

# **Линейная регрессия**

# Ковариация

$$cov(X, Y) = \frac{1}{n - 1} \sum (Y_i - \bar{Y})(X_i - \bar{(X)})$$

**Величина больше 0, если положительным отклонениям от среднего в X соответствуют положительные отклонения от среднего в Y, а отрицательным - отрицательные**

# Ковариация

```
ssize <- 1000  
  
x <- rnorm(ssize)  
y <- x - 10  
print(cov(x, y))
```

```
## [1] 1.017494
```

```
x <- rnorm(ssize)  
y <- 2 * x - 10  
print(cov(x, y))
```

```
## [1] 2.066004
```

```
x <- rnorm(ssize)  
y <- -2 * x - 10  
print(cov(x, y))
```

```
## [1] -2.107525
```

```
y <- -5 * x - 10  
print(cov(x, y))
```

```
## [1] -5.268812
```

**Какие минусы?**

# Ковариация

```
ssize <- 1000  
  
x <- rnorm(ssize)  
y <- x - 10  
print(cov(x, y))
```

```
## [1] 1.017494
```

```
x <- rnorm(ssize)  
y <- 2 * x - 10  
print(cov(x, y))
```

```
## [1] 2.066004
```

```
x <- rnorm(ssize)  
y <- -2 * x - 10  
print(cov(x, y))
```

```
## [1] -2.107525
```

```
y <- -5 * x - 10  
print(cov(x, y))
```

```
## [1] -5.268812
```

**Какие минусы?**

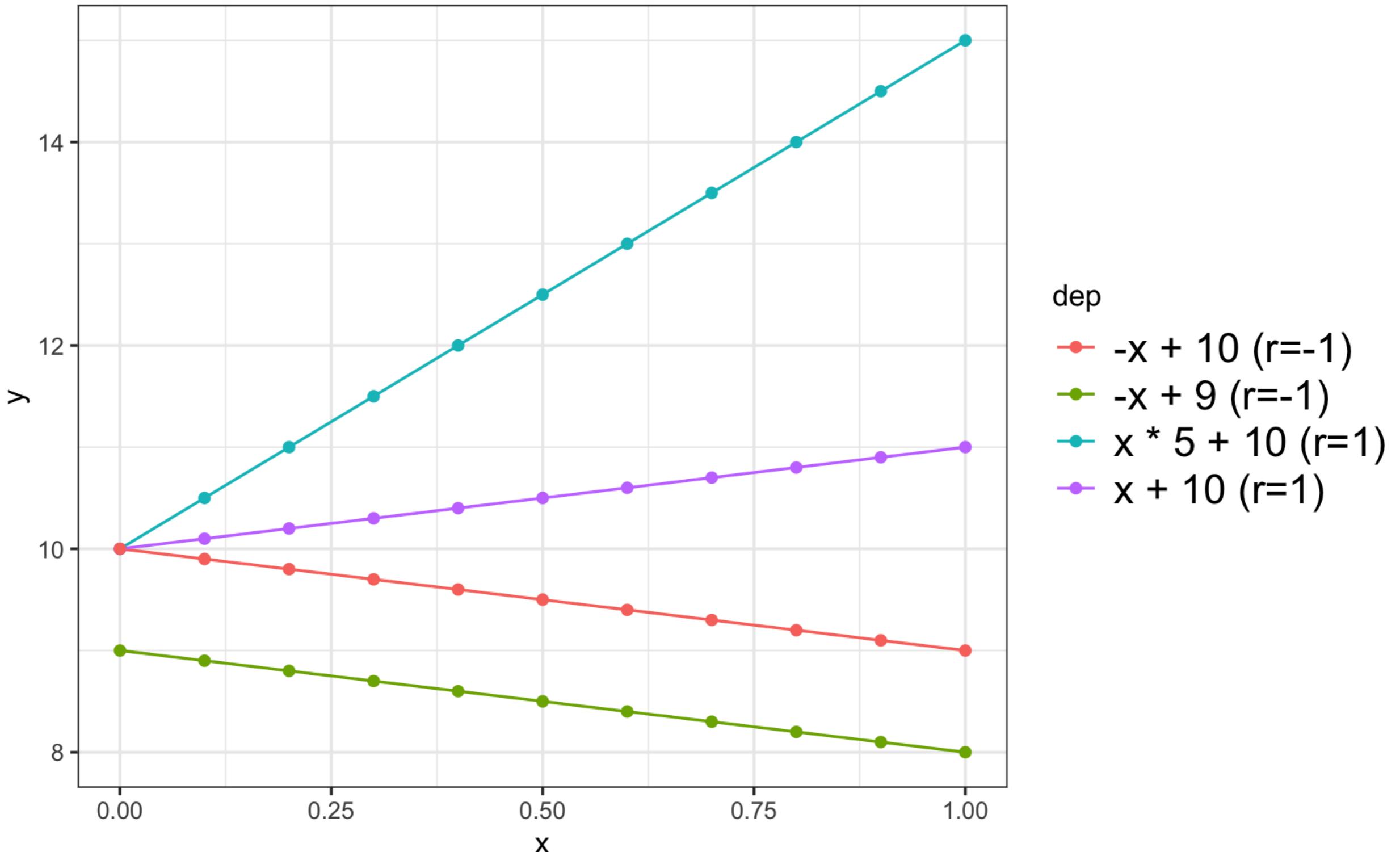
**Главный - зависит от шкалы измерения X и Y**

# Корреляция (=корреляция Пирсона)

$$cor(X, Y) = \frac{\frac{1}{n-1} \sum (Y_i - \bar{Y})(X_i - \bar{X})}{sd(X)sd(Y)}$$

