

# **Лекции по эконометрике № 4, 5**

**Классическая линейная регрессия.  
Проверка гипотез о конкретном  
значении коэффициентов регрессии.**

Демидова  
Ольга Анатольевна  
[https://www.hse.ru/staff/demidova\\_olga](https://www.hse.ru/staff/demidova_olga)  
E-mail: [demidova@hse.ru](mailto:demidova@hse.ru)  
23.09.2019, 30.09.2019

# План лекции № 4

- **Классическая линейная регрессия**
- **Проверка гипотез о конкретном значении коэффициентов парной регрессии**
- **Доверительные интервалы для коэффициентов парной регрессии**
- **Прогнозирование по модели парной регрессии**
- **Доверительные интервалы для среднего и индивидуального прогноза**
- **Проверка нормальности распределения**

# Классическая линейная регрессия

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$\varepsilon$  – сумма влияния многих факторов, каждый из которых незначительно влияет на  $Y$ . По Центральной предельной теореме такая случайная величина имеет нормальное распределение.

## Классическая линейная регрессия

Если  $\varepsilon_i$ ,  $i = 1, \dots, n$  распределены нормально,

т.е.  $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ ,

То оценки параметров  $\beta_0$  и  $\beta_1$  тоже распределены нормально, причем

$$\hat{\beta}_0 \sim N \left( \beta_0, \frac{\sum_{i=1}^n X_i^2}{n \sum_{i=1}^n x_i^2} \sigma_\varepsilon^2 \right)$$

$$\hat{\beta}_1 \sim N \left( \beta_1, \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n x_i^2} \right) \quad \text{где } x_i = X_i - \bar{X}, \\ i = 1, \dots, n$$

## Классическая линейная регрессия

Дисперсия возмущений  $\sigma_\varepsilon^2$  неизвестна, для нее используется оценка

$$\hat{\sigma}_\varepsilon^2 = \frac{RSS}{n - 2}$$

Случайная величина

$$\chi^2(n - 2)$$

$$\frac{RSS}{\hat{\sigma}_\varepsilon^2}$$

имеет распределение

## Классическая линейная регрессия

$$\hat{\beta}_0 \sim N(\beta_0, \sigma_{\hat{\beta}_0}^2)$$

$$\sigma_{\hat{\beta}_0}^2 = \frac{\sum_{i=1}^n X_i^2}{n \sum_{i=1}^n x_i^2} \sigma_{\varepsilon}^2$$

## Классическая линейная регрессия

$$\hat{\beta}_1 \sim N\left(\beta_1, \sigma_{\hat{\beta}_1}^2\right), \quad \sigma_{\hat{\beta}_1}^2 = \frac{\sigma_{\varepsilon}^2}{\sum_{i=1}^n x_i^2}$$

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \sim N(0,1), \quad \sigma_{\hat{\beta}_1}^2 = \frac{\sigma_{\varepsilon}^2}{\sum_{i=1}^n x_i^2}$$

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim ???, \quad \frac{\hat{\sigma}_{\hat{\beta}_1}^2}{\sigma_{\hat{\beta}_1}^2} = \frac{\hat{\sigma}_{\varepsilon}^2}{\sigma_{\varepsilon}^2}$$

## Классическая линейная регрессия

$$\hat{\sigma}_{\varepsilon}^2 = \frac{RSS}{n-2}, \quad \frac{RSS}{\sigma_{\varepsilon}^2} \sim \chi^2(n-2),$$

$$\frac{\hat{\sigma}_{\varepsilon}^2}{\sigma_{\varepsilon}^2} (n-2) \sim \chi^2(n-2), \quad \frac{\hat{\sigma}_{\hat{\beta}_1}^2}{\sigma_{\hat{\beta}_1}^2} = \frac{\hat{\sigma}_{\varepsilon}^2}{\sigma_{\varepsilon}^2}$$

$$\frac{\hat{\sigma}_{\hat{\beta}_1}^2}{\sigma_{\hat{\beta}_1}^2} (n-2) \sim \chi^2(n-2),$$



## Классическая линейная регрессия

$$t(k) \sim \frac{N(0,1)}{\sqrt{\chi^2(k)/k}} \qquad \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim ???,$$

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{\frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}}}{\sqrt{\frac{\frac{\sigma_{\hat{\beta}_1}^2}{\sigma_{\hat{\beta}_1}^2} (n-2)}{(n-2)}}} \sim t(n-2),$$

# Проверка гипотез

## Проверка гипотез состоит из

- Выбора основной и альтернативной гипотезы
- Вычисления некоторой тестовой статистики
- Выбора уровня значимости  $\alpha$  (числа между 0 и 1),  
Самые распространенные уровни значимости 0.05 и 0.01
- Разбиения множества значений тестовой статистики на две области: там, где основная гипотеза отвергается и там, где основная гипотеза не отвергается

