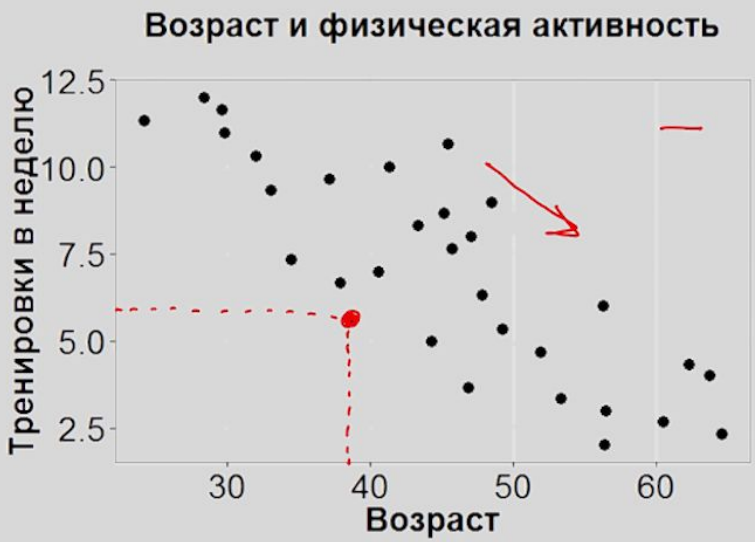
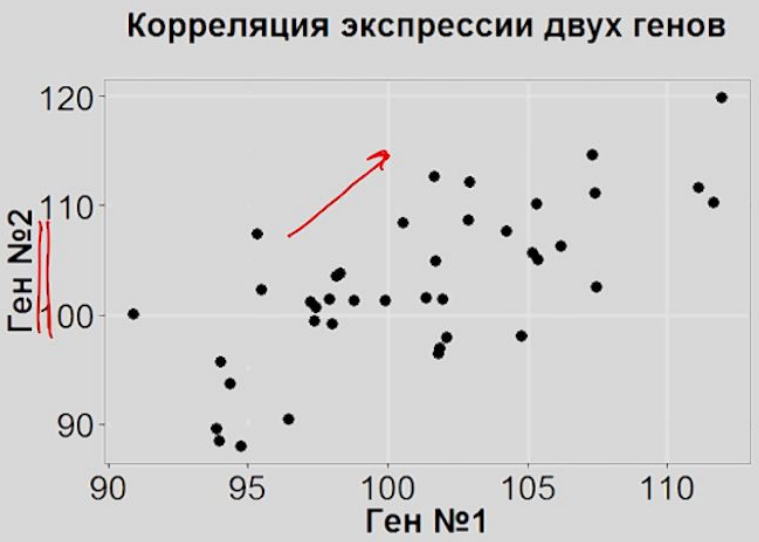
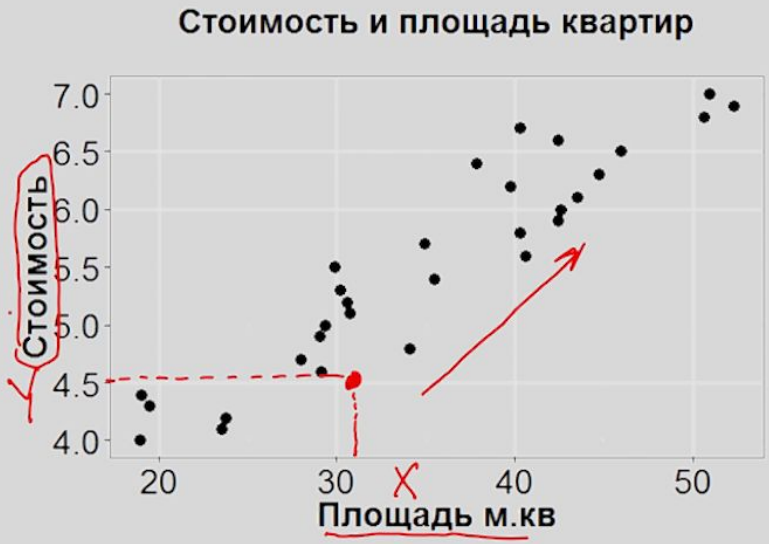
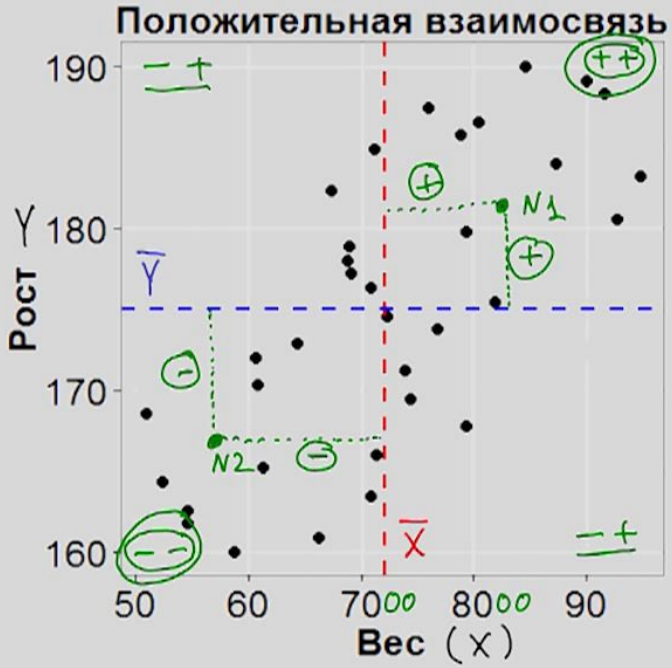

Возраст и физическая активность
Тренировки в неделю
12.5
10.0
7.5
5.0
2.5
30
40
50
60
Возраст


Корреляция экспрессии двух генов
Ген №2
120
110
100
90
90
95
100
105
110
Ген №1


Стоимость и площадь квартир
Стоимость
7.0
6.5
6.0
5.5
5.0
4.5
4.0
20
30
40
50
Площадь м.кв



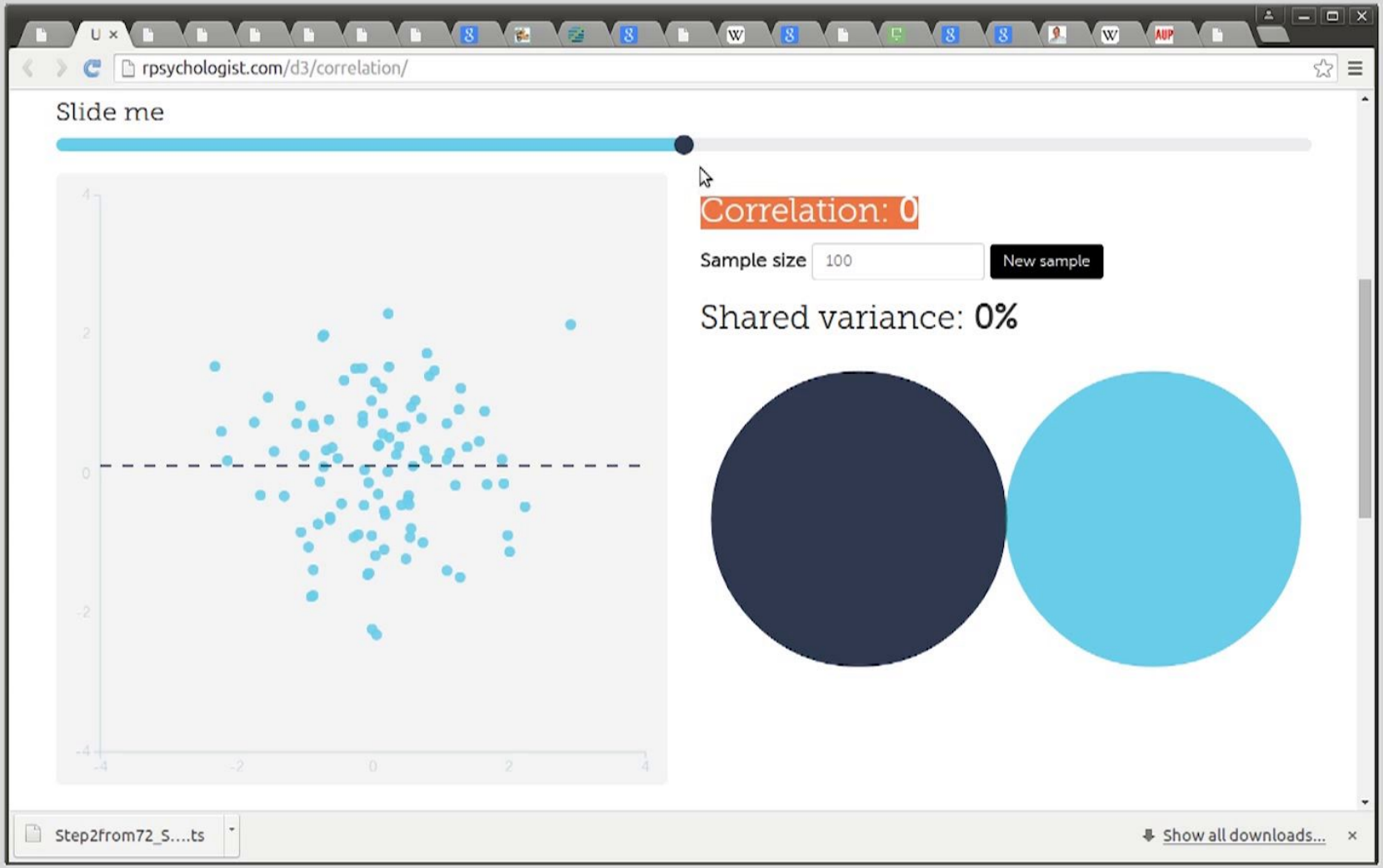
$$+ \begin{cases} (x_1 - \bar{x}) \cdot (y_1 - \bar{y}) > 0 \\ (x_2 - \bar{x}) \cdot (y_1 - \bar{y}) > 0 \\ \vdots \\ (x_i - \bar{x}) \cdot (y_i - \bar{y}) \end{cases}$$

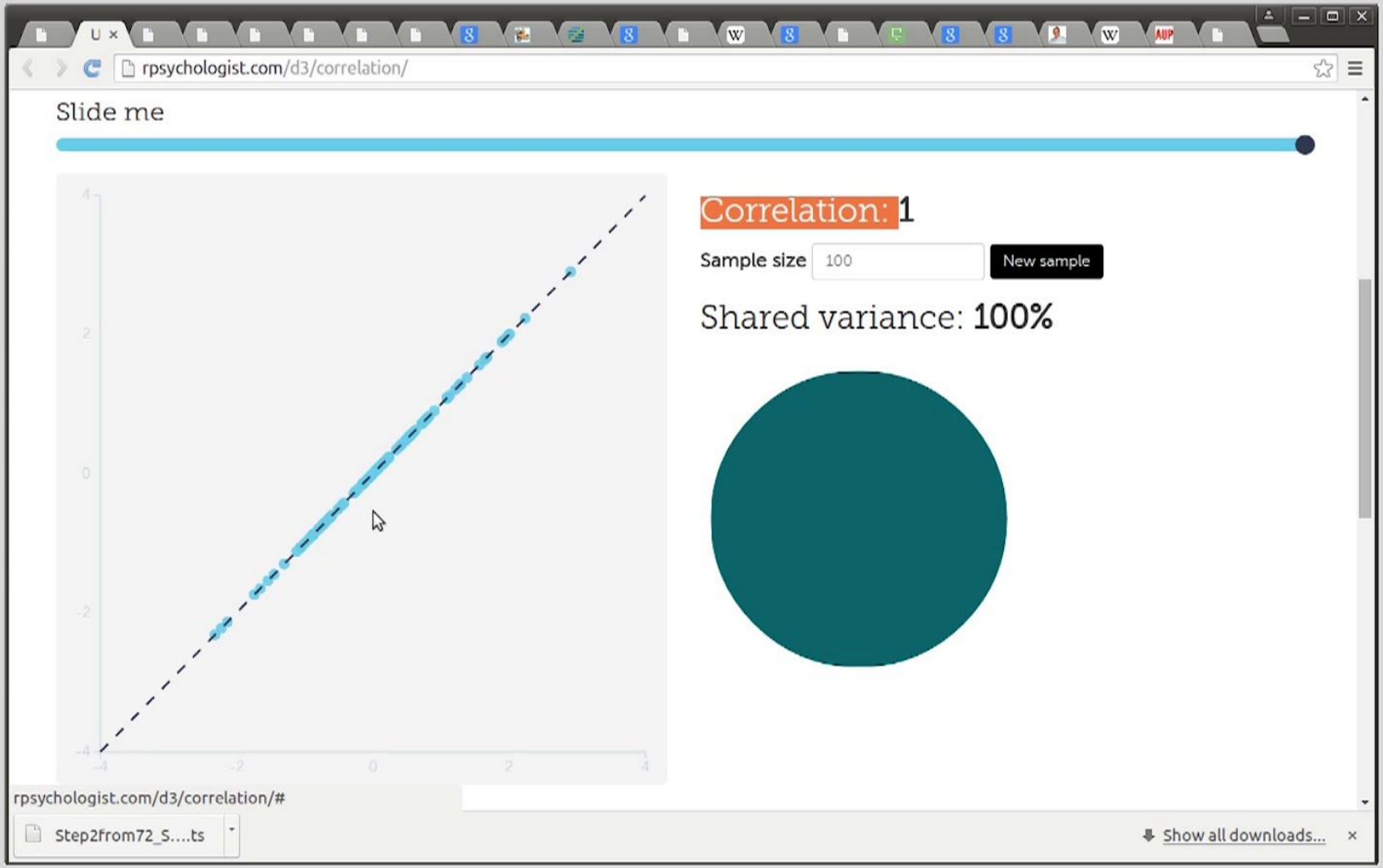
$$cov = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{N - 1} \quad \begin{matrix} + \\ - \end{matrix}$$