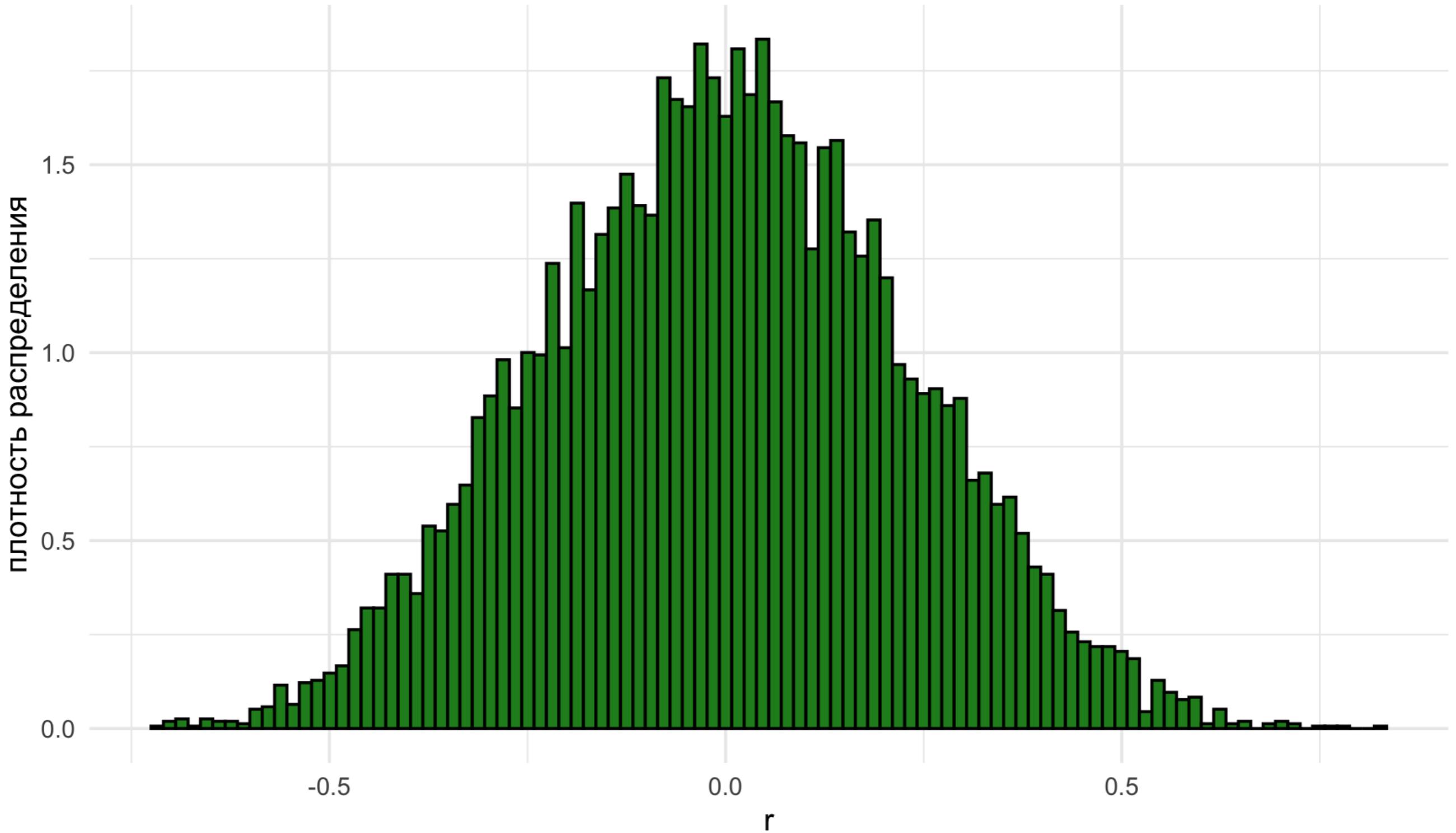
1. Всегда лежит в пределах от -1 до 1
2. Если к х или у прибавить одно и то же число, то значение корреляции не изменится
3. Если х или у умножить на одно и то же положительное число, то значение корреляции не изменится, если на отрицательное - поменяется на противоположное
4. Корреляция выборок - случайная величина, оценка корреляции генеральных совокупностей
5. Если говорить о генеральных совокупностях, то если у **зависит от** х линейно , то корреляция будет равна -1 или 1, верно и обратное
6. Если х и у нормально распределены, то из равенства корреляции 0 следует **независимость** переменных
7. **В общем случае из равенства корреляции 0 не следует независимость переменных**
8. **Из значимой корреляции не следует причинно-следственная связь**
9. **Всегда когда считаем корреляцию на реальных данных и рассуждаем на основании этих корреляций надо строить графики.**
10. **Плохо учитывает или учитывает иных зависимостей кроме линейных**

1. Всегда лежит в пределах от -1 до 1
2. Если к  $x$  или  $y$  прибавить одно и то же число, то значение корреляции не изменится
3. Если  $x$  или  $y$  умножить на одно и то же положительное число, то значение корреляции не изменится, если на отрицательное - поменяется на противоположное

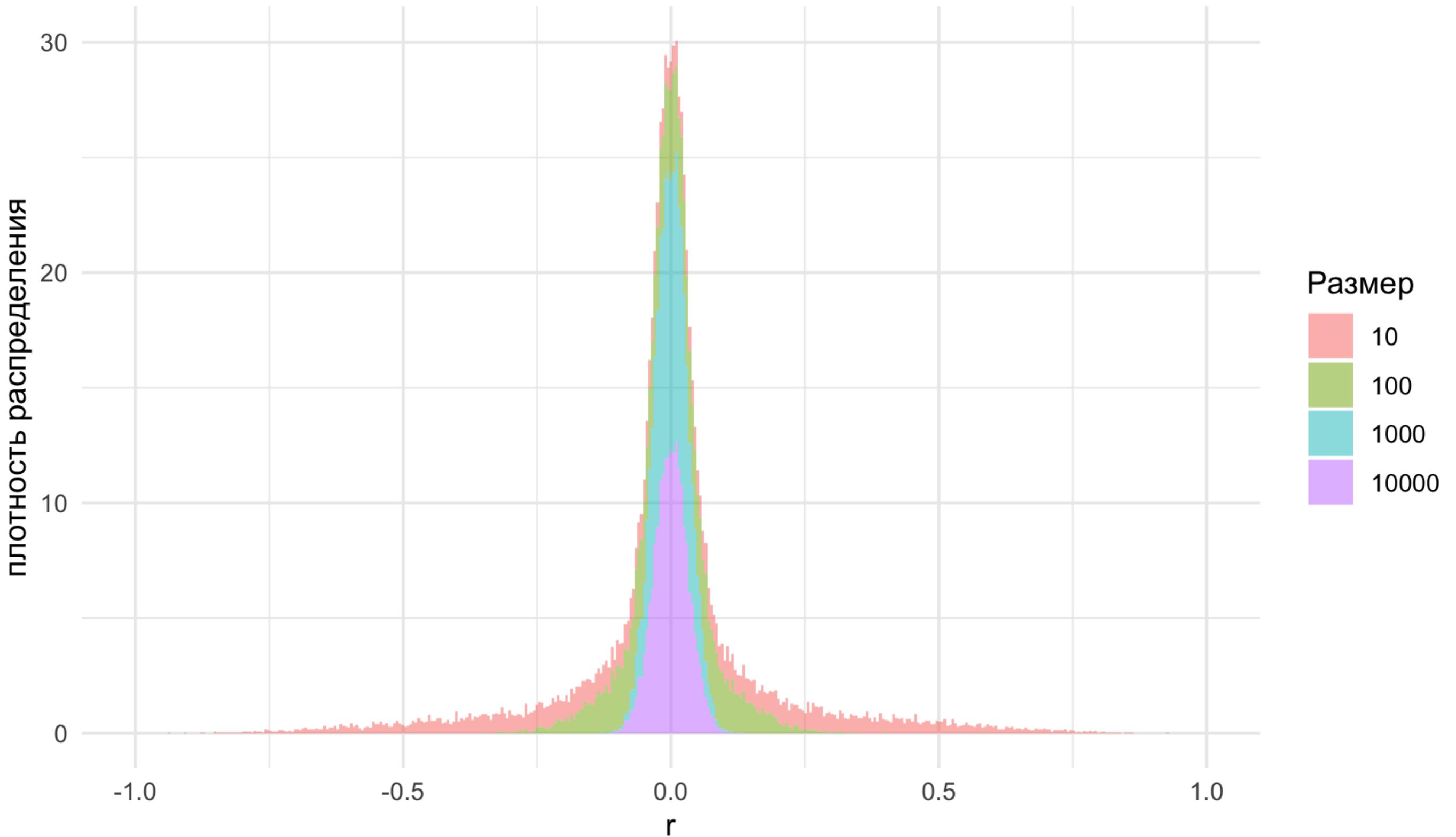


#### 4. Корреляция выборок - случайная величина, оценка корреляции генеральных совокупностей

Распределение корреляции:  $x \sim N(0, 1)$ ,  $y \sim N(0,1)$ , размер выборки 20