## Проверка гипотез о конкретном значении коэффициентов регрессии при двусторонней альтернативной гипотезе

Модель:  $Y = \beta_0 + \beta_1 X + \varepsilon$

Нулевая гипотеза:  $H_0 : \beta_1 = \beta_1^0$

Альтернативная гипотеза:  $H_1 : \beta_1 \neq \beta_1^0$

## **Проверка гипотез о конкретном значении коэффициентов регрессии при двусторонней альтернативной гипотезе**

**Сначала необходимо оценить по  $n$  наблюдениям модель:**

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

**Если нулевая гипотеза не отвергается, то тестовая статистика**

$$t = \frac{\hat{\beta}_1 - \beta_1^0}{s.e.(\hat{\beta}_1)} \sim t(n-2)$$

**Имеет  $t$  – распределение с  $(n - 2)$  степенями свободы.**

# Таблицы для $t$ - распределения

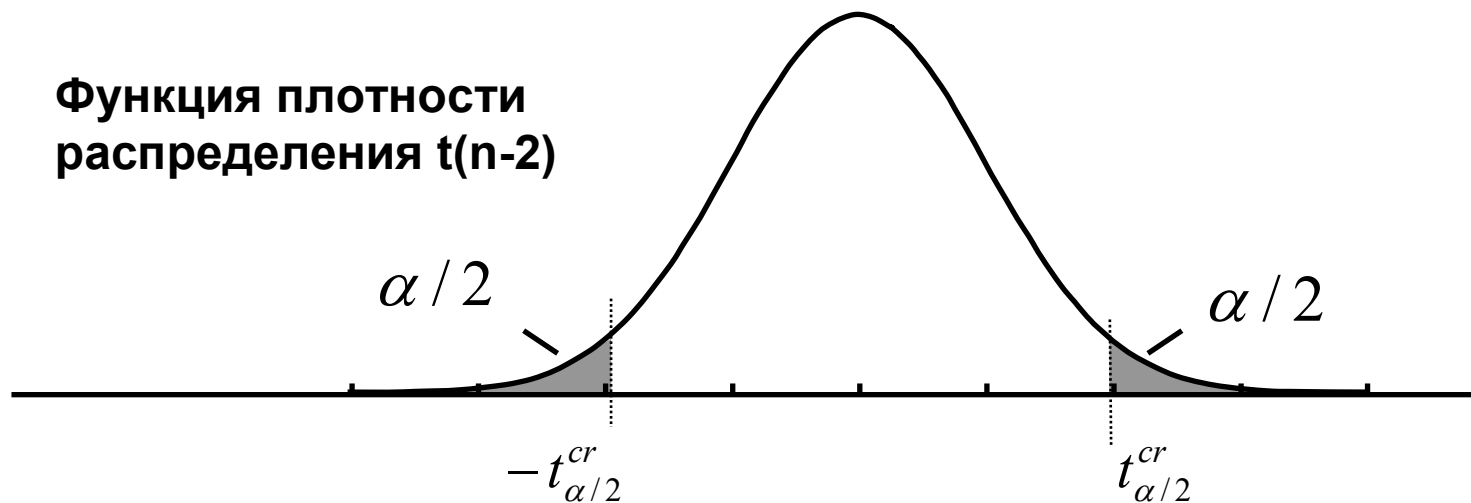
$t$  Distribution: Critical values of  $t$

Degrees of freedom	Two-tailed test One-tailed test	10% 5%	5% 2.5%	2% 1%	1% 0.5%	0.2% 0.1%	0.1% 0.05%
1		6.314	12.706	31.821	63.657	318.31	636.62
2		2.920	4.303	6.965	9.925	22.327	31.598
3		2.353	3.182	4.541	5.841	10.214	12.924
4		2.132	2.776	3.747	4.604	7.173	8.610
5		2.015	2.571	3.365	4.032	5.893	6.869
...		...	...	...	...	...	...
...		...	...	...	...	...	...
18		1.734	2.101	2.552	2.878	3.610	3.922
19		1.729	2.093	2.539	2.861	3.579	3.883
20		1.725	2.086	2.528	2.845	3.552	3.850
...		...	...	...	...	...	...
...		...	...	...	...	...	...
120		1.658	1.980	2.358	2.617	3.160	3.373
$\infty$		1.645	1.960	2.326	2.576	3.090	3.291