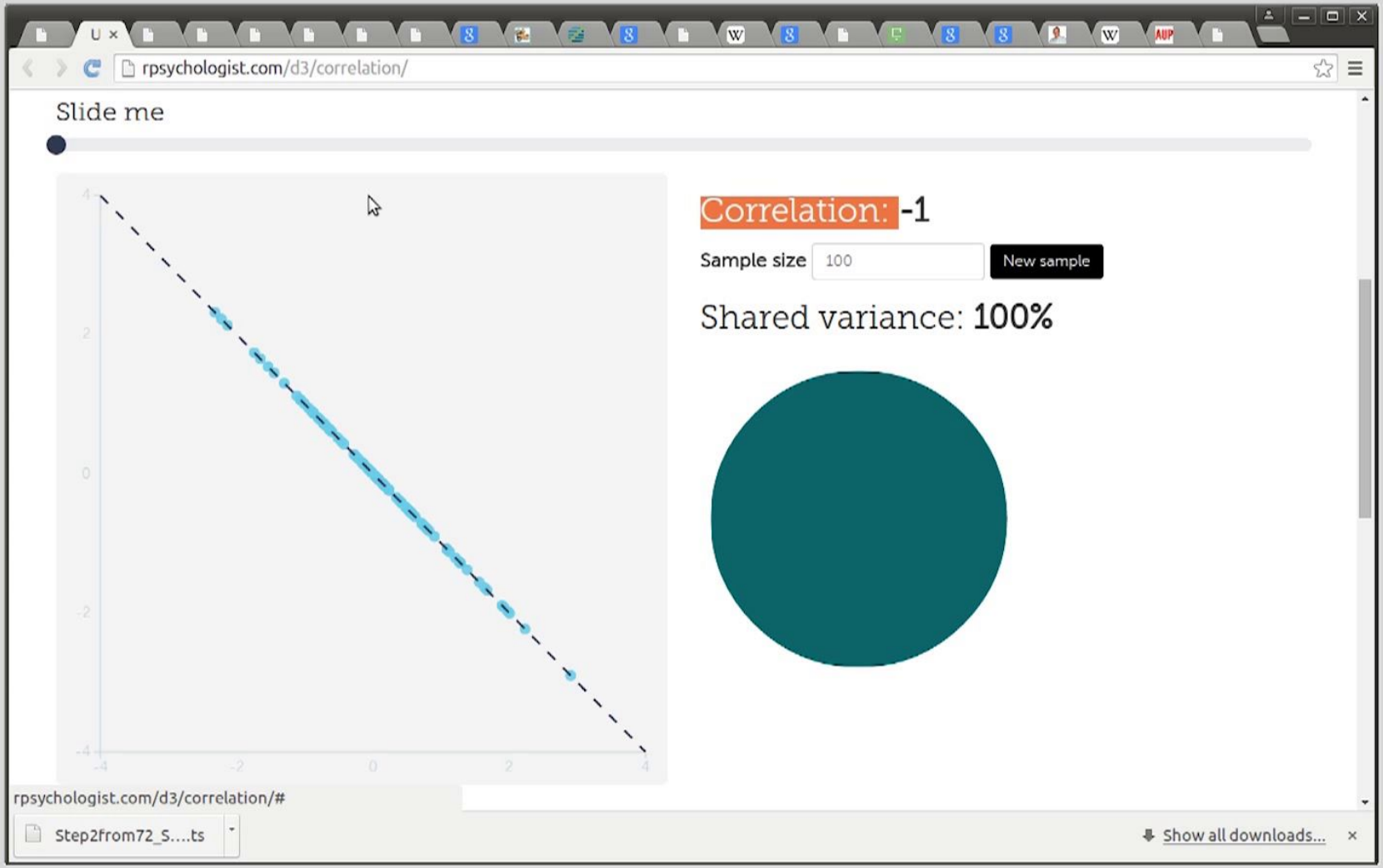
7200

$$r_{xy} = \frac{cov}{\sigma_x \cdot \sigma_y} \qquad [-1 \; ; \; 1]$$

"-" "+"


rpsychologist.com/d3/correlation/
Slide me
Correlation: 0
Sample size
100
New sample
Shared variance: 0%
Step2from72_S....ts
Show all downloads...


rpsychologist.com/d3/correlation/
Slide me
Correlation: 1
Sample size 100
New sample
Shared variance: 100%
rpsychologist.com/d3/correlation/#
Step2from72_S....ts
Show all downloads...


rpsychologist.com/d3/correlation/
Slide me
Correlation: -1
Sample size 100
New sample
Shared variance: 100%
rpsychologist.com/d3/correlation/#
Step2from72_S....ts
Show all downloads...

Коэффициент корреляции

$\mathbf{r_{xy}}$ — показатель силы и направления взаимосвязи двух количественных переменных

Принимает значения [-1, 1]

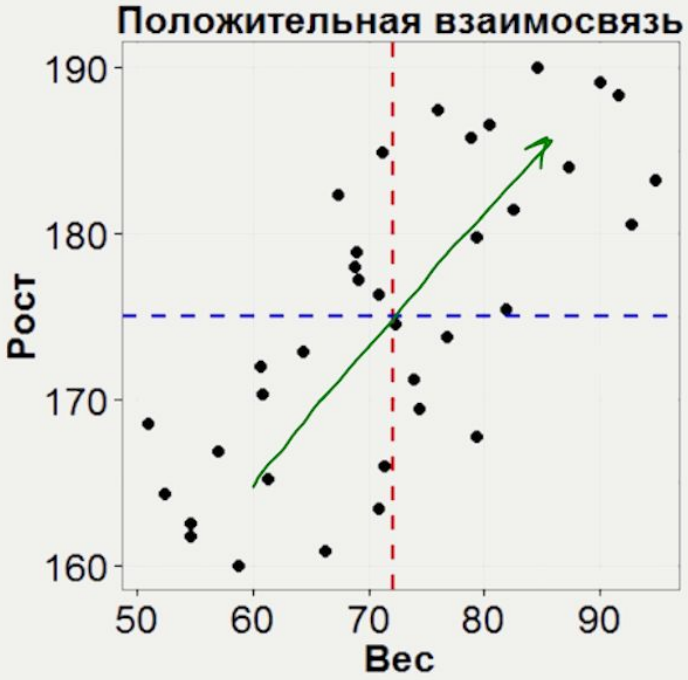
Знак коэффициента корреляции показывает направление взаимосвязи

Коэффициент детерминации

R^2 — показывает, в какой степени дисперсия одной переменной обусловлена влиянием другой переменной

Равен квадрату коэффициента корреляции

Принимает значения [0, 1]



$H_0 \quad r_{XY} = 0$

$H_1 \quad r_{XY} \neq 0$

$t \quad df = N - 2 \quad p$

RStudio

File Edit Code View Plots Session Build Debug Tools Help

Go to file/function

Project: (None)

r_plot.R × Untitled1* ×

Source on Save

Run Source

```
1  N=50
2
3  plot(x,y, lwd=5)
4
5  cor.test(x,y)
6
```

6:1 (Top Level) R Script

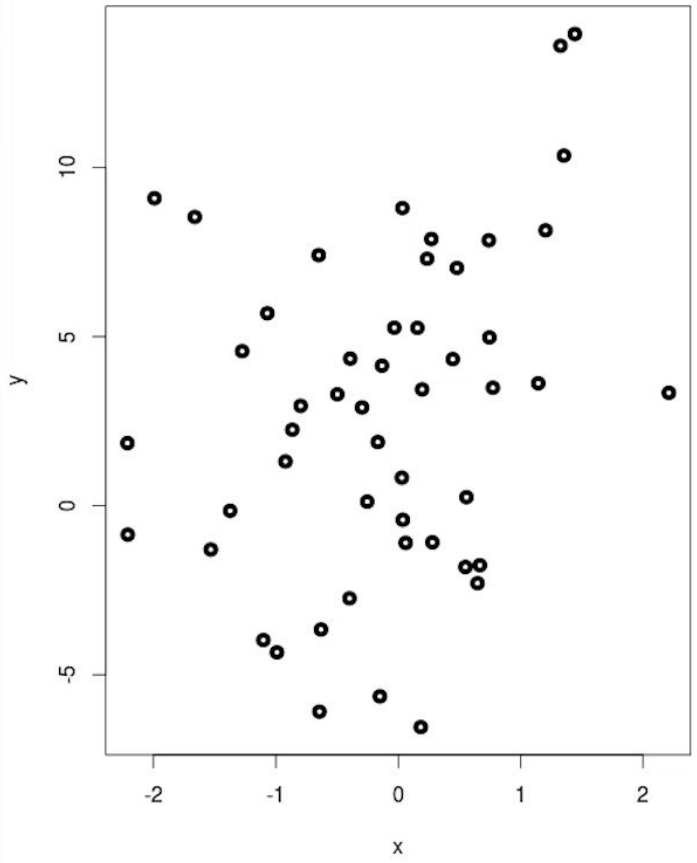
Console ~/

```
> cor.test(x,y)

 Pearson's product-moment correlation

data:  x and y
t = 2.067, df = 48, p-value = 0.04415
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.008193015 0.522646138
sample estimates:
      cor
0.2858888

>
```

Environment History

Files Plots Packages Help Viewer

Zoom Export Clear All



RStudio

File Edit Code View Plots Session Build Debug Tools Help

Go to file/function

Project: (None)

r_plot.R × Untitled1* ×

Source on Save

Run Source

```
1  N=50
2
3  plot(x,y, lwd=5)
4
5  cor.test(x,y)
6
```

6:1 (Top Level) R Script

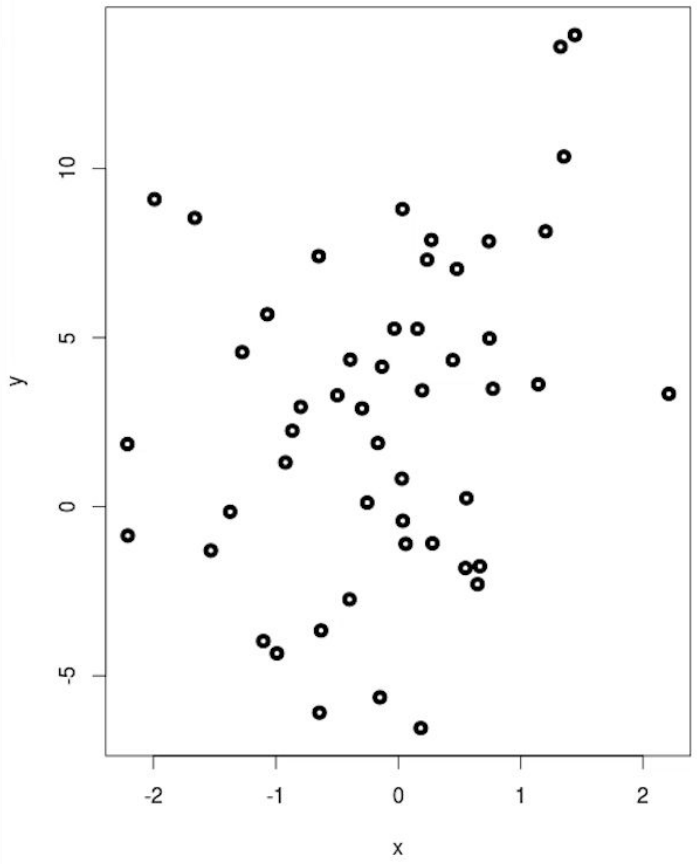
Console ~/

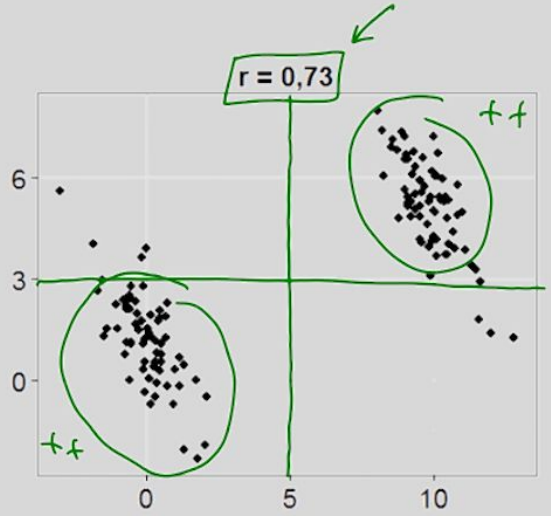
```
> cor.test(x,y)

 Pearson's product-moment correlation

data:  x and y
t = 2.067, df = 48, p-value = 0.04415
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.008193015 0.522646138
sample estimates:
      cor
0.2858888

>
```