Распределение корреляции:  $x \sim N(0, 1)$ ,  $y \sim N(0, 1)$



Если мы предполагаем, что настоящий коэффициент корреляции (посчитанный для генеральных совокупностей) равен 0 (наше  $H_0$ ), то вот такая величина распределена по Стьюденту

$$t = \frac{r}{\sqrt{1 - r^2}} \sqrt{n - 2} \sim t_{n-2}$$

**Тест на значимость корреляции в R - cor.test**

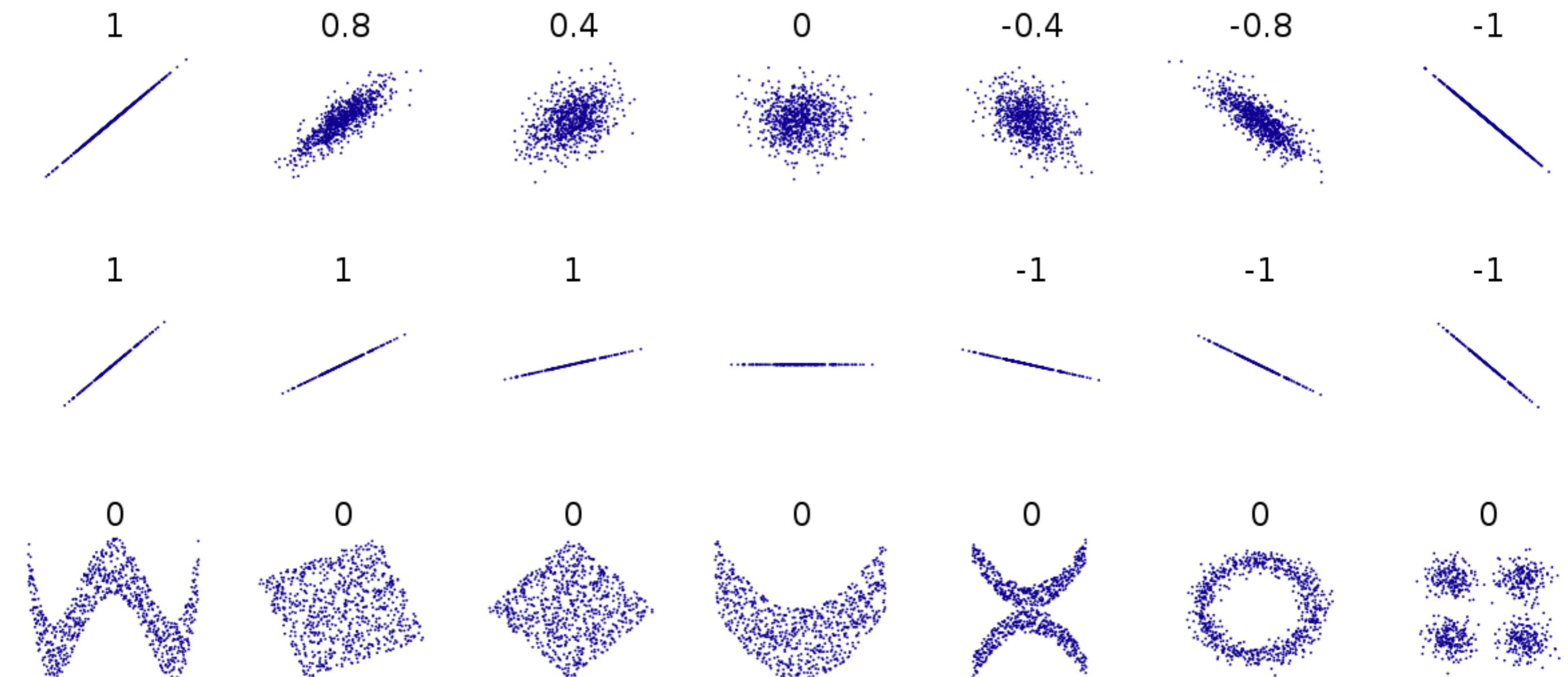
6. Если  $x$  и  $y$  нормально распределены, то из равенства корреляции 0 следует **независимость** переменных

Это не значит, что корреляция применима только для нормально распределенных  $x$  и  $y$

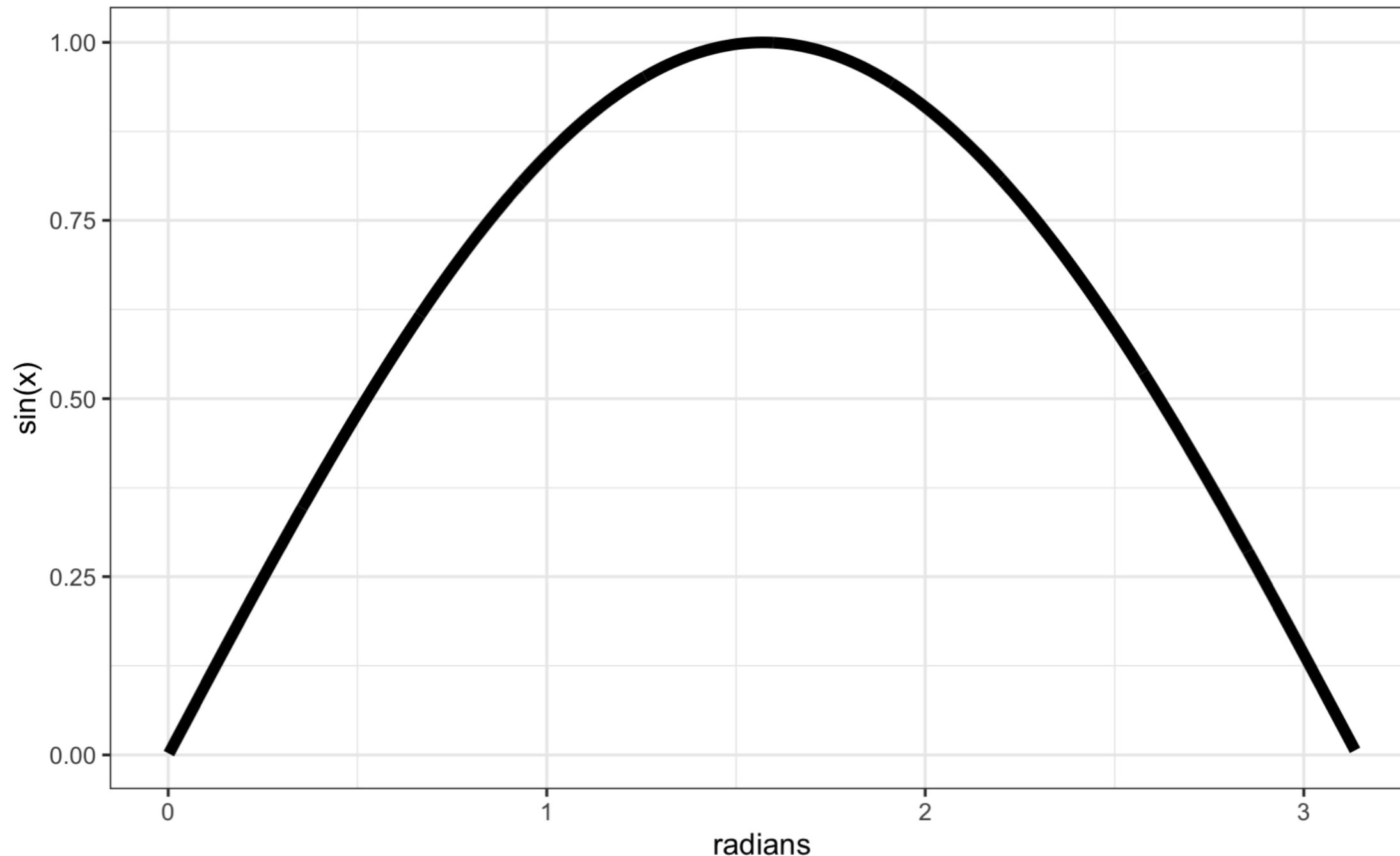
```
x <- runif(1000, 0, 1)
y <- 5 * x + rnorm(1000, sd=0.01)
cor(x, y)
```

```
## [1] 0.9999773
```