**Правило принятия решения при двусторонней альтернативной гипотезе и уровне значимости  $\alpha$ :**

**Нулевая гипотеза  $H_0 : \beta_1 = \beta_1^0$  отвергается**

**если  $|t| > t_{\alpha/2}^{cr}$**



**Серым цветом выделена область отвержения нулевой гипотезы при двусторонней альтернативной гипотезе.**

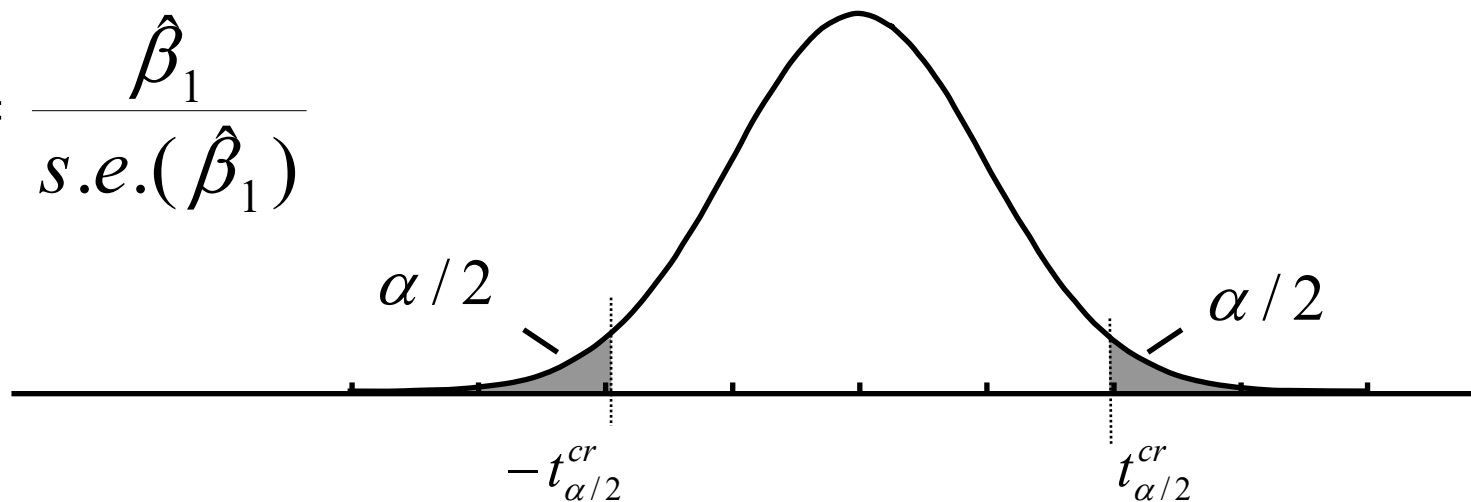
# Проверка гипотезы о значимости коэффициента

Модель  $Y = \beta_0 + \beta_1 X + \varepsilon$

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

$$t = \frac{\hat{\beta}_1}{s.e.(\hat{\beta}_1)}$$



Если нулевая гипотеза отвергается, то говорят, что коэффициент  $\beta_1$  значим. Если нулевая гипотеза не отвергается, то коэффициент  $\beta_1$  называется незначимым. Серым цветом выделена область отвержения нулевой гипотезы.

# Проверка гипотезы о значимости коэффициента.

## t - статистика

Модель:  $Y = \beta_0 + \beta_1 X + \varepsilon$

```
. reg EARNINGS S
```

Source	SS	df	MS	Number of obs = 570		
Model	3977.38016	1	3977.38016	F( 1, 568)	=	65.64
Residual	34419.6569	568	60.5979875	Prob > F	=	0.0000
Total	38397.0371	569	67.4816117	R-squared	=	0.1036
				Adj R-squared	=	0.1020
				Root MSE	=	7.7845

EARNINGS	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
S	1.073055	.1324501	8.102	0.000	.8129028	1.333206
_cons	-1.391004	1.820305	-0.764	0.445	-4.966354	2.184347

t – статистика коэффициента наклона выделена красным цветом.



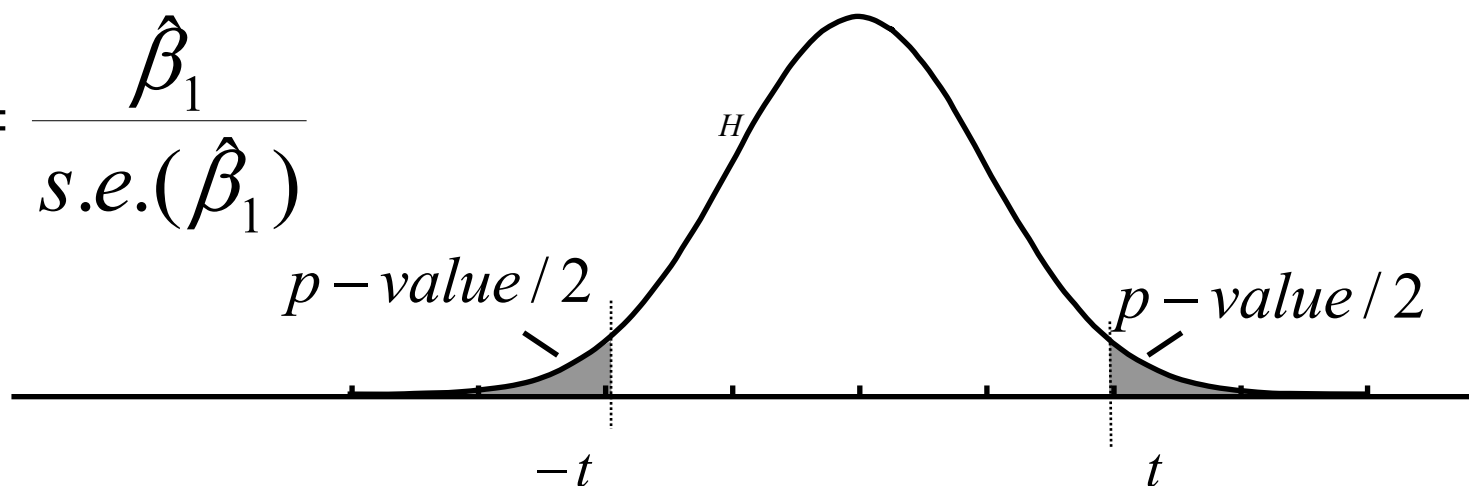
## P – VALUE (P – Значение) для проверки гипотезы о значимости коэффициента

Модель  $Y = \beta_0 + \beta_1 X + \varepsilon$

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

$$t = \frac{\hat{\beta}_1}{s.e.(\hat{\beta}_1)}$$



**P – value** – минимальный уровень значимости, при котором нулевая гипотеза отвергается. На рисунке это площадь всей заштрихованной области.

