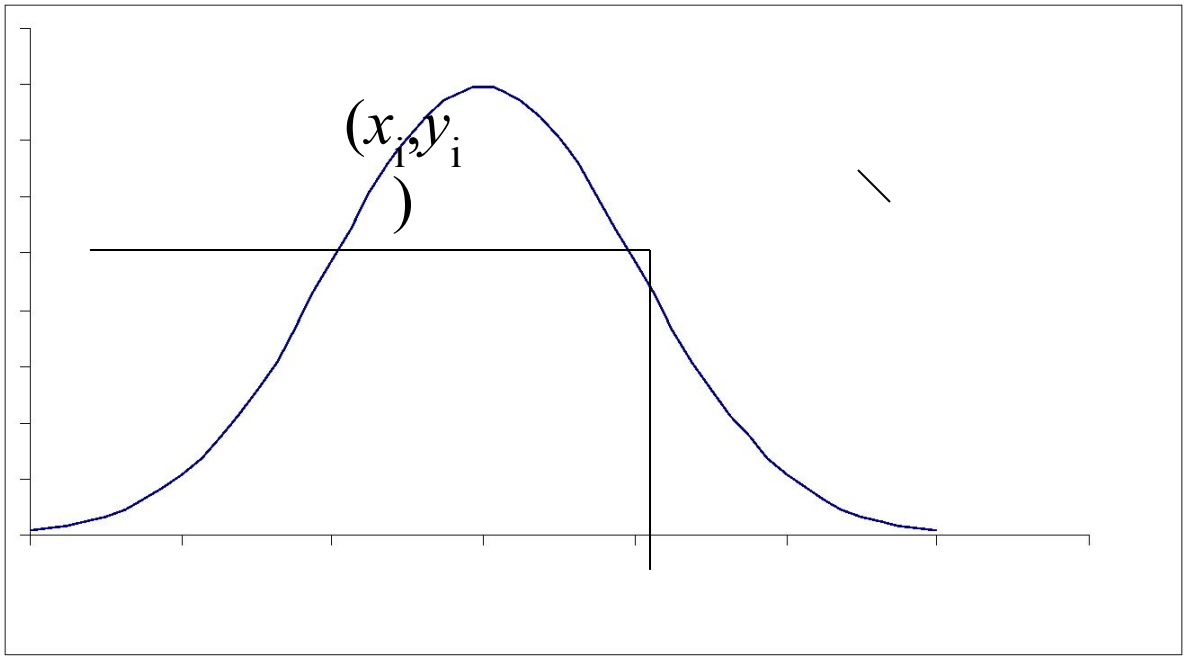
- выборочная ковариация XиY

С учетом этих обозначений получим:

(4)

МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

Из (3): прямая проходит через точку



МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

При $M\varepsilon_i=0$, $i=1,\ldots,n$, (отсутствии систематических ошибок) уравнение является уравнением регрессии, т. е.

Внимание! При получении МНК-оценок параметров b и m не использовалось никаких предположений о распределении X и Y.

КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

Представим последнее соотношение в эквивалентном виде:

где s_x, s_y - выборочные средние квадратичные отклонения *x* и *y*.
Здесь используются *нормированные и центрированные* значения *x, y.*

Нормирование позволяет избежать зависимости от их единиц измерения.

Центрирование позволяет работать с приращениями.

КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

Обозначив

r - выборочный коэффициент корреляции

получим

Коэффициент корреляции показывает, насколько величин s_y в среднем изменится y, если x изменится на s_x.

Коэффициент корреляции характеризует тесноту связи X и Y.

КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

С учетом (4) получаем:

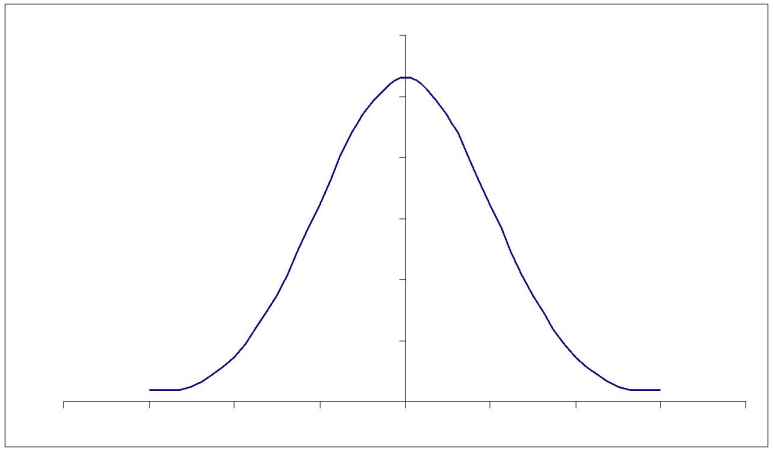
Эта формула обычно используется как определение выборочного коэффициента корреляции