## 7. В общем случае из равенства корреляции 0 не следует независимость переменных

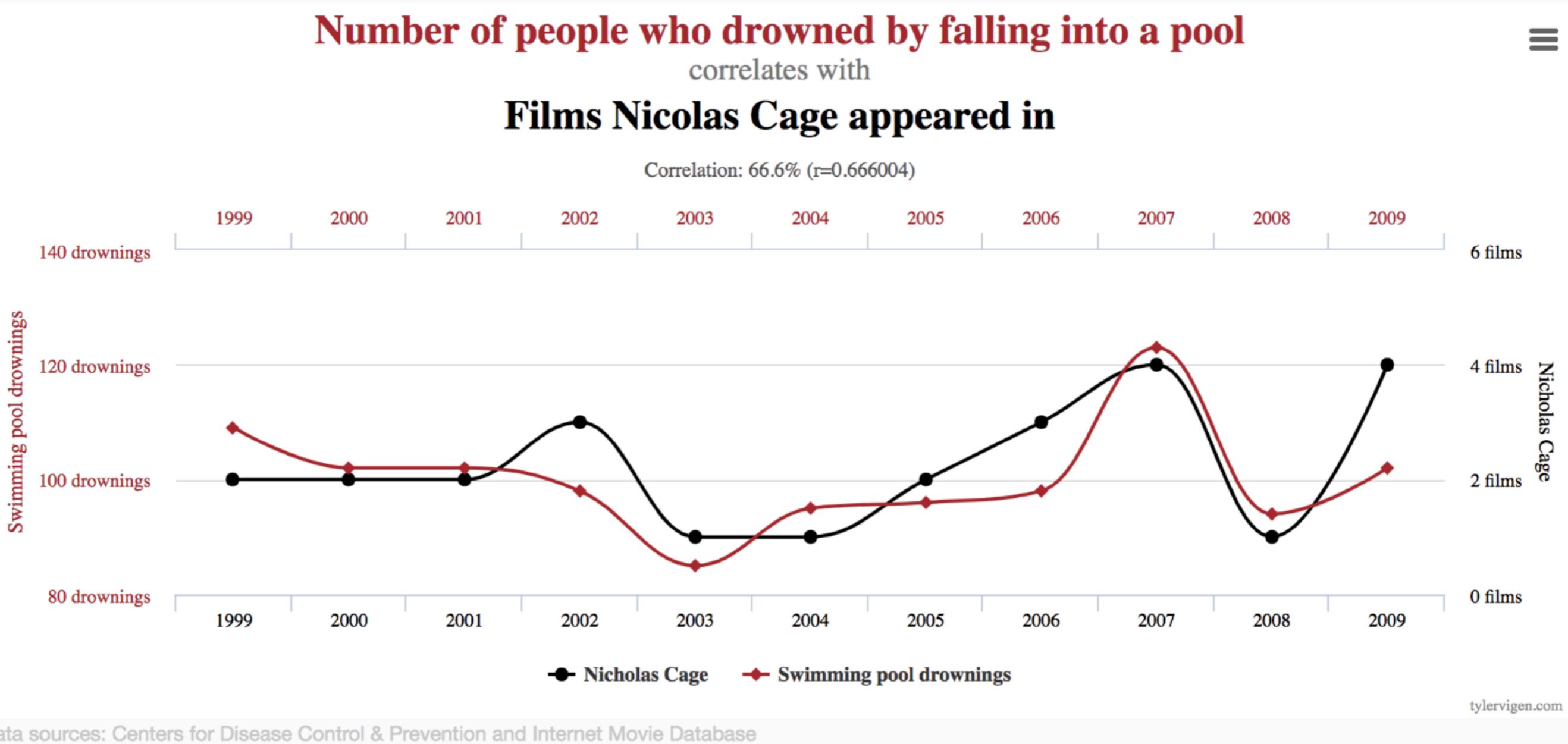


## 7. В общем случае из равенства корреляции 0 не следует независимость переменных



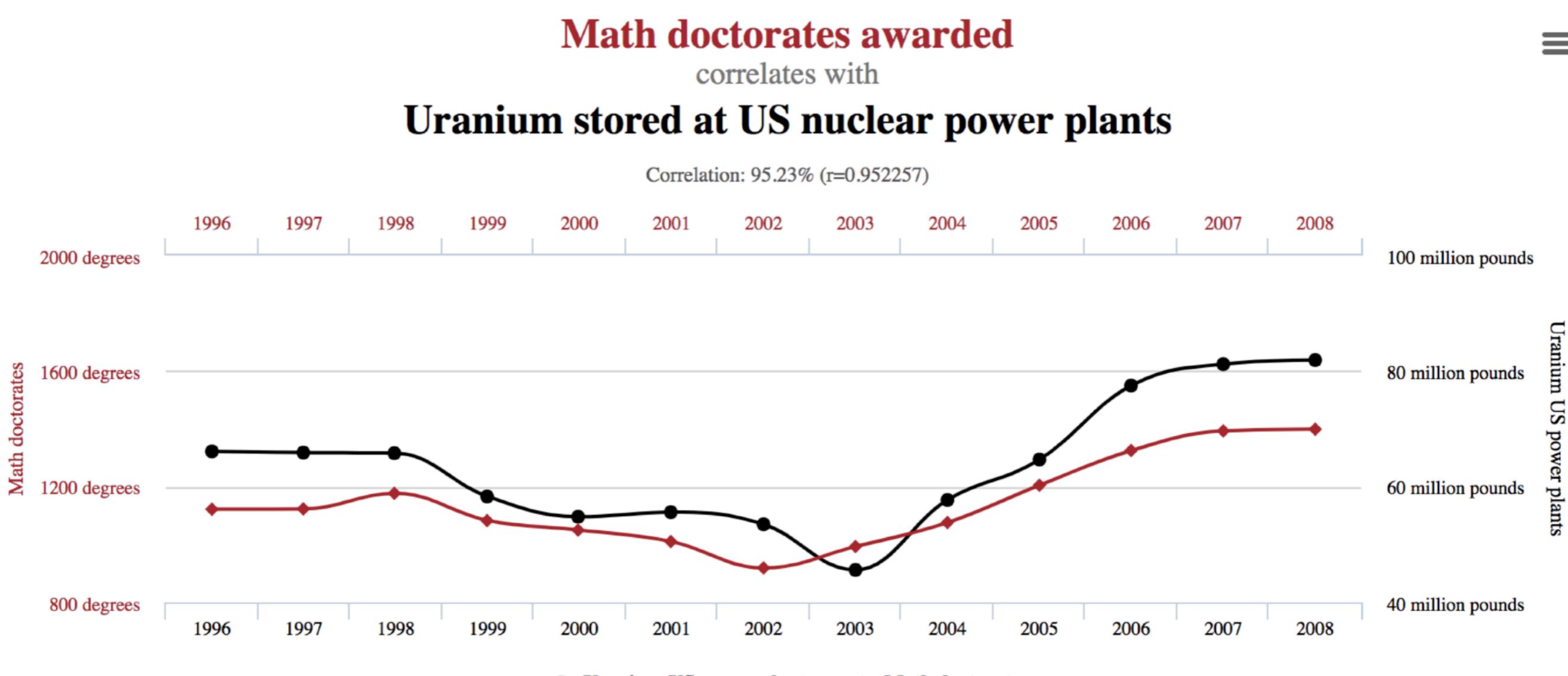
Берем  $x$  равномерно на отрезке  $[0; \pi]$ . Корреляция в среднем - 0