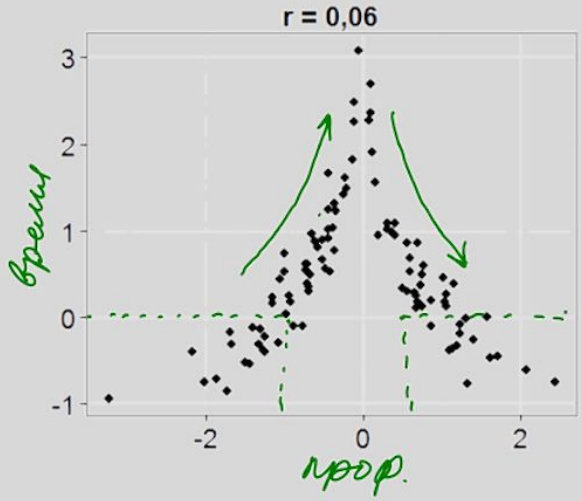
Environment History

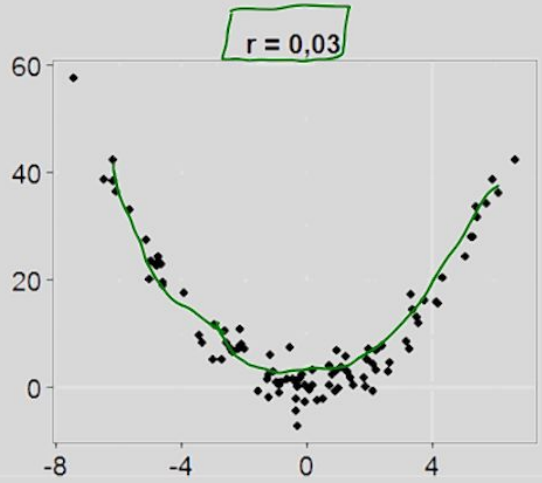
Files Plots Packages Help Viewer

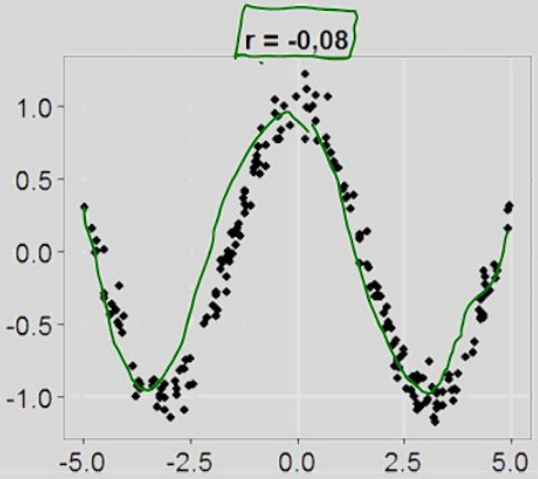
Zoom Export Clear All



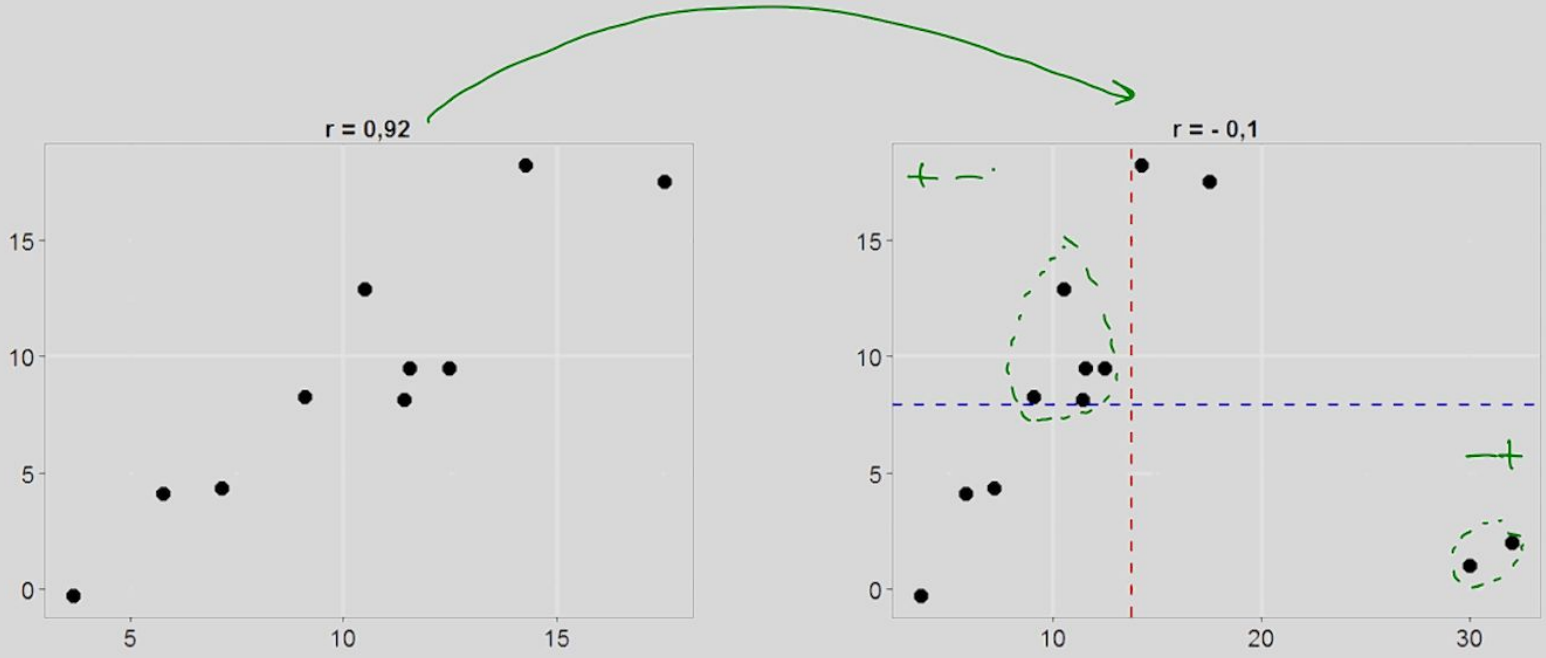

r = 0,73
++
++
6
3
0
0
5
10


r = 0,06
3
2
1
0
-1
-2
0
2

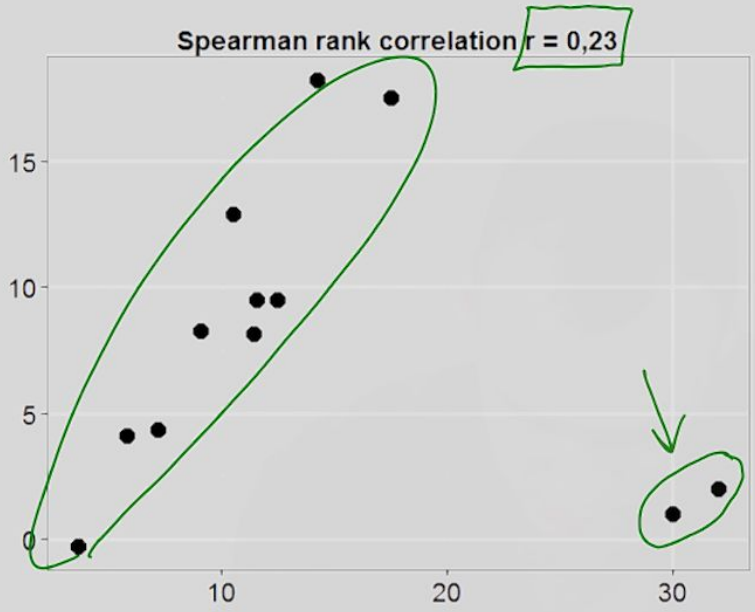

r = 0,03
60
40
20
0
-8
-4
0
4


r = -0,08
1.0
0.5
0.0
-0.5
-1.0
-5.0
-2.5
0.0
2.5
5.0

Опасность выбросов!!!



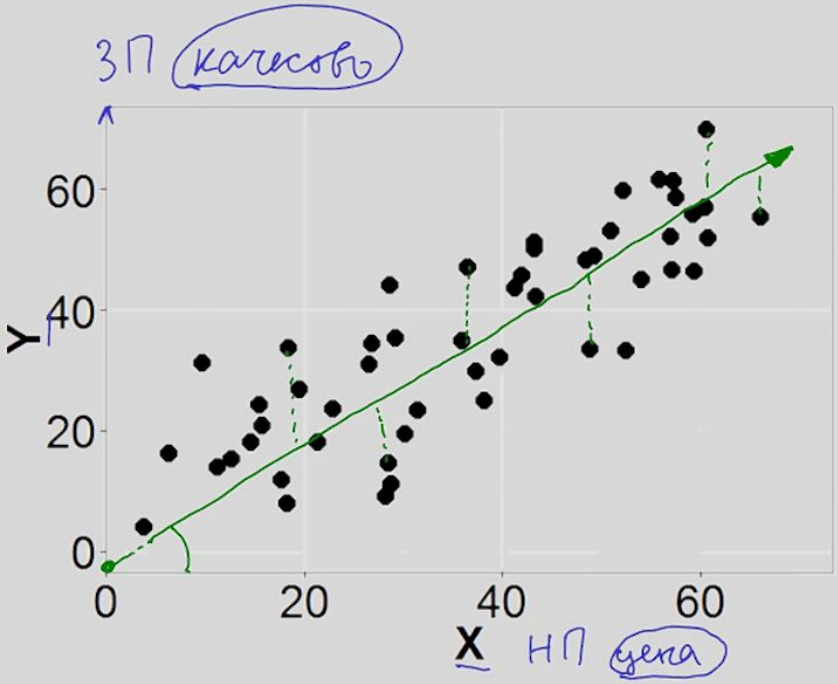
Коэффициент корреляции Спирмена



$$r_s = 1 - \frac{6\sum_i d_i^2}{N(N^2-1)}$$

X	Y		X	Y	d^2
3,7	-0,3		1	1	0
5,8	4,1		2	4	4
7,1	4,3		3	5	4
9,1	8,3		4	7	9
10,5	12,9		5	10	25
11,4	8,1		6	6	0
11,6	9,5		7	9	4
12,5	9,5		8	8	0
14,3	18,2		9	12	9
17,5	17,5		10	11	1
30,0	1,0		11	2	81
32,0	2,0		12	3	81

Линия регрессии



Statistic	N	Mean	St. Dev.
x	51	36.7	17.3
y	51	35.9	16.9

$y = b_0 + b_1 x$

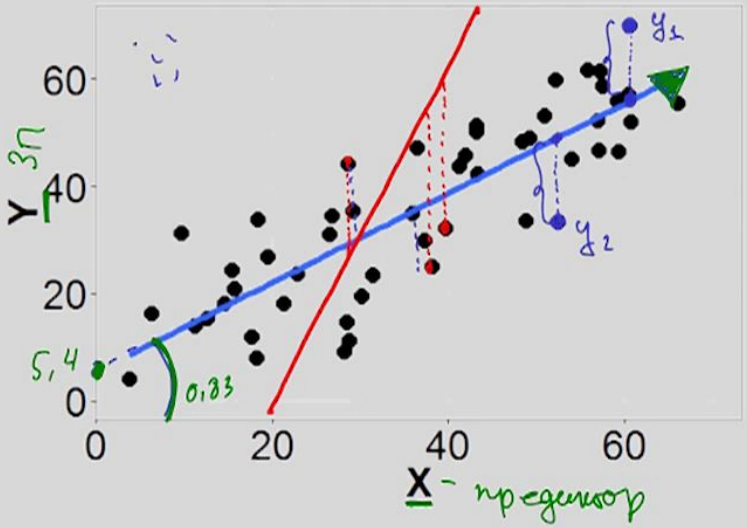
intercept slope

Метод наименьших квадратов

b_0 b_1

МНК – метод нахождения оптимальных параметров линейной регресии, таких, что сумма квадратов ошибок (остатков) была минимальна