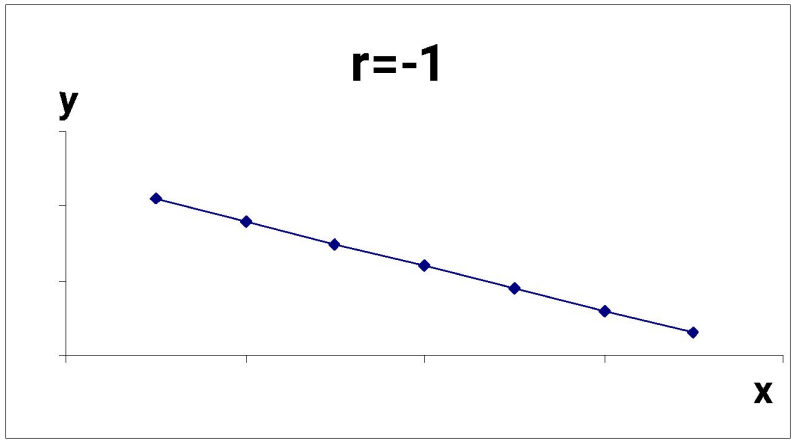
Для расчетов по таблице наблюдений применяется формула:

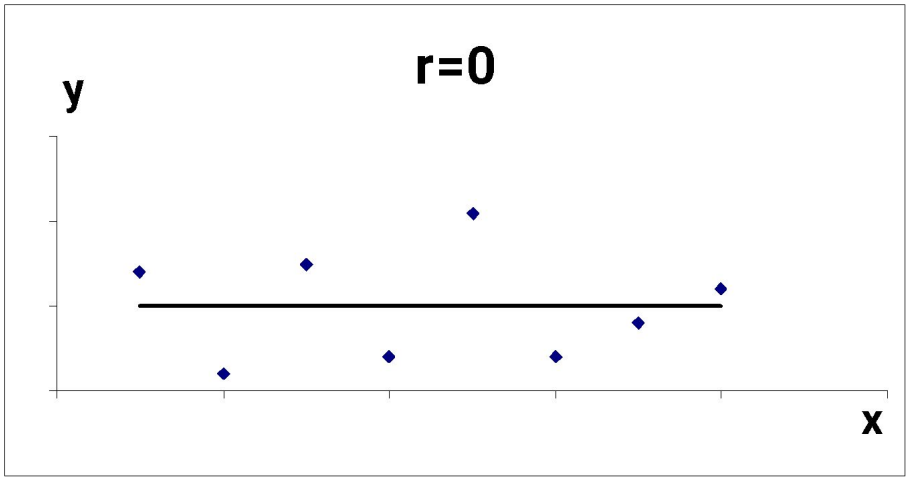
Свойства коэффициента корреляции

- 1. $-1 \le r \le 1$. Чем ближе $|r|$ к 1, тем теснее связь.
- 2. При $r=\pm 1$ корреляционная связь - линейная (наблюдения располагаются на прямой)
- 3. При $r=0$ связь отсутствует, линия регрессии параллельна оси OX.