## 8. Из значимой корреляции не следует причинно-следственная связь



<https://www.tylervigen.com/spurious-correlations>

## 8. Из значимой корреляции не следует причинно-следственная связь

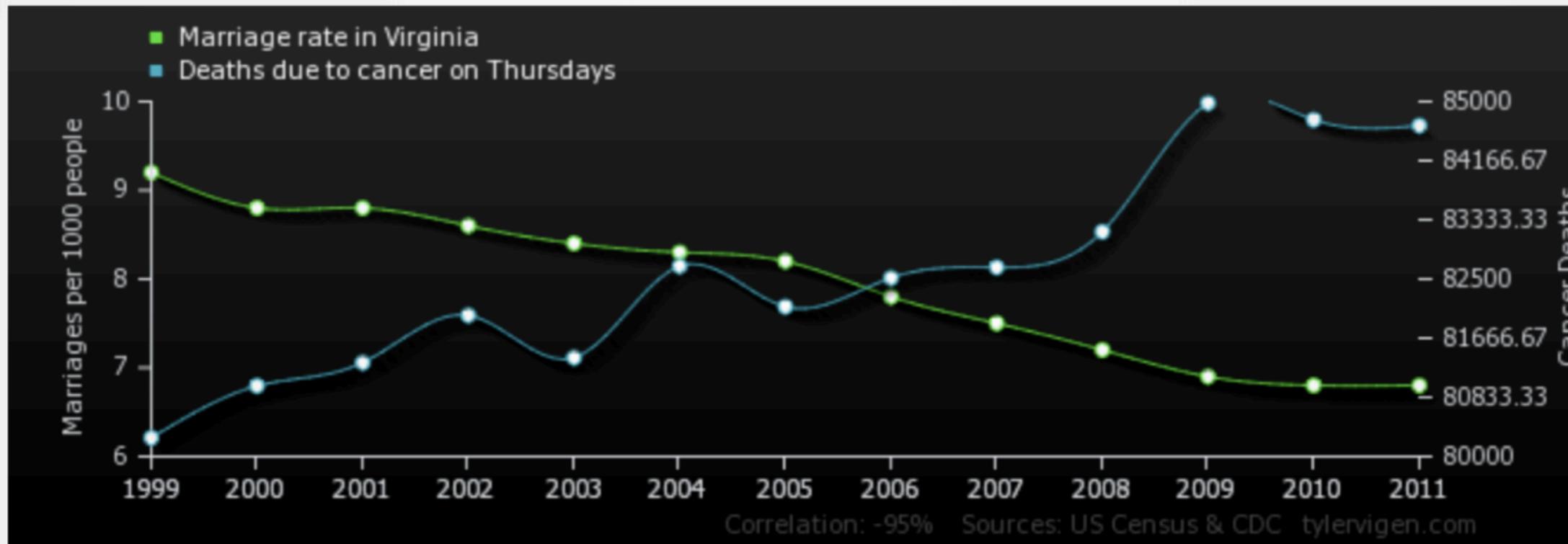


Data sources: National Science Foundation and Dept. of Energy

tylervigen.com

## 8. Из значимой корреляции не следует причинно-следственная связь

### Marriage rate in Virginia inversely correlates with Deaths due to cancer on Thursdays



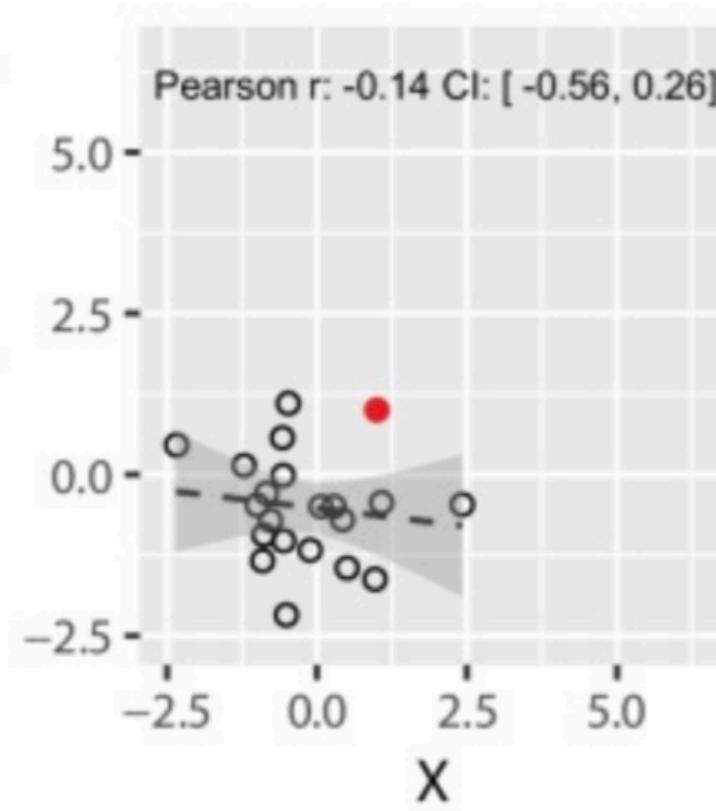
Upload this image to imgur

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
<i>Marriage rate in Virginia</i> Marriages per 1000 people (US Census)	9.2	8.8	8.8	8.6	8.4	8.3	8.2	7.8	7.5	7.2	6.9	6.8	6.8
<i>Deaths due to cancer on Thursdays</i> Cancer Deaths (CDC)	80,262	80,994	81,321	81,988	81,390	82,682	82,109	82,516	82,663	83,166	84,978	84,742	84,660