$\hat{y} = 5{,}4 + 0{,}83 \cdot x$



$e_1 = y_1 - \hat{y}_1$ $\quad \sum e_i^2$

$e_2 = y_2 - \hat{y}_2$

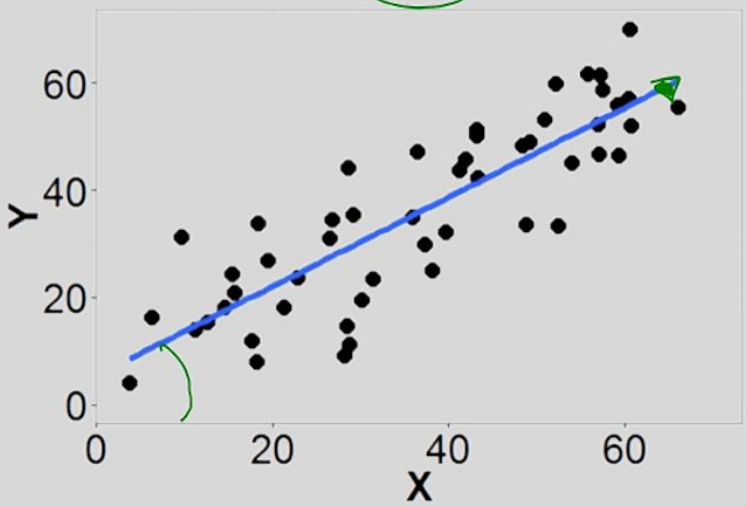
$$b_1 = \frac{sd_y}{sd_x} \cdot r_{xy} \qquad b_1 = 0{,}83$$

$$b_0 = \bar{Y} - b_1 \cdot \bar{X} \qquad b_0 = 5{,}4$$

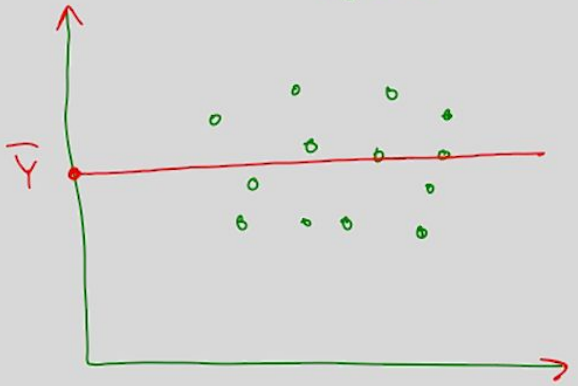
Коэффициенты линейной регрессии

b_1

$\hat{y} = 5.4 + 0.83 x$



$r_{xy} = 0$

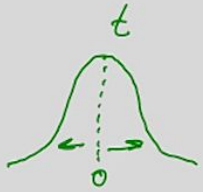


$b_1 = \frac{sd_y}{sd_x} \cdot r_{xy} = 0$

$b_0 = \bar{Y} - b_1 \cdot \bar{X} = \bar{Y}$

$H_0 \quad \beta_1 = 0$

$H_1 \quad \beta_1 \neq 0$



$$t = \frac{b_1 - 0}{se}$$

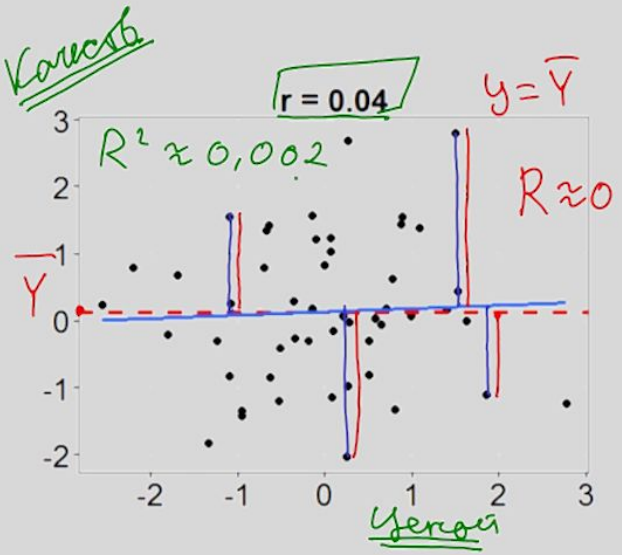
$$t = \frac{b_1}{se}$$

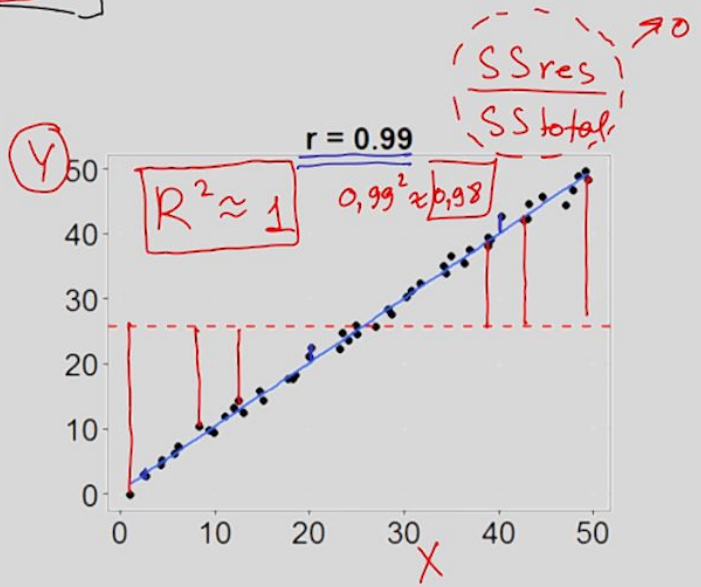
$df = N - 2$

Коэффициент детерминации

R^2 – доля дисперсии зависимой переменной (Y), объясняемая регрессионной моделью.

$$R^2 = 1 - \frac{SS_{res}}{SS_{total}}$$





Условия применения

- Линейная взаиомсвязь X и Y
- Нормальное распределение остатков
- Гомоскедастичность - постоянная изменчивость остактков на всех уровнях независимой переменной

hhtp://bitly.com/slr_diag

spark.rstudio.com/minebocek/slr_diag/

Diagnostics for simple linear regression

Select a trend:

- (•) Linear up
- ( ) Linear down
- ( ) Curved up
- ( ) Curved down
- ( ) Fan-shaped

[x] Show residuals

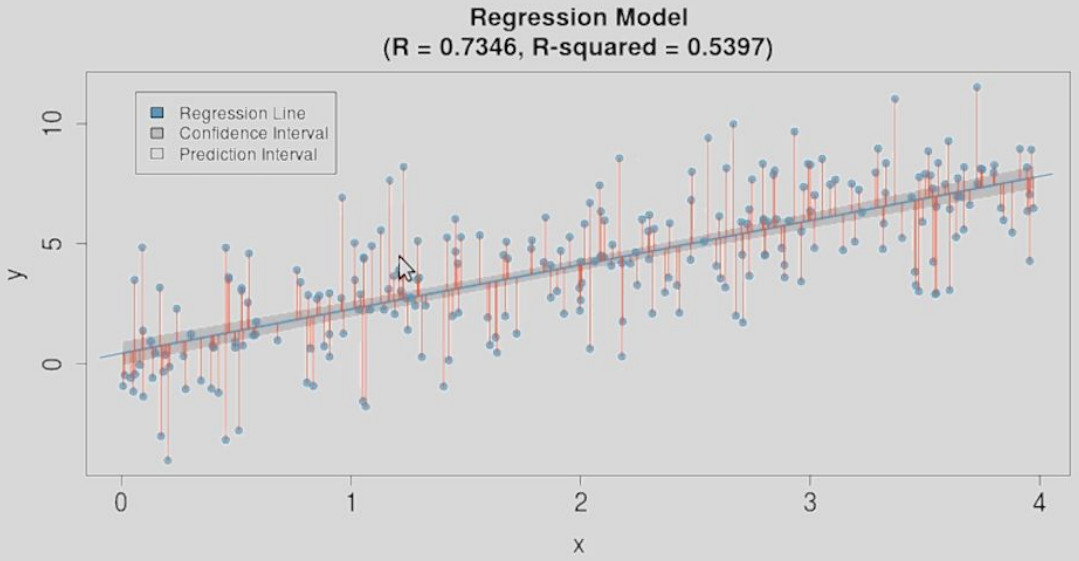
This applet uses ordinary least squares (OLS) to fit a regression line to the data with the selected trend. The applet is designed to help you practice evaluating whether or not the linear model is an appropriate fit to the data. The three diagnostic plots on the lower half of the page are provided to help you identify undesirable patterns in the residuals that may arise from non-linear trends in the data.

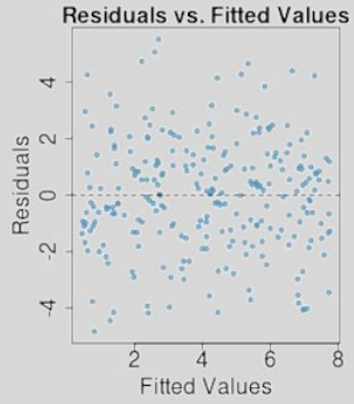
Rate this app!

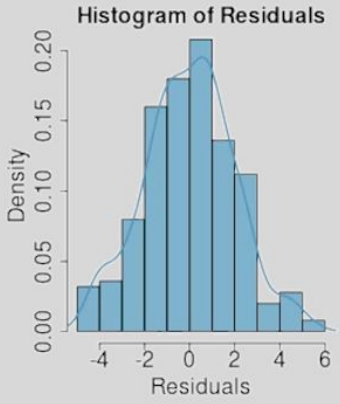
View code

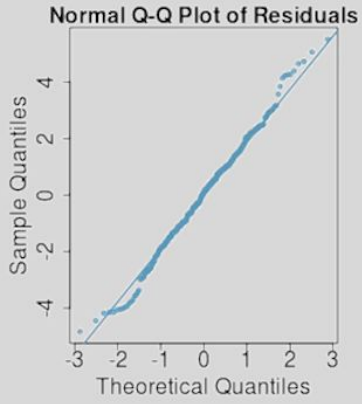
Check out other apps

Want to learn more for free?









spark.rstudio.com/minebocek/slr_diag/

Diagnostics for simple linear regression

Select a trend:

- ○ Linear up
- ◉ Linear down
- ○ Curved up
- ○ Curved down
- ○ Fan-shaped

☑ Show residuals

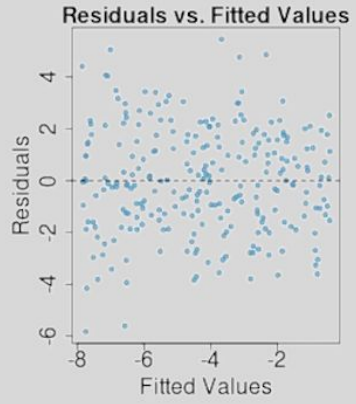
This applet uses ordinary least squares (OLS) to fit a regression line to the data with the selected trend. The applet is designed to help you practice evaluating whether or not the linear model is an appropriate fit to the data. The three diagnostic plots on the lower half of the page are provided to help you identify undesirable patterns in the residuals that may arise from non-linear trends in the data.

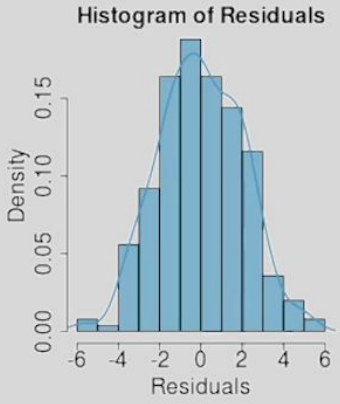
Rate this app!