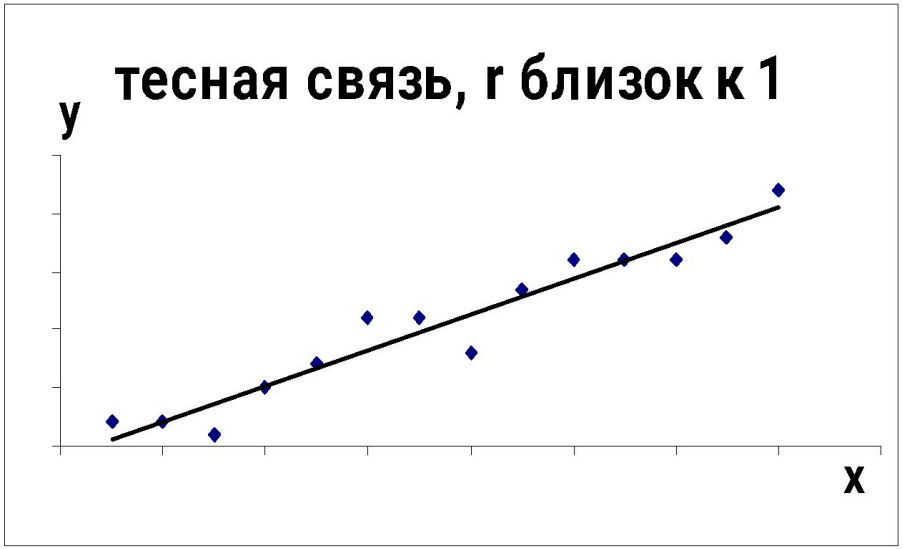
Коэффициент корреляции характеризует тесноту связи

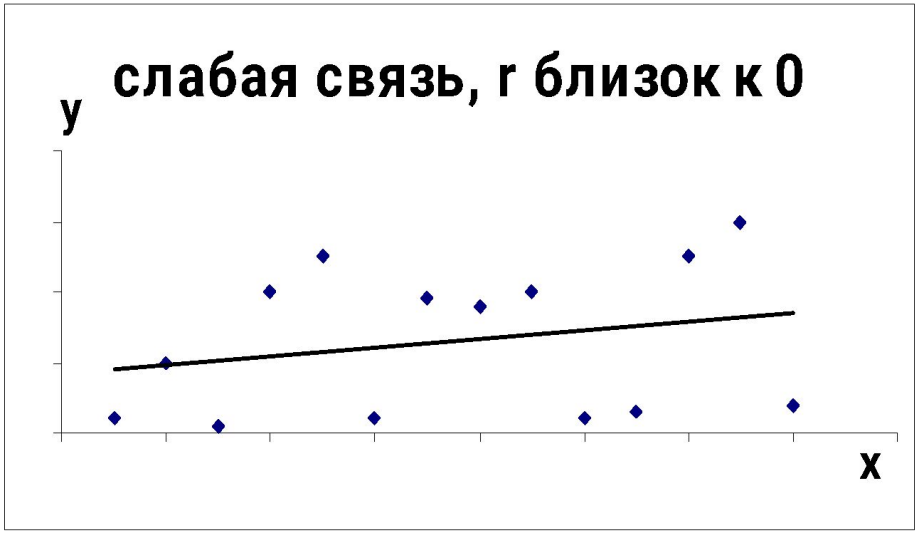






Коэффициент корреляции характеризует тесноту связи





Классическая нормальная линейная регрессионная модель

Предположим, что:

X - детерминированная величина;

$\varepsilon_1, \ldots, \varepsilon_n$ - независимые нормальные одинаково распределенные случайные величины: $\varepsilon_i \sim N(0, \sigma^2)$.

В этих предположениях соотношение

$$Y = mX + b + \varepsilon$$

называется *классической нормальной линейной регрессионной моделью*
(Classical Normal Linear Regression model).

МЕТОД МАКСИМАЛЬНОГО ПРАВДОПОДОБИЯ

Для упрощения выкладок можно вместо функции (5) максимизировать ее логарифм (т. к. логарифм - монотонная функция):

Из (6) следует, что **при известной дисперсии σ^2** для нахождения оценок МП достаточно минимизировать Q_e, и, следовательно, **МП-оценка совпадает с НК-оценкой.**

Свойства оценок: несмещенность

Оценка является **несмещенной** оценкой параметра θ, если:

Математическое ожидание оценки равно оцениваемому параметру.

МП-оценки могут иметь смещение!

Свойства оценок: состоятельность

Обозначим -оценка параметра , полученная по n наблюдениям.

называется **состоятельной оценкой**, если сходится по вероятности к :

для состоятельной несмещенной оценки выполняется закон больших чисел

Свойства оценок: состоятельность

Другая формулировка закона больших чисел - **неравенство Чебышева:**

Может использоваться для определения необходимого числа наблюдений, если задано допустимое отклонение оценки от оцениваемого параметра и допустимая вероятность отклонения.