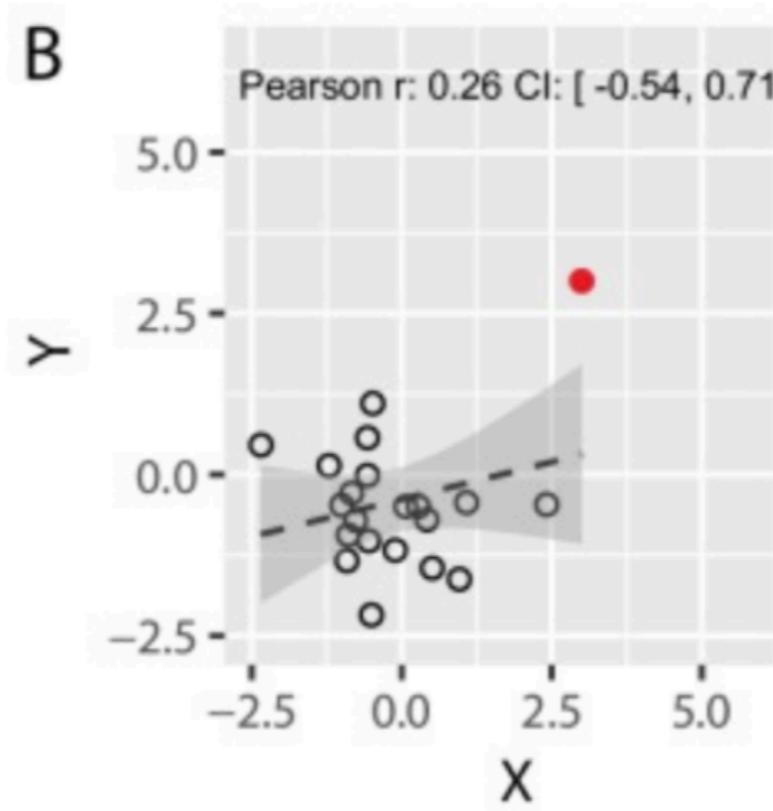
Correlation: -0.951857

## 9. Всегда когда считаем корреляцию на реальных данных и рассуждаем на основании этих корреляций надо строить графики.

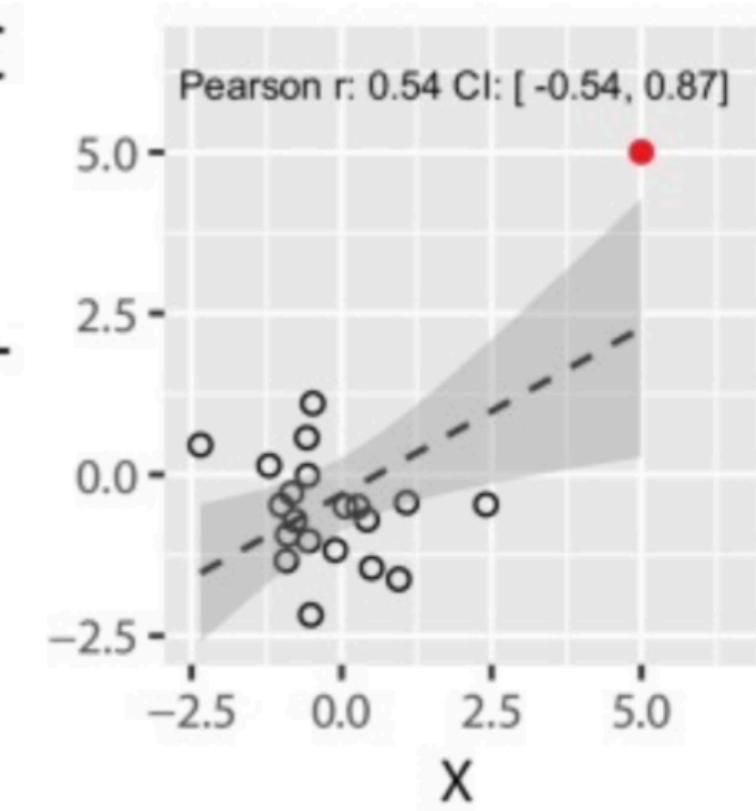
A



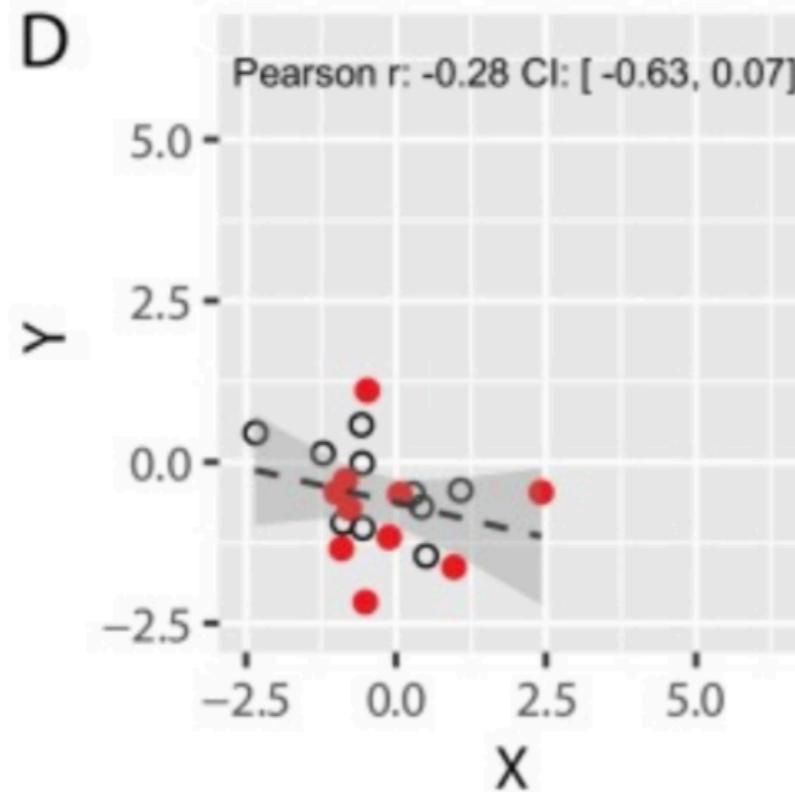
B