# Проверка гипотезы о значимости коэффициента.

## P-value

```
. reg EARNINGS S
```

Source	SS	df	MS	Number of obs = 570			
Model	3977.38016	1	3977.38016	F( 1, 568)	=	65.64	
Residual	34419.6569	568	60.5979875	Prob > F	=	0.0000	
Total	38397.0371	569	67.4816117	R-squared	=	0.1036	
				Adj R-squared	=	0.1020	
				Root MSE	=	7.7845	

EARNINGS	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
S	1.073055	.1324501	8.102	0.000	.8129028	1.333206
_cons	-1.391004	1.820305	-0.764	0.445	-4.966354	2.184347

В таблице выделены P-value для проверки гипотез о значимости коэффициентов регрессии.

# Проверка гипотезы о значимости коэффициента. Связь P-value и уровня значимости $\alpha$ .

```
. reg EARNINGS S
```

Source	SS	df	MS	Number of obs = 570		
Model	3977.38016	1	3977.38016	F( 1, 568)	=	65.64
Residual	34419.6569	568	60.5979875	Prob > F	=	0.0000
Total	38397.0371	569	67.4816117	R-squared	=	0.1036
				Adj R-squared	=	0.1020
				Root MSE	=	7.7845

EARNINGS	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
S	1.073055	.1324501	8.102	0.000	.8129028	1.333206
_cons	-1.391004	1.820305	-0.764	0.445	-4.966354	2.184347

Если P-value коэффициента регрессии меньше, чем выбранный уровень значимости  $\alpha$ , то нулевая гипотеза отвергается и соответствующий коэффициент является значимым. В приведенном примере при любом разумном уровне значимости константа незначима, а коэффициент наклона значим.

**Проверка гипотез о конкретном значении  
коэффициентов регрессии при односторонней  
альтернативной гипотезе (>)**

**Модель:** 
$$Y = \beta_0 + \beta_1 X + \varepsilon$$

**Основная гипотеза:** 
$$H_0 : \beta_1 = \beta_1^0$$

**Альтернативная гипотеза:** 
$$H_1 : \beta_1 > \beta_1^0$$

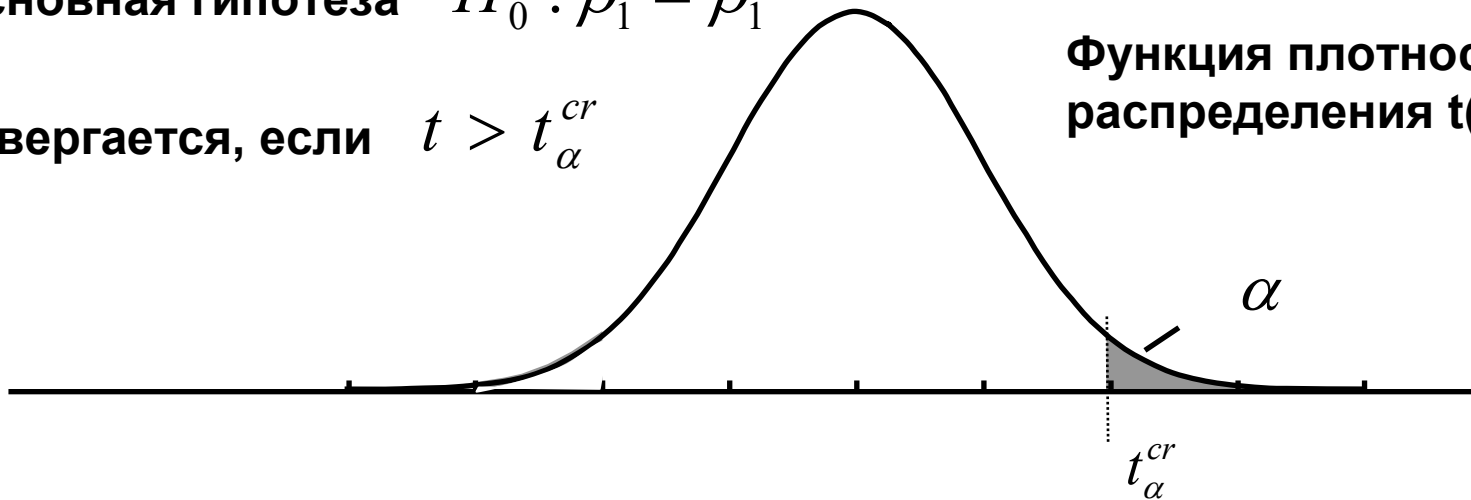
# Проверка гипотез о конкретном значении коэффициента регрессии при односторонней альтернативной гипотезе ( $>$ )

Правило отвержения нулевой гипотезы при  
односторонней альтернативной гипотезе ( $>$ )  
и уровне значимости  $\alpha$ .