Свойства оценок: эффективность

Эффективной называется оценка, обладающая минимальной дисперсией:

-любая другая оценка.

Для несмещенных оценок:

Несмещенность оценок параметров регрессии

формула (3)

Доказательство несмещенности параметра *m*:

Несмещенность оценок параметров регрессии

В силу детерминированности X, свойств математического ожидания и определения

Так как $My_i = mx_i + b$, то, после приведения подобных членов, получаем

что и требовалось доказать

Несмещенность оценок параметров регрессии

Доказательство несмещенности параметра *b*:

В силу свойств математического ожидания и детерминированности X, получаем:

Так как и по определению имеем:

Несмещенность оценок параметров регрессии

Так как $My_i = mx_i + b$, и в силу свойств математического ожидания получаем:

что и требовалось доказать

Теорема Гаусса-Маркова

В условиях классической нормальной регрессионной модели оценки (3) имеют наименьшую дисперсию в классе всех линейных несмещенных оценок.

(3) - самые эффективные оценки -

Best Linear Unbiased Estimates (BLUE)

ОЦЕНИВАНИЕ ДИСПЕРСИИ ВОЗМУЩЕНИЙ

Оценка максимального правдоподобия

Несмещенная оценка

Число неизвестных параметров (m, b)

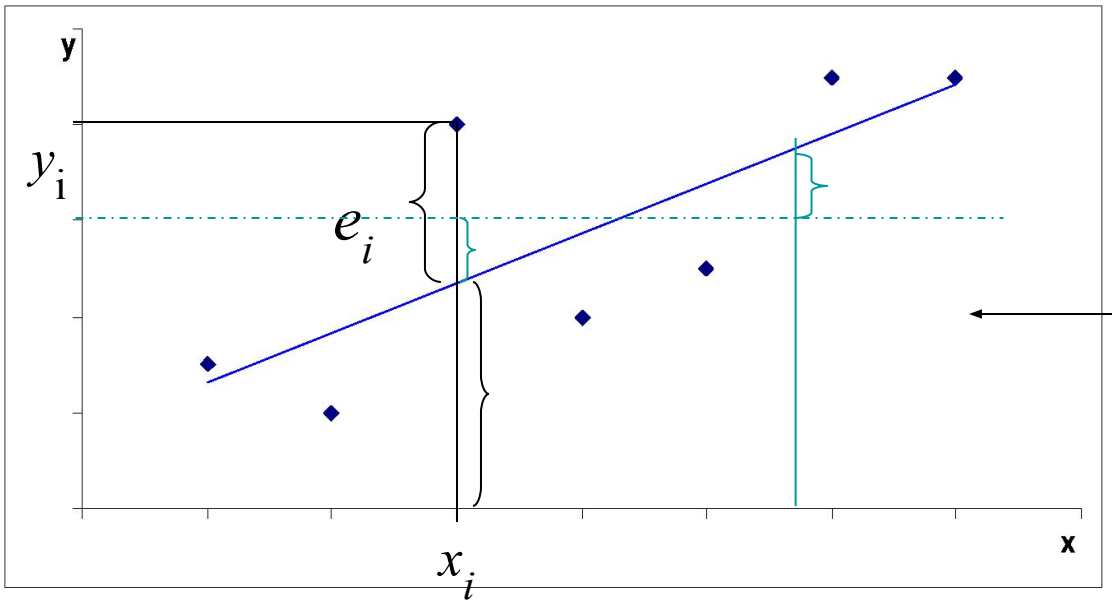
где

Оценки b, m, σ независимы

?для нормированных наблюдений

ОЦЕНКА ЗНАЧИМОСТИ УРАВНЕНИЯ РЕГРЕССИИ

Обозначим:



Q_e - *остаточная сумма*

поле корреляции

ОЦЕНКА ЗНАЧИМОСТИ УРАВНЕНИЯ РЕГРЕССИИ

Компоненты дисперсии	Сумма квадратов	Число степеней свободы	Средние квадраты
Регрессионная	$Q_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	*p*	$s_R^2 = \frac{Q_R}{p}$
Остаточная	$Q_e = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	*n-(p+1)*	$s^2 = \frac{Q_e}{n-(p+1)}$
Общая	$Q = \sum_{i=1}^{n} (y_i - \bar{y})^2$	*n-1*	

ОЦЕНКА ЗНАЧИМОСТИ УРАВНЕНИЯ РЕГРЕССИИ

- Чем меньше остаточная сумма Q_e, тем выше качество модели.
- Чем больше регрессионная сумма Q_R, тем выше качество модели.
- Чем больше отношение Q_R/Q_e, тем выше качество модели.
- Для перехода к стандартному распределению следует рассматривать не суммы, а средние квадраты.