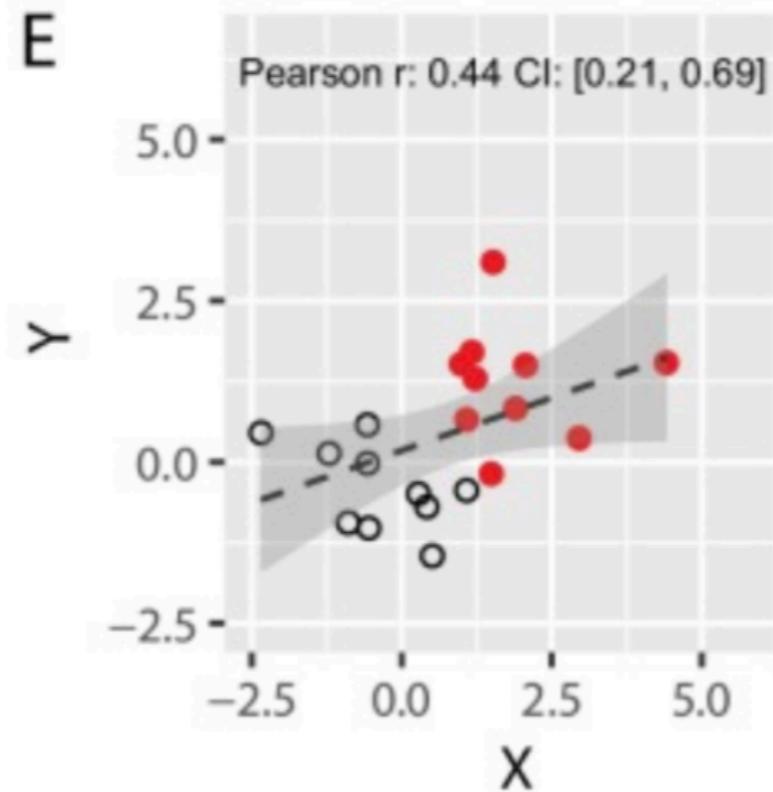
C



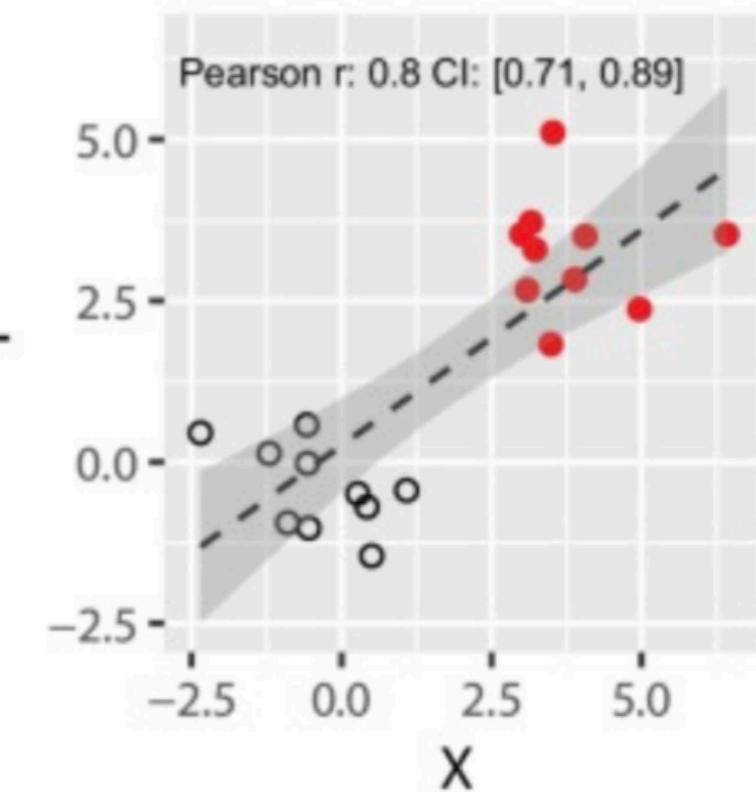
D



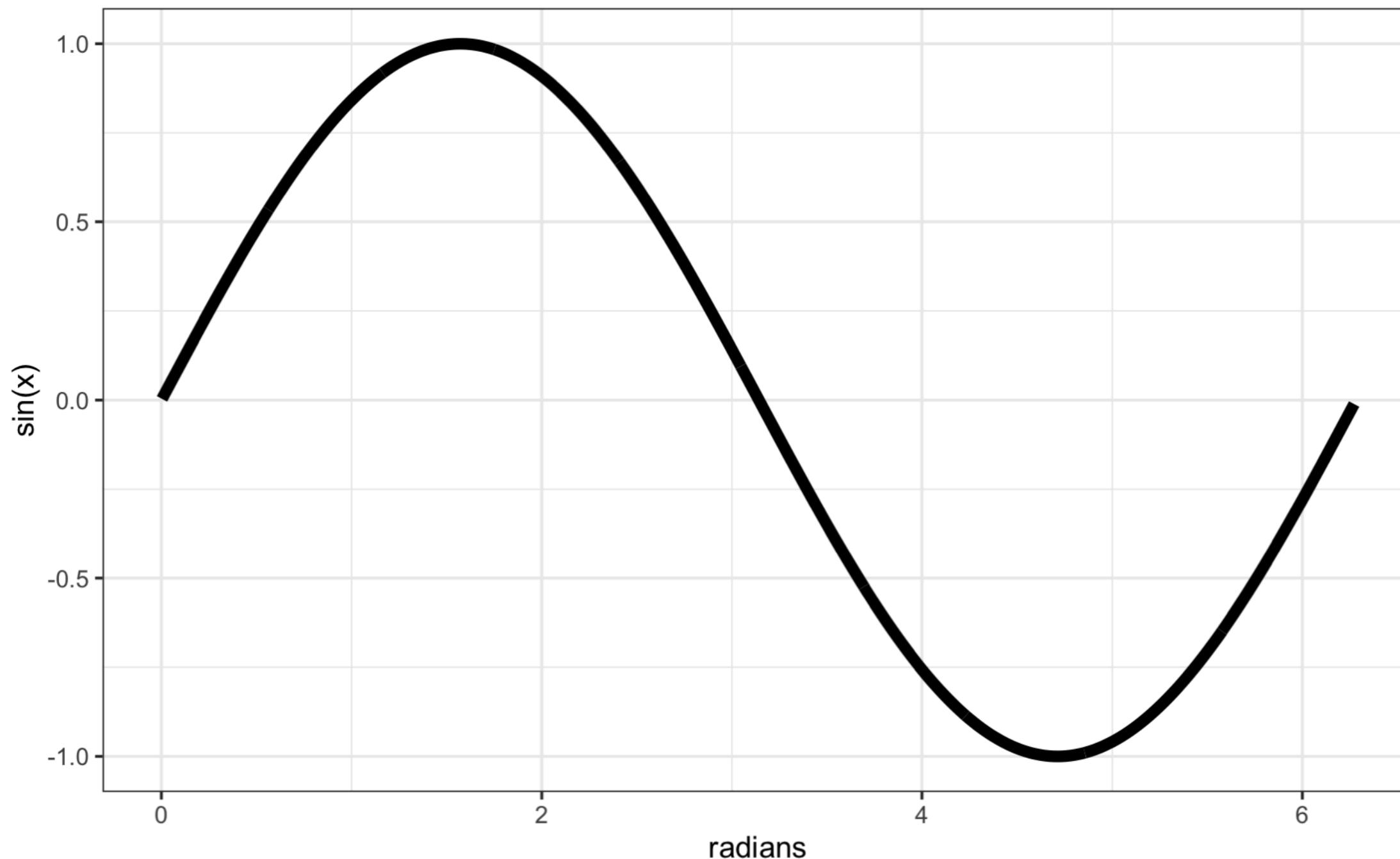
E



F

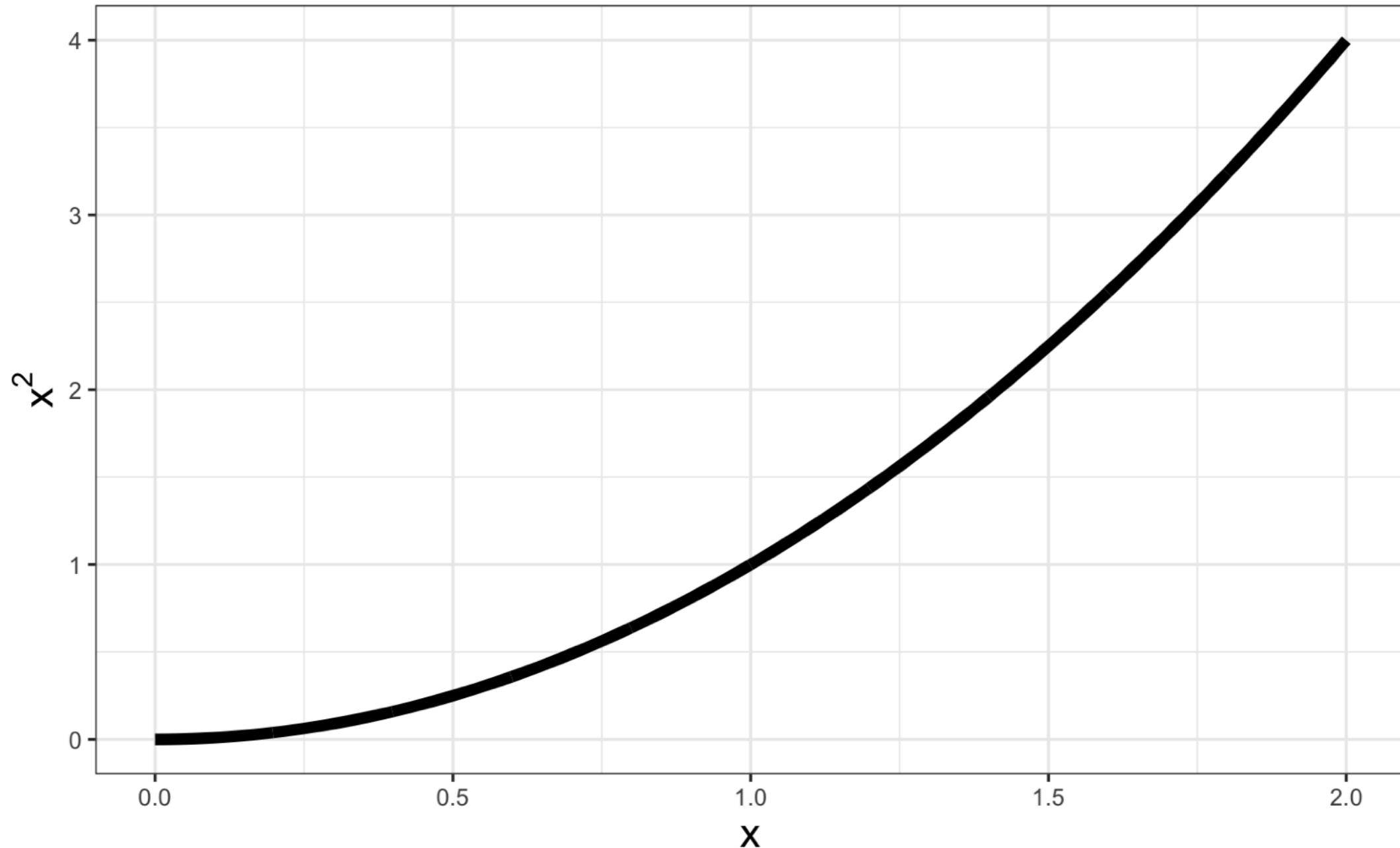


**9. Всегда когда считаем корреляцию на реальных данных и рассуждаем на основании этих корреляций надо строить графики.**