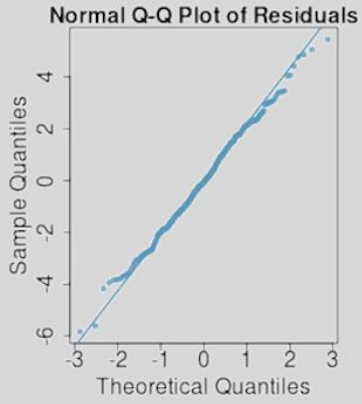
View code

Check out other apps

Want to learn more for free?







Diagnostics for simple linear regression

Select a trend:

- Linear up
- Linear down
- Curved up
- Curved down
- Fan-shaped

☑ Show residuals

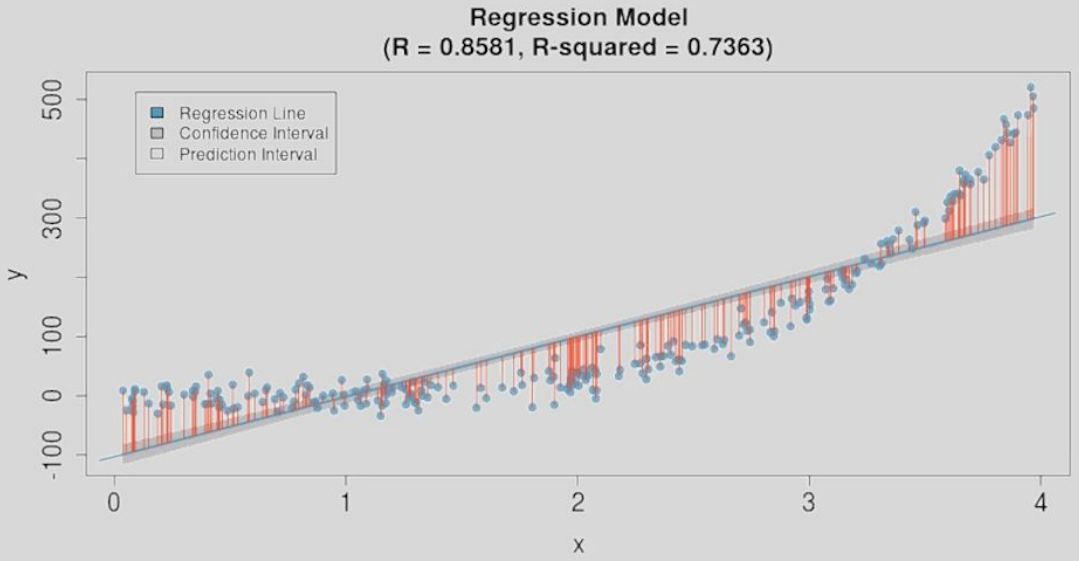
This applet uses ordinary least squares (OLS) to fit a regression line to the data with the selected trend. The applet is designed to help you practice evaluating whether or not the linear model is an appropriate fit to the data. The three diagnostic plots on the lower half of the page are provided to help you identify undesirable patterns in the residuals that may arise from non-linear trends in the data.

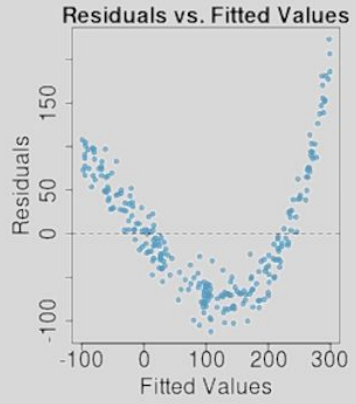
Rate this app!

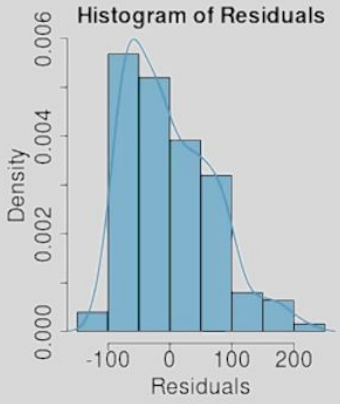
View code

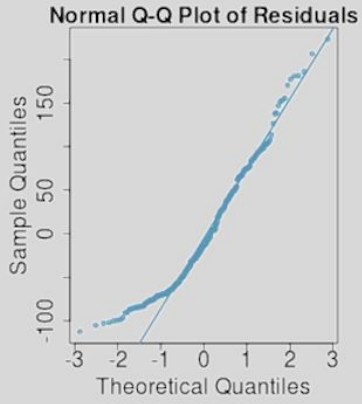
Check out other apps

Want to learn more for free?









spark.rstudio.com/minebocek/slr_diag/

Diagnostics for simple linear regression

Select a trend:

- Linear up
- Linear down
- Curved up
- Curved down
- Fan-shaped

☑ Show residuals

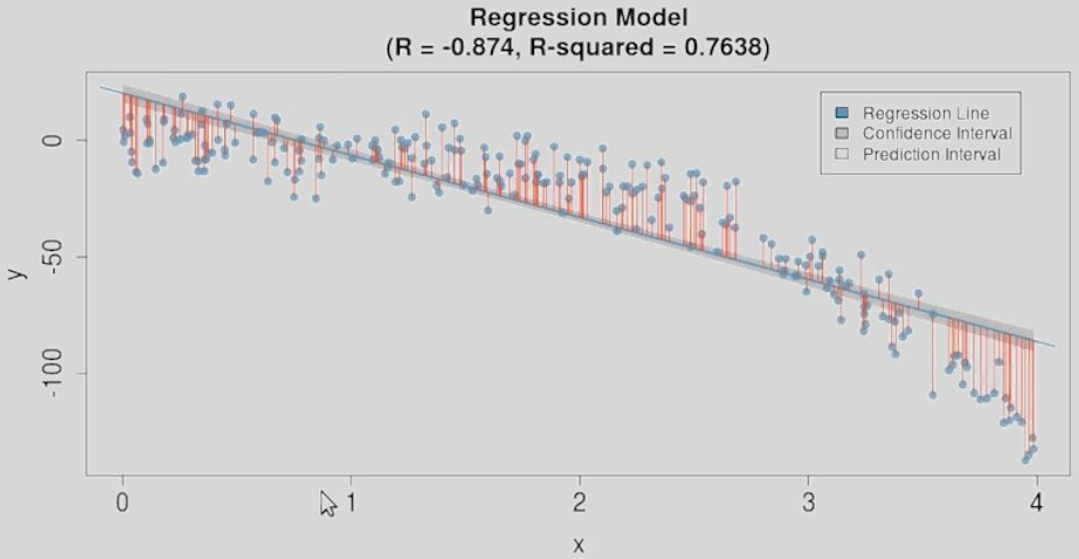
This applet uses ordinary least squares (OLS) to fit a regression line to the data with the selected trend. The applet is designed to help you practice evaluating whether or not the linear model is an appropriate fit to the data. The three diagnostic plots on the lower half of the page are provided to help you identify undesirable patterns in the residuals that may arise from non-linear trends in the data.

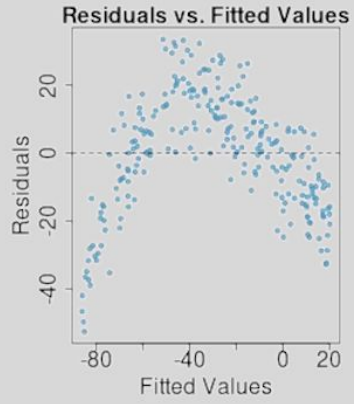
Rate this app!

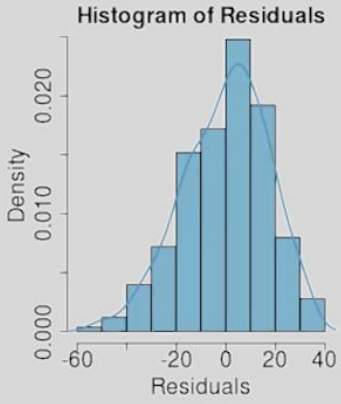
View code

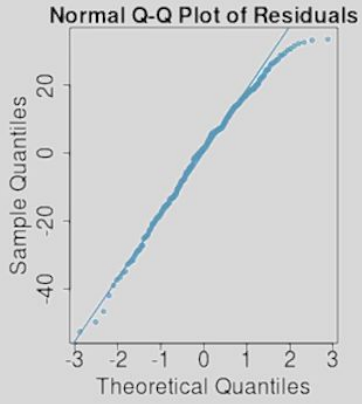
Check out other apps

Want to learn more for free?









spark.rstudio.com/minebocek/slr_diag/

Diagnostics for simple linear regression

Select a trend:

- ○ Linear up
- ○ Linear down
- ○ Curved up
- ○ Curved down
- ◉ Fan-shaped

☑ Show residuals

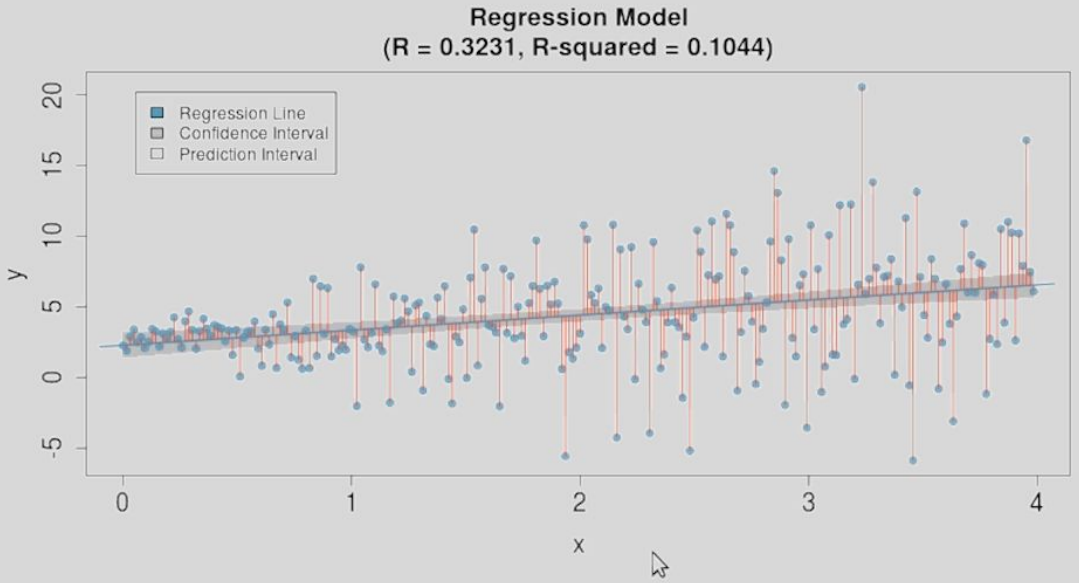
This applet uses ordinary least squares (OLS) to fit a regression line to the data with the selected trend. The applet is designed to help you practice evaluating whether or not the linear model is an appropriate fit to the data. The three diagnostic plots on the lower half of the page are provided to help you identify undesirable patterns in the residuals that may arise from non-linear trends in the data.

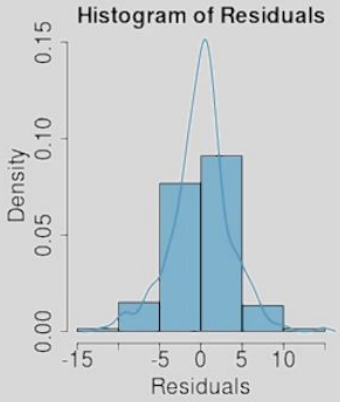
Rate this app!

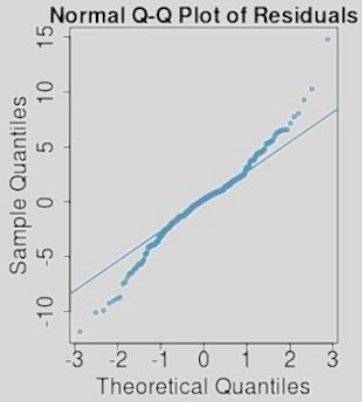
View code

Check out other apps

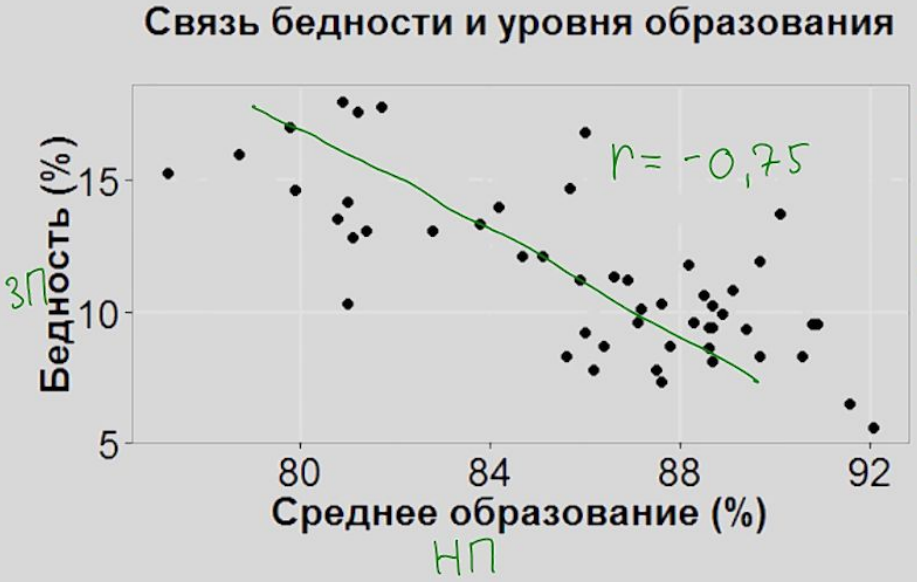
Want to learn more for free?