ОЦЕНКА ЗНАЧИМОСТИ УРАВНЕНИЯ РЕГРЕССИИ

Можно доказать, что

В условиях теоремы Гаусса-Маркова

ОЦЕНКА ЗНАЧИМОСТИ УРАВНЕНИЯ РЕГРЕССИИ

F показывает, в какой мере регрессия лучше оценивает значение зависимой переменной по сравнению с ее средним значением

При

(!)

гипотеза (предположение) о незначимости регрессии отклоняется

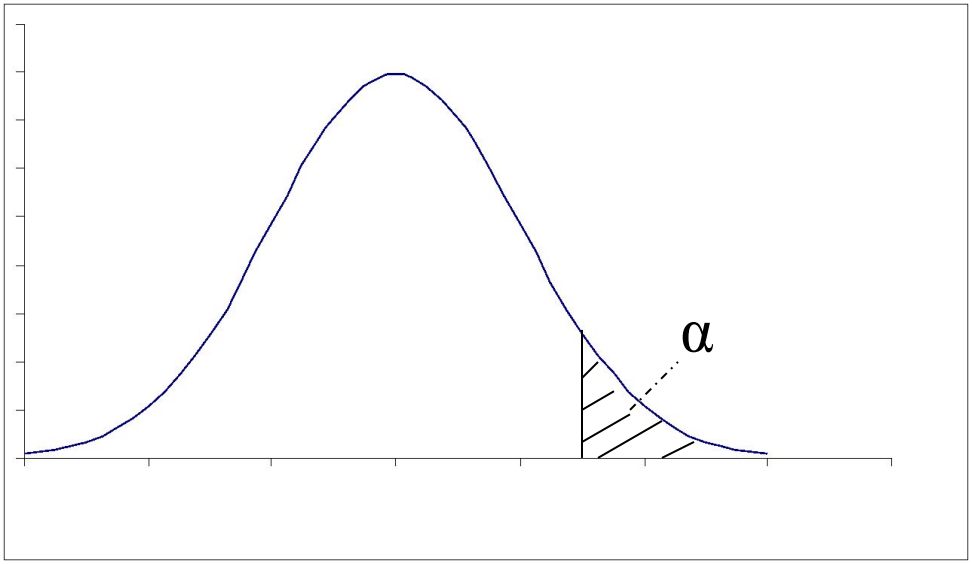
с уровнем значимости α.

Неравенство **(!) - правило (критерий) проверки гипотезы о незначимости линейной регрессии,**

$f_{\alpha,k1,k2}$ — квантиль распределения Фишера, критическое (пороговое) значение F,

α - вероятность отклонения гипотезы при условии, что она верна - вероятность ошибки первого рода.

Квантили F-распределения Фишера-Снедекора



MS Excel 2010: $f_{\alpha,k1,k2}$=F.Обр.ПХ(α, k1,k2)

α - вероятность «хвоста», которую также называют уровнем значимости или вероятностью ошибки 1 рода (вероятность отклонить гипотезу о незначимости при условии, что она верна).

ОЦЕНКА ЗНАЧИМОСТИ УРАВНЕНИЯ РЕГРЕССИИ

Коэффициент детерминации:

Коэффициент детерминации показывает, какая часть изменения зависимой переменной объясняется изменением объясняющей переменной.

$$0 \le R^2 \le 1$$

1. Чем ближе R^2 к единице, тем лучше регрессия аппроксимирует наблюдения.
2. Если $R^2=1$, то наблюдения лежат на линии регрессии.
3. Если $R^2=0$, то изменение зависимой переменной полностью обусловлено неучтенными в модели факторами, и линия регрессии параллельна оси OX.