Основная гипотеза  $H_0 : \beta_1 = \beta_1^0$

отвергается, если  $t > t_{\alpha}^{cr}$

Функция плотности  
распределения  $t(n-2)$



Серым цветом выделена область отвержения нулевой гипотезы  
при односторонней альтернативной гипотезе ( $>$ )

**Проверка гипотез о конкретном значении  
коэффициента регрессии при односторонней  
альтернативной гипотезе (<)**

**Модель:**

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

**Основная гипотеза:**

$$H_0 : \beta_1 = \beta_1^0$$

**Альтернативная гипотеза:**

$$H_1 : \beta_1 < \beta_1^0$$

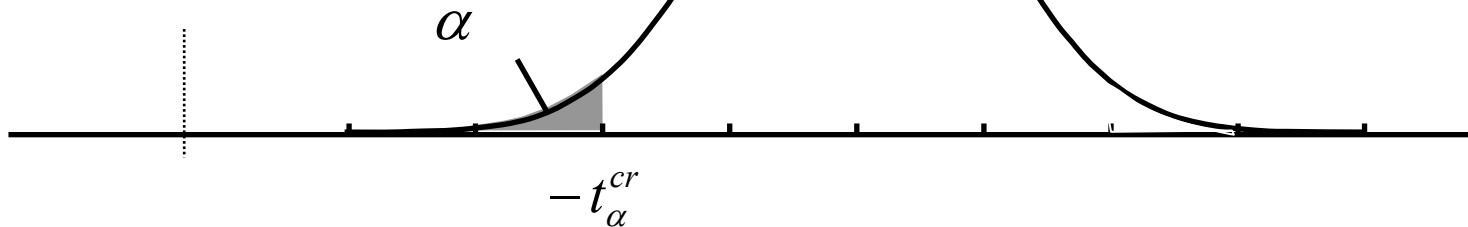
# Проверка гипотез о конкретном значении коэффициента регрессии при односторонней альтернативной гипотезе ( $<$ ).

Правило отвержения нулевой гипотезы при односторонней альтернативной гипотезе ( $<$ ) и уровне значимости  $\alpha$ .

Основная гипотеза  $H_0 : \beta_1 = \beta_1^0$

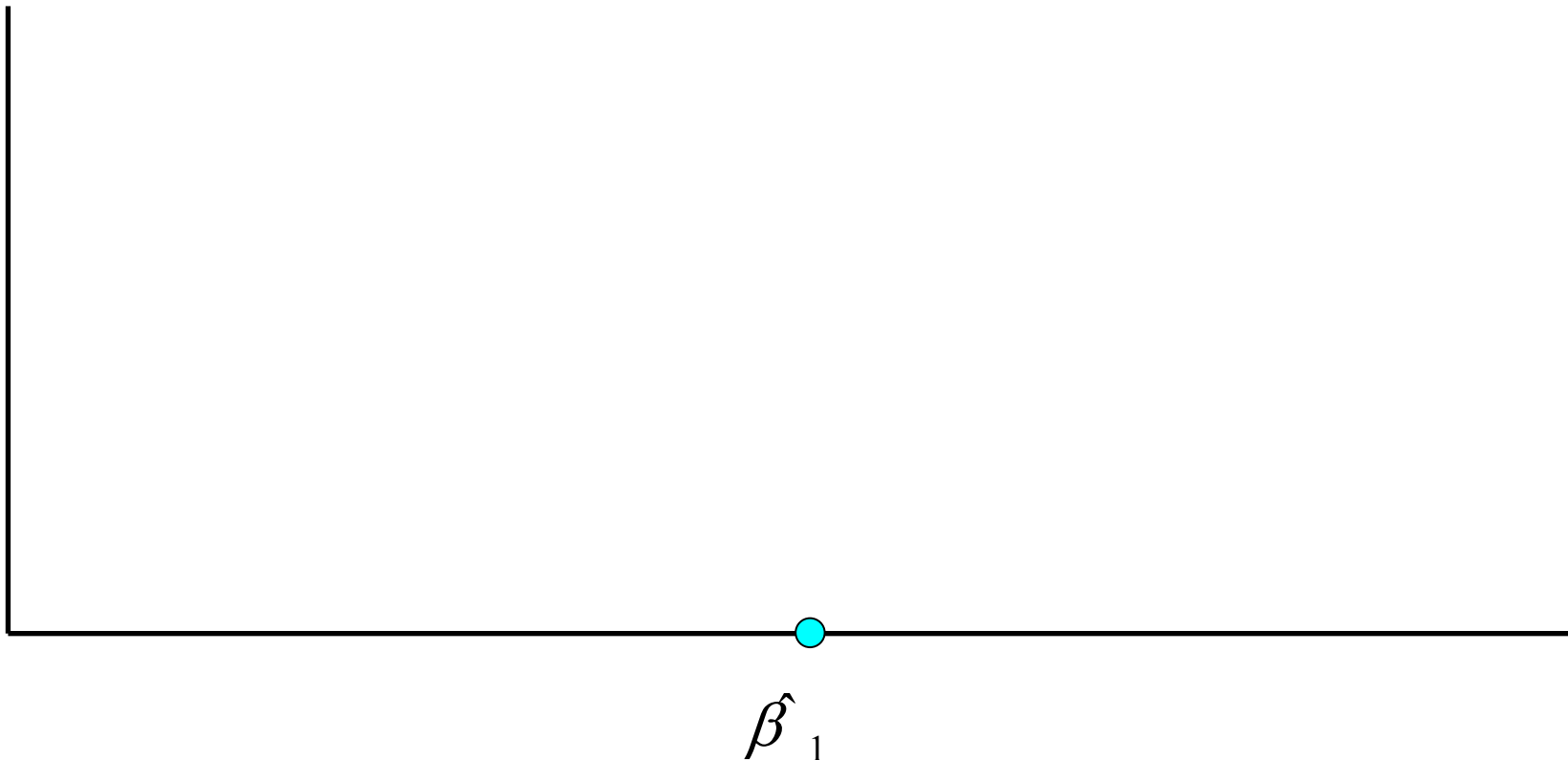
Функция плотности распределения  $t(n-2)$