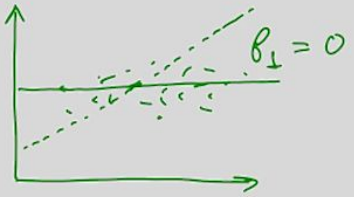
Регрессионный анализ с одной независимой переменной



1. $\hat{y} = b_0 + b_1 x$

2. R^2

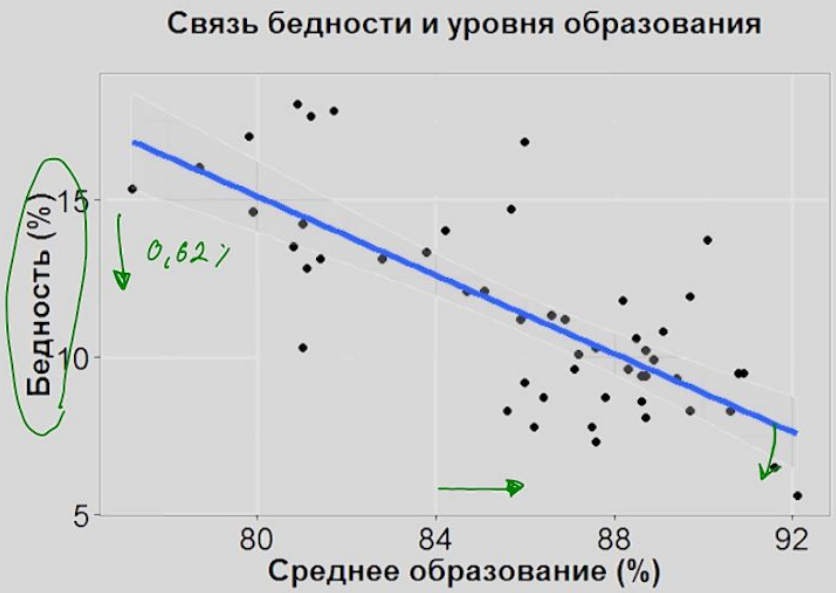
$b_1 = 0$: H_0



3. НП ⟶ ЗП

Descriptive statistics

Statistic	N	Mean	St. Dev.	Min	Max
poverty	51	11.3	3.1	5.6	18.0
hs_grad	51	86.0	3.7	77.2	92.1



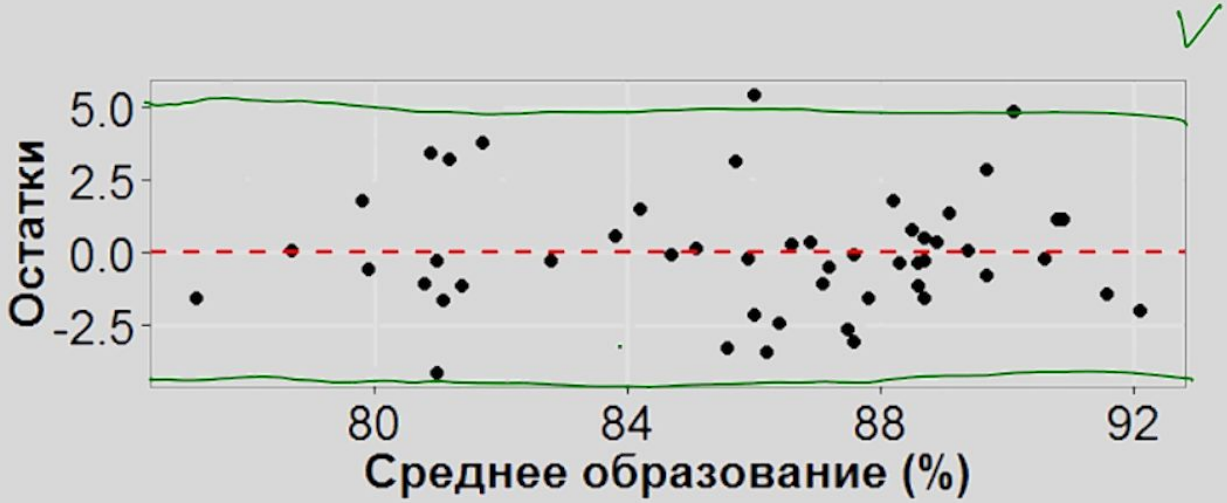
$\hat{y} = 64{,}78 - 0{,}62 \cdot hs\_grad$

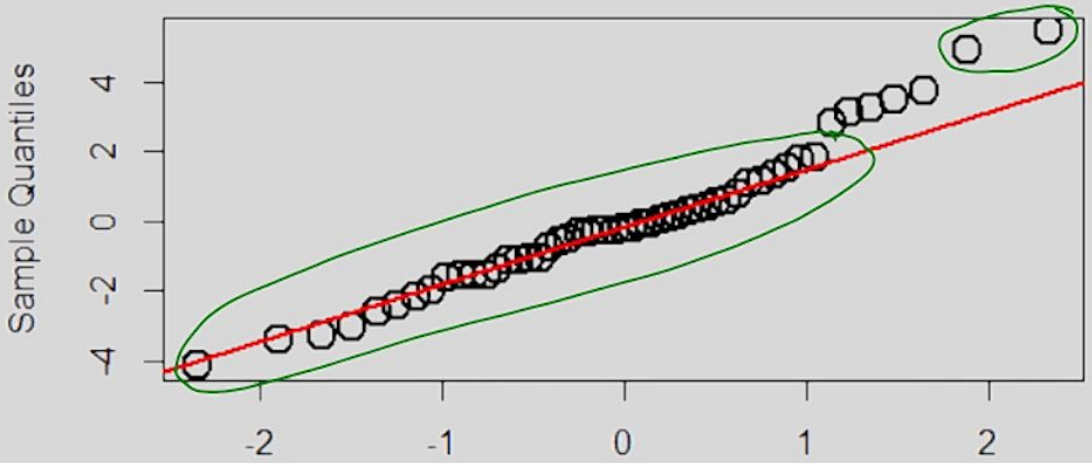
$H_0 \quad b_1 = 0$

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept) b_0	64.7810	6.8026	9.52	0.0000	$p < 0{,}05$
hs_grad b_1	-0.6212	0.0790	-7.86	0.0000	$p < 0{,}05$

Multiple R-squared: 0.5578,
F-statistic (1, 49) = 61.81, p-value < 0.01

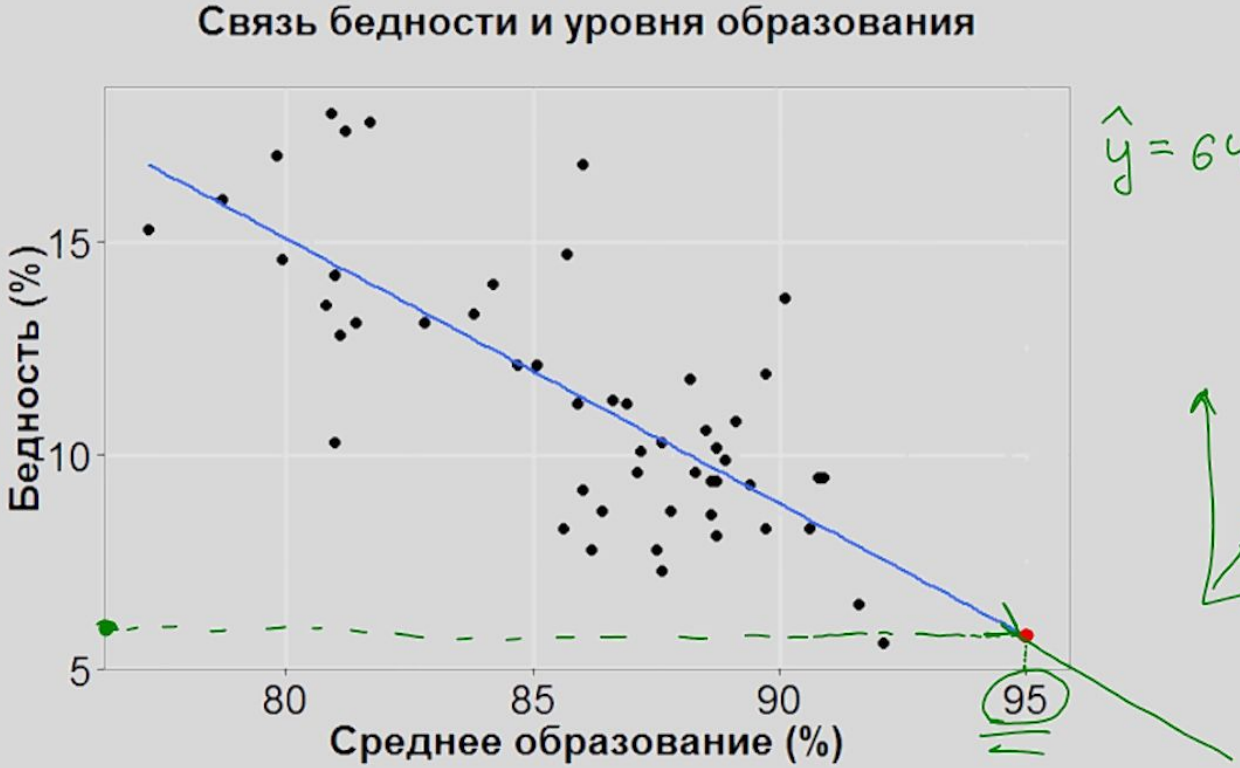
Анализ остатков



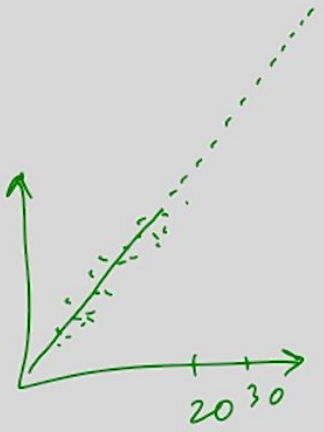


Предсказание значений

$$\widehat{\text{Бедность}} = 64{,}78 - 0{,}62 \cdot \text{образование}^{95}$$

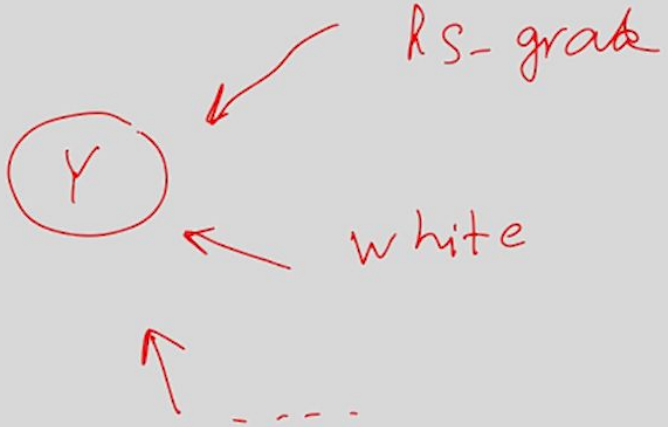


$$\hat{y} = 64{,}78 - 0{,}62 \cdot 95 = 5{,}77$$



Множественная регрессия
Multiple regression

Множественная регрессия позволяет исследовать влияние сразу нескольких независимых переменных на одну зависимую переменную.