ОЦЕНКА ЗНАЧИМОСТИ УРАВНЕНИЯ РЕГРЕССИИ

Можно доказать, что в случае парной регрессии:

$$R^2 = r^2$$

ОЦЕНКА ЗНАЧИМОСТИ УРАВНЕНИЯ РЕГРЕССИИ

- $Q_R > Q_e$ – регрессионная модель значима;
- $Q_R < Q_e$ – регрессионная модель незначима;
- $Q_R = Q_e$ – граничный случай; $R^2 = 0.5$;
- $R^2 \geq 0.5$ ($r \geq 0.7$) – регрессионная модель значима.

ОЦЕНКА ЗНАЧИМОСТИ УРАВНЕНИЯ РЕГРЕССИИ

Критерий (!) проверки гипотезы о незначимости регрессии может использовать значение R^2 :

Статистика *F* Фишера (Фишера-Снедекора) и коэффициент детерминации R^2 связаны друг с другом

ОЦЕНКА ЗНАЧИМОСТИ УРАВНЕНИЯ РЕГРЕССИИ

Другой способ оценки значимости уравнения ПАРНОЙ регрессии -

проверка гипотезы $m=0$

если $m=0$, то y не зависит от x.

Можно ли по значению **оценки** m судить о справедливости этой гипотезы?

Если гипотеза верна, то

большие значения оценки маловероятны

x_α - квантиль стандартного нормального распределения, α - суммарная вероятность двух «хвостов» (в эконометрике α=0.05)

СКО оценки m не знаем, используем выборочное СКО:

ОЦЕНКА ЗНАЧИМОСТИ УРАВНЕНИЯ РЕГРЕССИИ

СКО оценки *m* не знаем, используем выборочное СКО:

-статистика Стьюдента с числом степеней свободы n-2, $t_{\alpha;n-2}$ – ее квантиль уровня α (суммарная вероятность **двух** «хвостов»)

Правило проверки гипотезы о незначимости уравнения регрессии: гипотеза отклоняется при

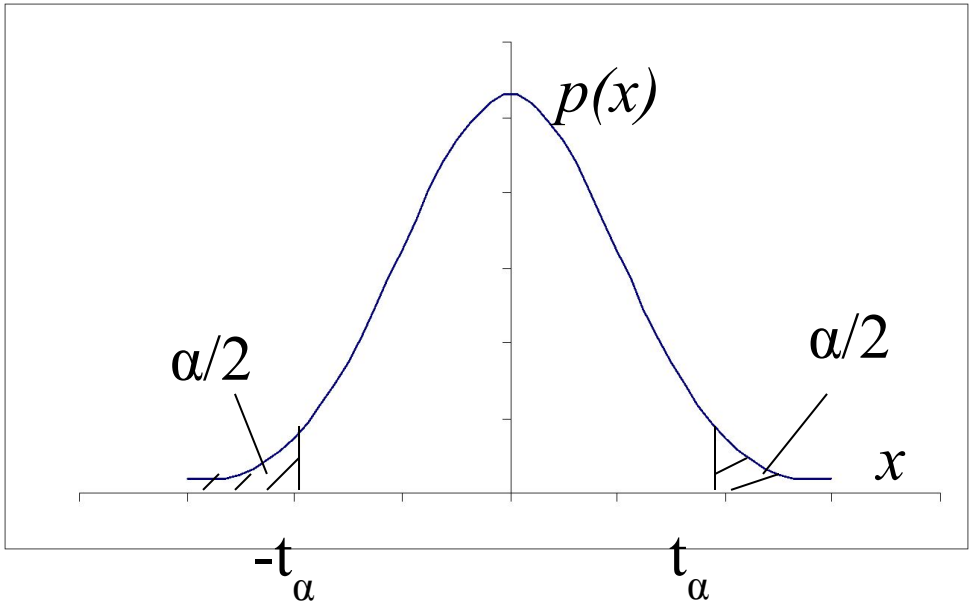
$$|T_{n-2}| > t_{\alpha,n-2}, \quad (!!)$$

ОЦЕНКА ЗНАЧИМОСТИ УРАВНЕНИЯ РЕГРЕССИИ

Для парной регрессии правила(!) и (!!) эквиваленты и $F=T^2$,

Для парной регрессии значимость уравнения регрессии эквивалентна значимости коэффициента регрессии

Квантили Т-распределения Стьюдента



Excel: t_{α}=Стьюдент.Обр.2Х(α, число степеней свободы)

α - заданная вероятность ошибки 1 рода (уровень значимости), суммарная верояность двух хвостов