отвергается, если  $t < t_{\alpha}^{cr}$



Серым цветом выделена область отвержения нулевой гипотезы при односторонней альтернативной гипотезе ( $<$ ).

## Доверительные интервалы для оценок коэффициентов регрессии



Найдем множество всех значений параметра  $\beta_1$ , гипотеза о равенстве которым при заданном уровне значимости  $\alpha$  и двусторонней альтернативной гипотезе не отвергается.



## Доверительные интервалы для оценок коэффициентов регрессии

Гипотеза  $H_0: \beta_1 = \beta_1^0$  не отвергается, если  $|t| \leq t_{\alpha/2}^{cr}$ , где

$$t = \frac{\hat{\beta}_1 - \beta_1^0}{s.e.(\hat{\beta}_1)}$$

$$\text{т.е.} \quad \left| \frac{\hat{\beta}_1 - \beta_1^0}{s.e.(\hat{\beta}_1)} \right| \leq t_{\alpha/2}^{cr} \quad \text{или} \quad \left| \hat{\beta}_1 - \beta_1^0 \right| \leq t_{\alpha/2}^{cr} \cdot s.e.(\hat{\beta}_1)$$

откуда

$$\hat{\beta}_1 - t_{\alpha/2}^{cr} s.e.(\hat{\beta}_1) \leq \beta_1^0 \leq \hat{\beta}_1 + t_{\alpha/2}^{cr} s.e.(\hat{\beta}_1)$$

## Доверительные интервалы для оценок коэффициентов регрессии

(1- $\alpha$ )100% доверительный интервал для  
коэффициента наклона  $\beta_1$  имеет вид:

$$\hat{\beta}_1 - t_{\alpha/2}^{cr} s.e.(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2}^{cr} s.e.(\hat{\beta}_1)$$

# Доверительные интервалы для оценок коэффициентов регрессии

Модель:  $Y = \beta_0 + \beta_1 X + u$

. reg EARNINGS S

Source	SS	df	MS	Number of obs = 570		
Model	3977.38016	1	3977.38016	F( 1, 568)	=	65.64
Residual	34419.6569	568	60.5979875	Prob > F	=	0.0000
Total	38397.0371	569	67.4816117	R-squared	=	0.1036
				Adj R-squared	=	0.1020
				Root MSE	=	7.7845

EARNINGS	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
S	1.073055	.1324501	8.102	0.000	.8129028	1.333206
_cons	-1.391004	1.820305	-0.764	0.445	-4.966354	2.184347

В последней колонке – 95% доверительные интервалы для коэффициентов регрессии.

# Проверка значимости коэффициентов с помощью доверительных интервалов

Модель:  $Y = \beta_0 + \beta_1 X + u$

```
. reg EARNINGS S
```

Source	SS	df	MS	Number of obs = 570		
Model	3977.38016	1	3977.38016	F( 1, 568)	=	65.64
Residual	34419.6569	568	60.5979875	Prob > F	=	0.0000
Total	38397.0371	569	67.4816117	R-squared	=	0.1036
				Adj R-squared	=	0.1020
				Root MSE	=	7.7845

EARNINGS	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
S	1.073055	.1324501	8.102	0.000	.8129028	1.333206
_cons	-1.391004	1.820305	-0.764	0.445	-4.966354	2.184347

Если 0 принадлежит доверительному интервалу для коэффициента, то этот коэффициент является незначимым.

В приведенном примере коэффициент  $\beta_0$  незначим, а коэффициент  $\beta_1$  - значим.

## Прогнозирование по модели парной регрессии

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \dots, n$$

Прогноз для  $X_{n+1}$  – ?

$$Y_{n+1} = \beta_0 + \beta_1 X_{n+1} + \varepsilon_{n+1}$$

$$\hat{Y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 X_{n+1}$$

Ошибка индивидуального прогноза

$$e_{n+1} = Y_{n+1} - \hat{Y}_{n+1} = (\beta_0 - \hat{\beta}_0) + X_{n+1}(\beta_1 - \hat{\beta}_1) + \varepsilon_{n+1}$$

$$\text{var}(e_{n+1}) = \text{var}(\hat{\beta}_0) + X_{n+1}^2 \text{var}(\hat{\beta}_1) +$$

$$+ 2 X_{n+1} \text{cov}(\hat{\beta}_0, \hat{\beta}_1) + \text{var}(\varepsilon_{n+1}) =$$

$$= \sigma_{\varepsilon}^2 \left[ \left( \frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n x_i^2} \right) + \frac{X_{n+1}^2}{\sum_{i=1}^n x_i^2} - 2 \frac{X_{n+1} \cdot \bar{X}}{\sum_{i=1}^n x_i^2} + 1 \right]$$