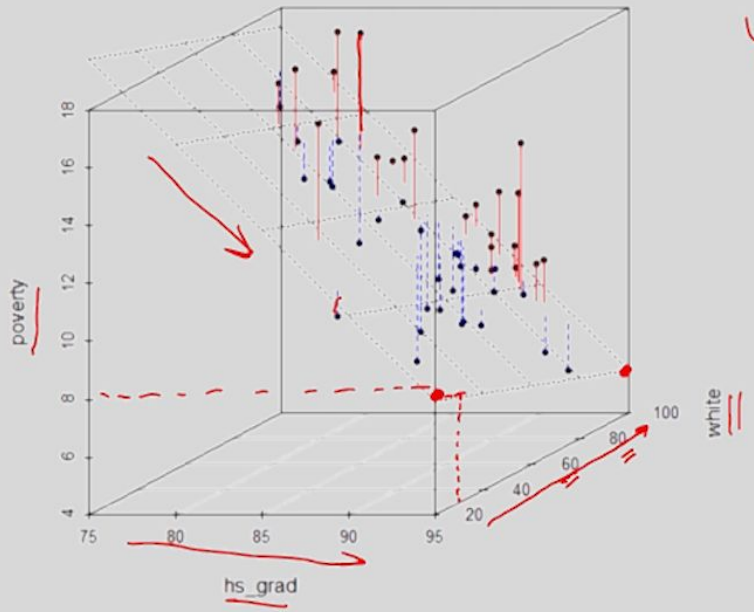
$$\hat{y} = b_0 + b_1 x_1$$

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

$$\hat{y} = b_0 + b_1 x_1 + \ldots + b_n x_n.$$

Множественная регрессия
Multiple regression

Множественная регрессия позволяет исследовать влияние сразу нескольких независимых переменных на одну зависимую переменную.



$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

Требования к данным

Линейная зависимость переменных

Нормальное распределение остатков

Гетероскедастичность

Проверка на мультиколлинеарность

Нормальное распределение переменных (желательно)

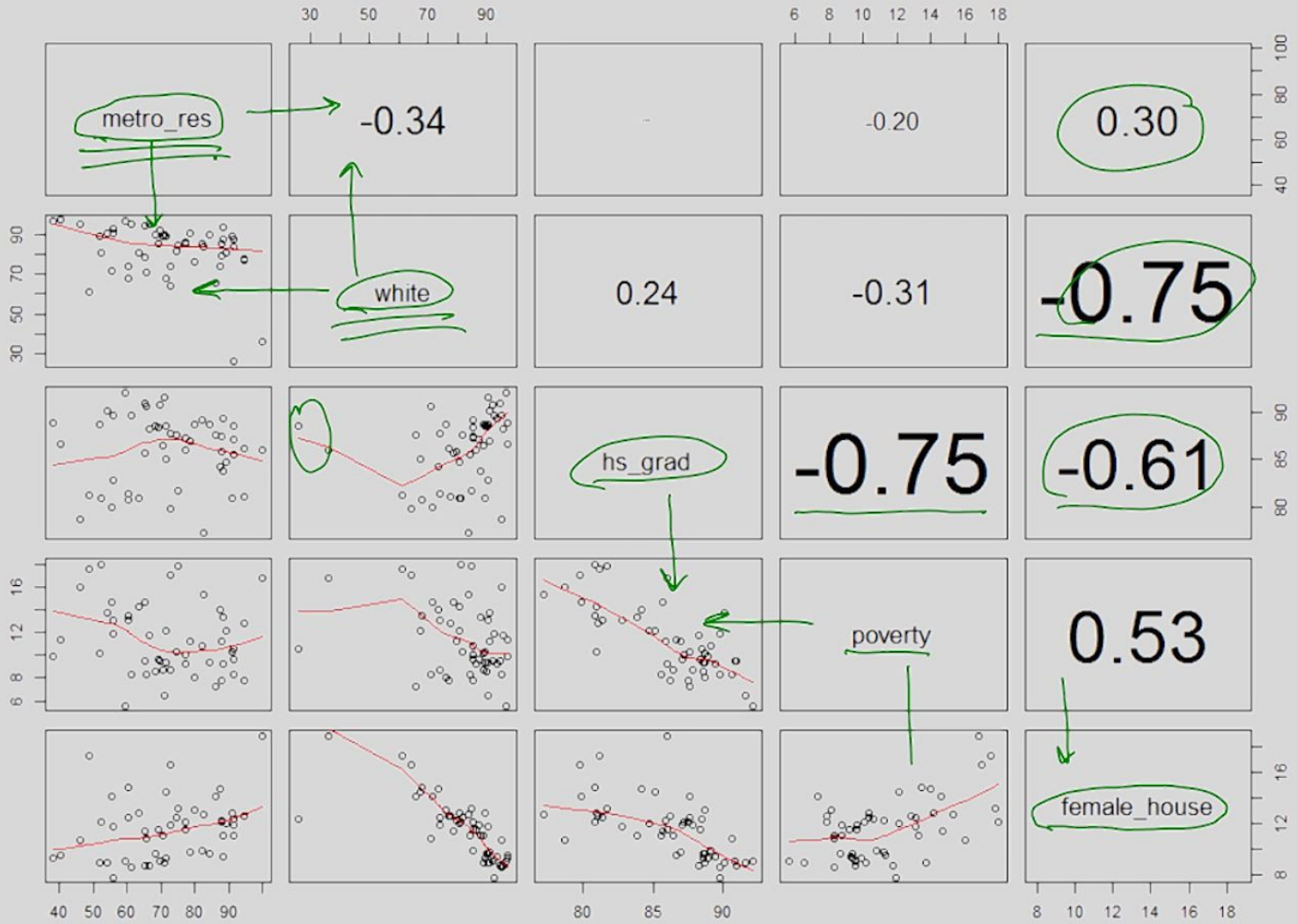
Множественная регрессия

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	66.47	12.5	5.28	0.0000
metro_res	-0.06	0.02	-2.88	0.0060
white	-0.05	0.03	-1.46	0.1522
hs_grad	-0.55	0.1	-5.29	0.0000
female_house	0.05	0.24	0.21	0.8363

Multiple R-squared: 0.6416, Adjusted R-squared: 0.6104

F-statistic (4, 46) = 20.58, p-value < 0,01

Исправленный R - квадрат (adjusted R-squared) - скорректированный коэффициент детерминации. Рассчитывается при включении в модель дополнительных независимых переменных.


metro_res
-0.34
-0.20
0.30
white
0.24
-0.31
-0.75
hs_grad
-0.75
-0.61
poverty
0.53
female_house

Выбор модели

Удаляем	Модель	Adj R-squared
	Poverty ~ hs_grad + white + metro_res + female_house	0,61
female_house	Poverty ~ hs_grad + white + metro_res	0,62
metro_res	Poverty ~ hs_grad + white + female_house	0,54
white	Poverty ~ hs_grad + metro_res + female_house	0,60
hs_grad	Poverty ~ white + metro_res + female_house	0,38

Удаляем	Модель	Adj R-squared
	Poverty ~ hs_grad + white + metro_res	0,62
metro_res	Poverty ~ hs_grad + white	0,55
white	Poverty ~ hs_grad + metro_res	0,57
hs_grad	Poverty ~ white + metro_res	0,17

Итоговая модель

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	68.7	6.39	10.76	0.0000
white	-0.05	0.02	-2.48	0.0167
metro_res	-0.05	0.02	-2.93	0.0053
hs_grad	-0.57	0.08	-7.57	0.0000

Multiple R-squared: 0.6412, Adjusted R-squared: 0.62

F-statistic (3,47) = 28, p-value < 0,01

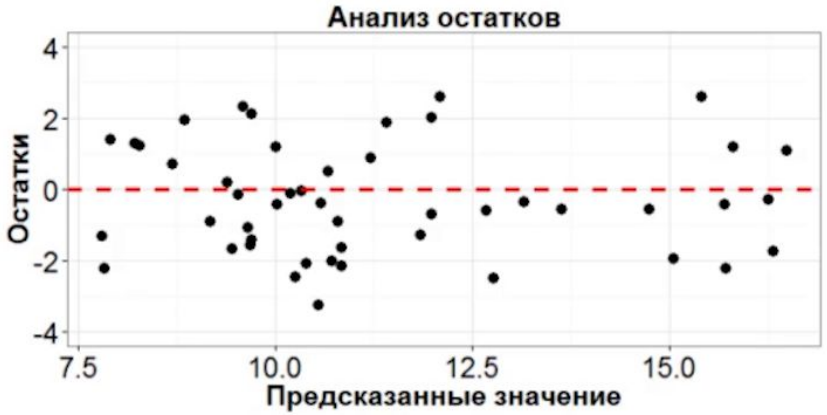
$$\hat{Y} = 68,7 - 0,05 \cdot white - 0,05 \cdot metro\_res - 0,57 \cdot hs\_grad$$

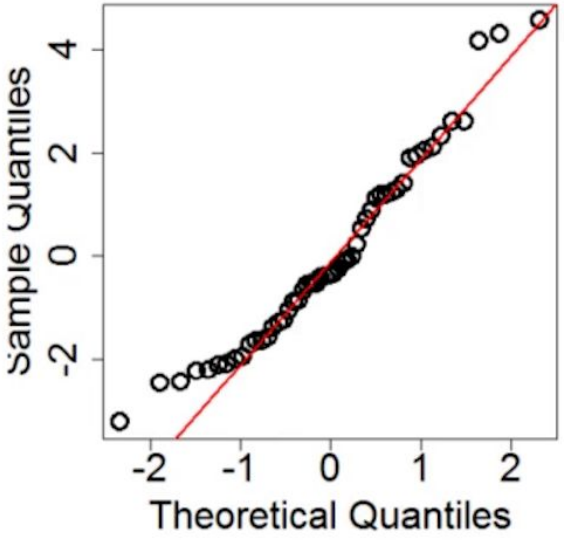
Итоговая модель

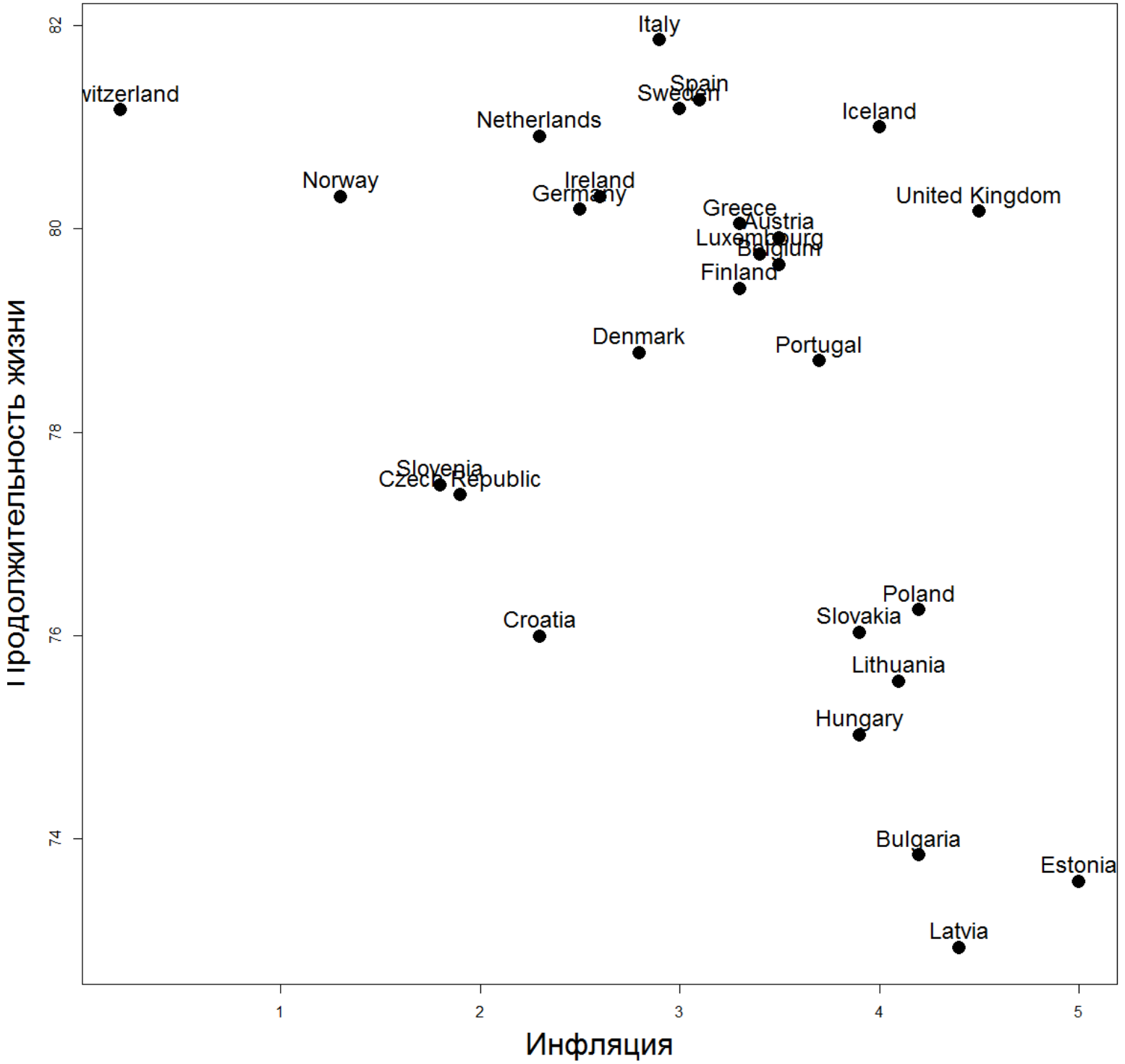
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	68.7	6.39	10.76	0.0000
white	-0.06	0.02	-2.48	0.0167
metro_res	-0.05	0.02	-2.93	0.0053
hs_grad	-0.57	0.08	-7.57	0.0000