## Прогнозирование по модели парной регрессии

$$\begin{aligned}\text{var}(e_{n+1}) &= \sigma_{\varepsilon}^2 \left[ \left( \frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n x_i^2} \right) + \frac{X_{n+1}^2}{\sum_{i=1}^n x_i^2} - 2 \frac{X_{n+1} \cdot \bar{X}}{\sum_{i=1}^n x_i^2} + 1 \right] \\ &= \sigma_{\varepsilon}^2 \left[ 1 + \frac{1}{n} + \frac{(X_{n+1} - \bar{X})^2}{\sum_{i=1}^n x_i^2} \right]\end{aligned}$$

## Прогнозирование по модели парной регрессии

$$\frac{Y_{n+1} - \hat{Y}_{n+1}}{\sqrt{\sigma_{\varepsilon}^2 \left[ 1 + \frac{1}{n} + \frac{(X_{n+1} - \bar{X})^2}{\sum_{i=1}^n x_i^2} \right]}} \sim N(0,1)$$

$$\frac{Y_{n+1} - \hat{Y}_{n+1}}{\sqrt{\hat{\sigma}_{\varepsilon}^2 \left[ 1 + \frac{1}{n} + \frac{(X_{n+1} - \bar{X})^2}{\sum_{i=1}^n x_i^2} \right]}} \sim t(n-2)$$

# Прогнозирование по модели парной регрессии

Доверительный интервал для индивидуального прогноза

$$\hat{\beta}_0 + \hat{\beta}_1 X_{n+1} \pm t_{\alpha/2} \sqrt{\hat{\sigma}_\varepsilon^2 \left[ 1 + \frac{1}{n} + \frac{(X_{n+1} - \bar{X})^2}{\sum_{i=1}^n x_i^2} \right]},$$

$$\hat{\sigma}_\varepsilon^2 = \frac{RSS}{n-2}$$



## Прогнозирование по модели парной регрессии

$Y_{n+1} = \beta_0 + \beta_1 X_{n+1} + \varepsilon_{n+1}$  – индивидуальный прогноз

$$E(Y_{n+1}) = \beta_0 + \beta_1 X_{n+1}$$

$\hat{Y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 X_{n+1}$  – "средний" прогноз

Ошибка "среднего" прогноза

$$\tilde{\varepsilon}_{n+1} = E(Y_{n+1}) - \hat{Y}_{n+1} = (\beta_0 - \hat{\beta}_0) + X_{n+1}(\beta_1 - \hat{\beta}_1) =$$

$$\text{var}(\tilde{\varepsilon}_{n+1}) = \text{var}(\hat{\beta}_0) + X_{n+1}^2 \text{var}(\hat{\beta}_1) +$$

$$+ 2 X_{n+1} \text{cov}(\hat{\beta}_0, \hat{\beta}_1) =$$

$$= \sigma_{\varepsilon}^2 \left[ \left( \frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n x_i^2} \right) + \frac{X_{n+1}^2}{\sum_{i=1}^n x_i^2} - 2 \frac{X_{n+1} \cdot \bar{X}}{\sum_{i=1}^n x_i^2} \right]$$

# Прогнозирование по модели парной регрессии

Доверительный интервал для "среднего" прогноза

$$\hat{\beta}_0 + \hat{\beta}_1 X_{n+1} \pm t_{\alpha/2} \sqrt{\hat{\sigma}_\varepsilon^2 \left[ \frac{1}{n} + \frac{(X_{n+1} - \bar{X})^2}{\sum_{i=1}^n x_i^2} \right]},$$

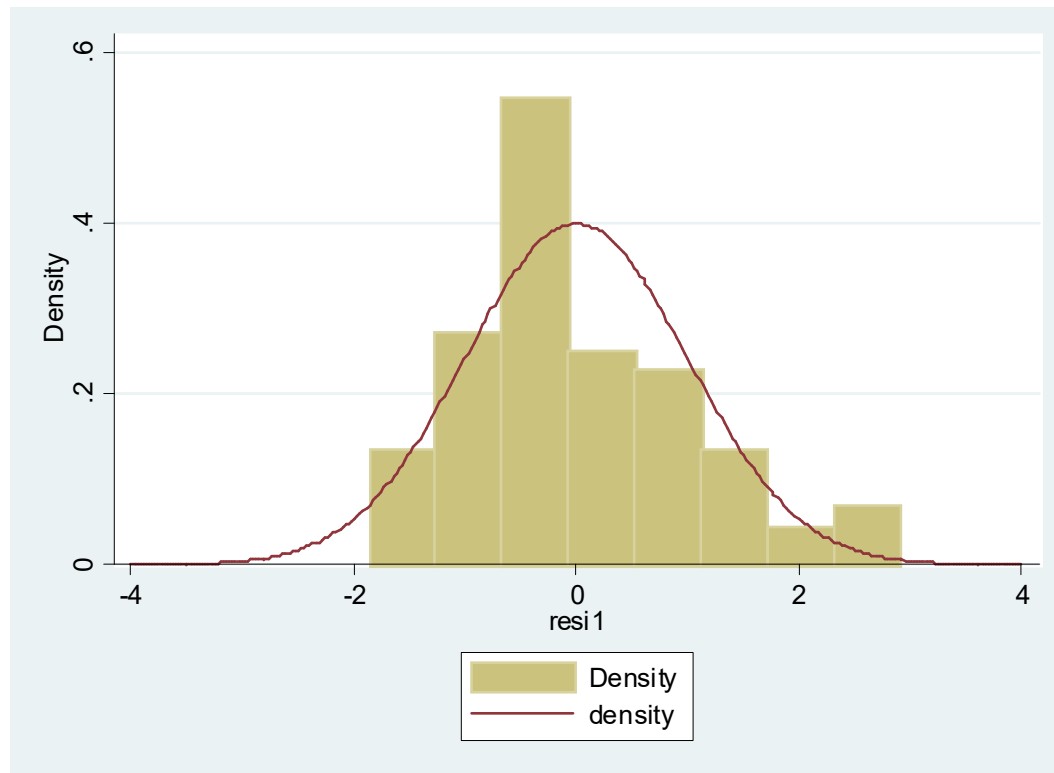
$$\hat{\sigma}_\varepsilon^2 = \frac{RSS}{n-2}$$

# **Тестирование регрессионных остатков на нормальность распределения**

# Проверка нормальности распределения остатков

## Визуальный анализ

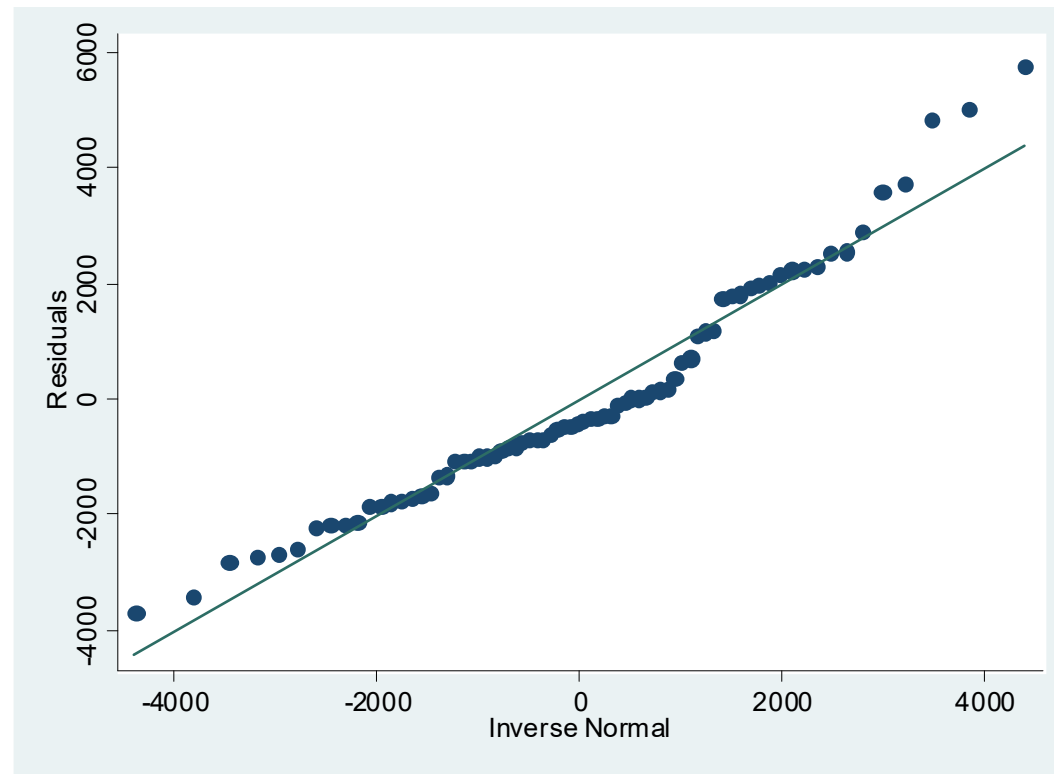
- Сравнение гистограммы остатков с гистограммой нормального распределения



# Проверка нормальности распределения остатков

Визуальный анализ

Q-Q plot (Q-norm plot)



# Проверка нормальности распределения остатков

## Тест Jarque-Bera

$$H_0 : e_i \sim N(., .)$$

$$H_1 : e_i \not\sim N(.,.)$$

$$JB = \frac{n}{6} \left( sk^2 + \frac{1}{4} (k - 3)^2 \right) \sim \chi^2(2)$$

**Sk – skewness, k – kurtosis (нормированные третий и четвертый центральные моменты)**

$$sk = \frac{1}{n} \sum_{i=1}^n \frac{(X_i - \bar{X})^3}{\sigma^3}; \quad \sigma^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2;$$

$$k = \frac{1}{n} \sum_{i=1}^n \frac{(X_i - \bar{X})^4}{\sigma^4}$$

# Проверка нормальности распределения остатков

**Недостаток: Тест Jarque-Bera применим только при большом числе наблюдений, при малом следует использовать тест Шапиро – Уилка.**

**Весьма популярным является тест Колмогорова-Смирнова проверки нормальности.**