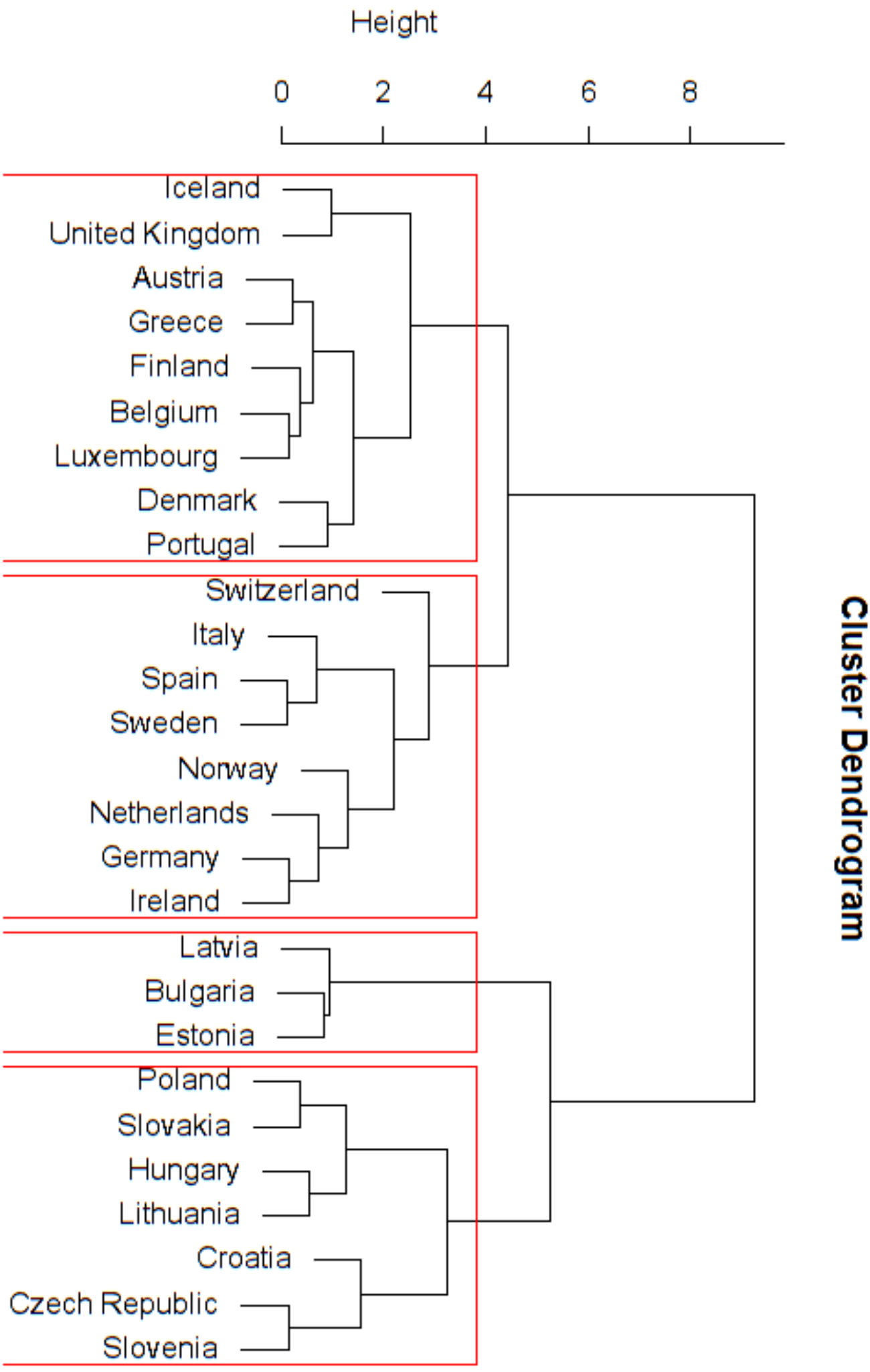
Multiple R-squared: 0.6412, Adjusted R-squared: 0.62

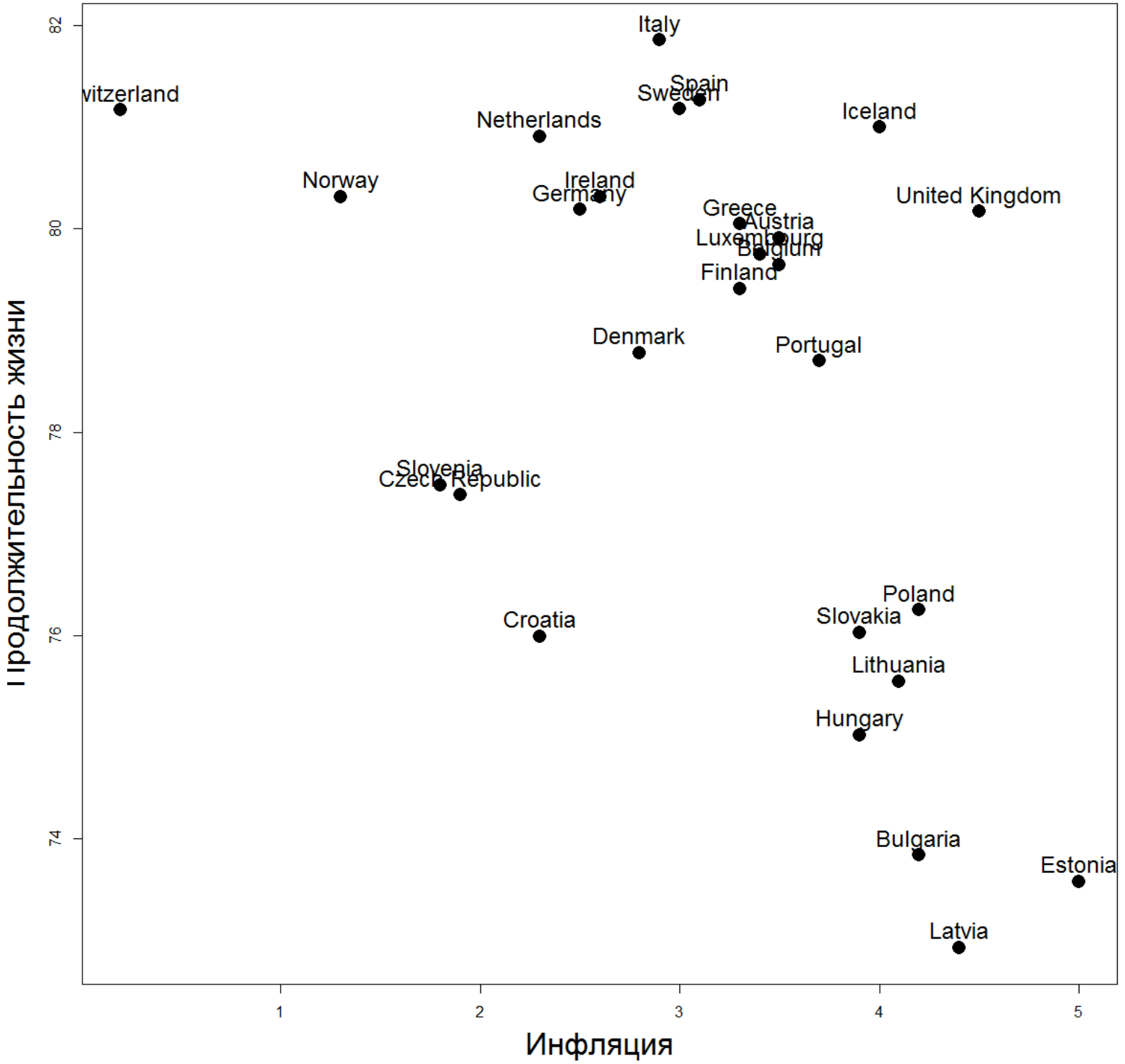
F-statistic (3,47) = 28, p-value < 0,01





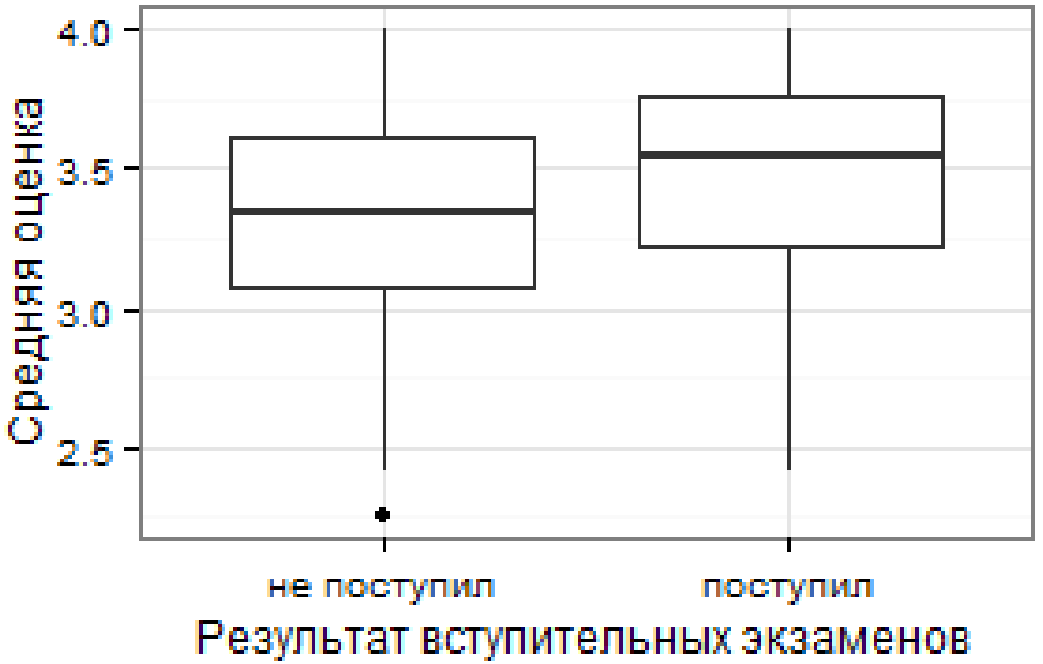

Italy
vitzerland
Spain
Sweden
Iceland
Netherlands
Norway
Ireland
Germany
United Kingdom
Greece
Austria
Luxembourg
Belgium
Finland
Denmark
Portugal
Slovenia
Czech Republic
Poland
Slovakia
Croatia
Lithuania
Hungary
Bulgaria
Estonia
Latvia
82
80
78
76
74
1
2
3
4
5
Продолжительность жизни
Инфляция


Height
0
2
4
6
8
Iceland
United Kingdom
Austria
Greece
Finland
Belgium
Luxembourg
Denmark
Portugal
Switzerland
Italy
Spain
Sweden
Norway
Netherlands
Germany
Ireland
Latvia
Bulgaria
Estonia
Poland
Slovakia
Hungary
Lithuania
Croatia
Czech Republic
Slovenia
Cluster Dendrogram


Italy
vitzerland
Spain
Sweden
Iceland
Netherlands
Norway
Ireland
Germany
United Kingdom
Greece
Austria
Luxembourg
Belgium
Finland
Denmark
Portugal
Slovenia
Czech Republic
Poland
Slovakia
Croatia
Lithuania
Hungary
Bulgaria
Estonia
Latvia
82
80
78
76
74
1
2
3
4
5
Продолжительность жизни
Инфляция

Логистическая регрессия

	admit	mark
1	0	3.61
2	1	3.67
3	1	4.00
4	1	3.19
5	0	2.93
6	1	3.00
7	1	2.98
8	0	3.08
9	1	3.39
10	0	3.92
11	0	4.00
12	0	3.22
13	1	4.00



	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-4.3576	1.0353	-4.21	0.0000
mark	1.0511	0.2989	3.52	0.0004