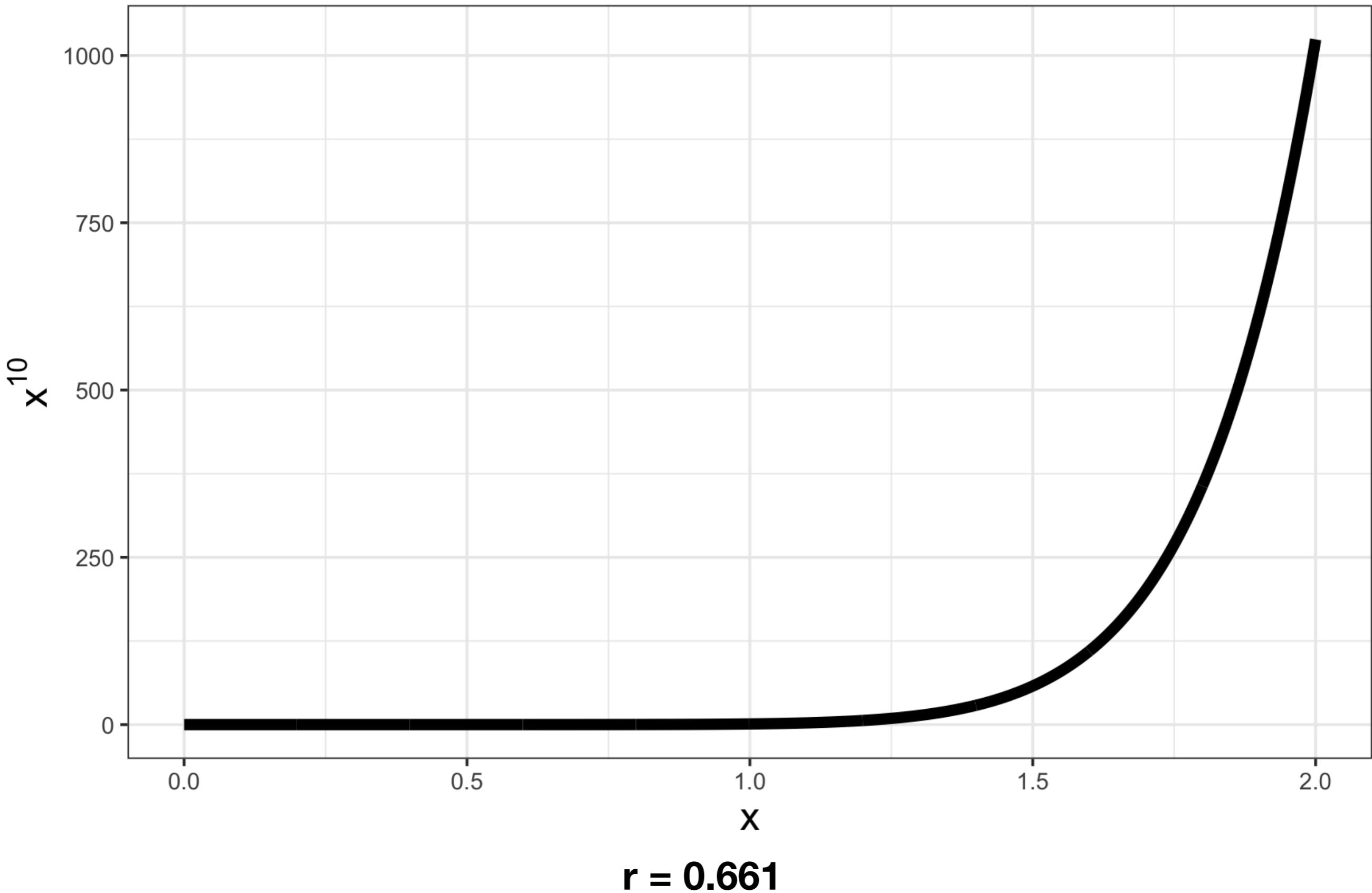
Берем  $x$  равномерно на отрезке  $[0; 2\pi]$ . Корреляция в среднем - -0.74

## 10. Плохо учитывает или учитывает иных зависимостей кроме линейных



$$r = 0.916$$

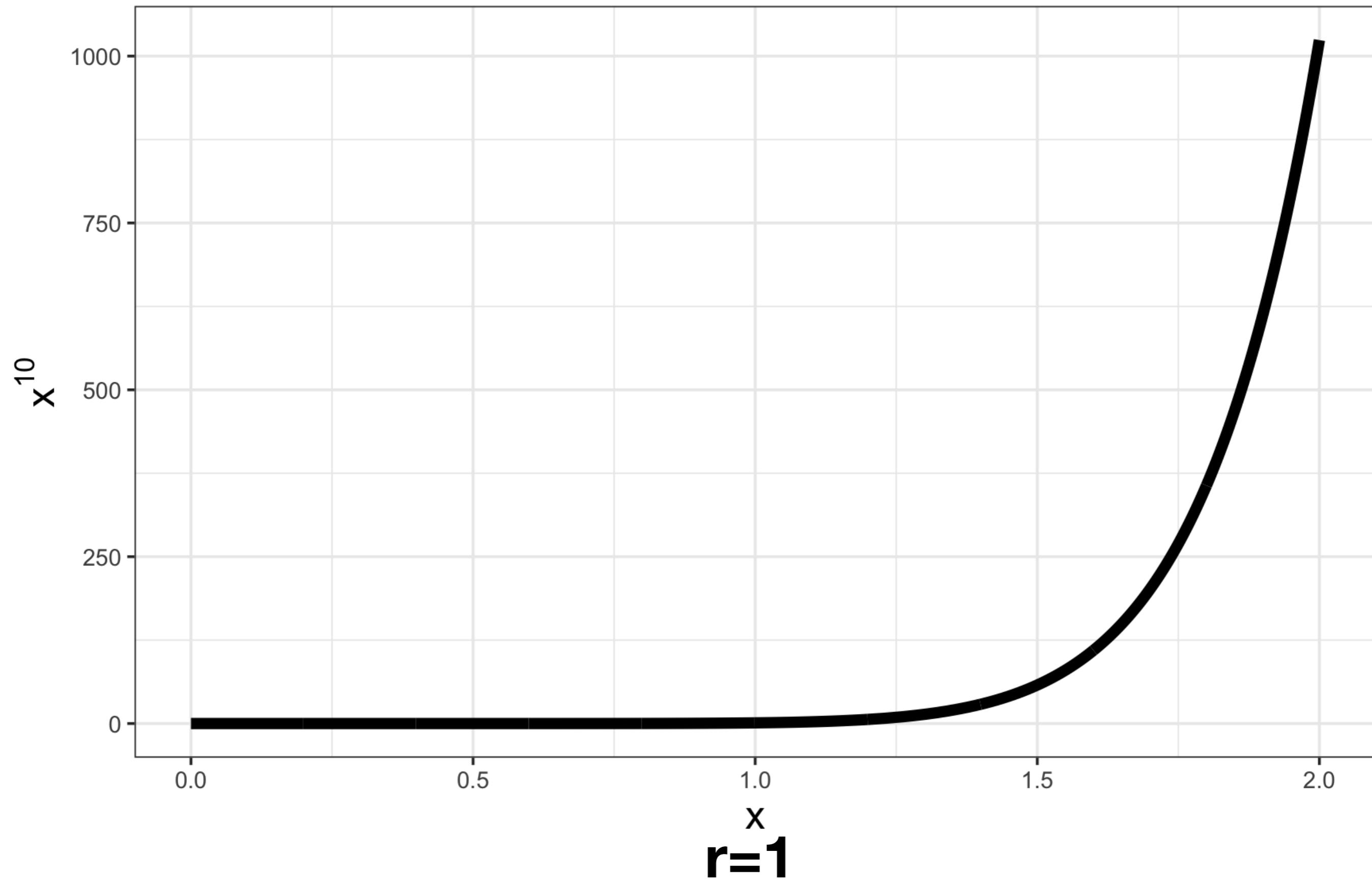
## 10. Плохо учитывает или учитывает иных зависимостей кроме линейных



# Корреляция Спирмена

**Считаем корреляцию Пирсона, но для рангов наблюдений. Учитывает монотонные зависимости**

# Корреляция Спирмена



У меня есть набор данных с целевой переменной  $y$  и 1000 признаков. Считаю корреляцию между  $y$  и каждым признаком. Нужна ли поправка на множественное тестирование?

У меня есть набор данных с целевой переменной  $y$  и 1000 признаков. Считаю корреляцию между  $y$  и каждым признаком. Нужна ли поправка на множественное тестирование?

Да

```
m <- matrix(rnorm(1001 * 100, mean = 0, sd=1), ncol=1001)
pvals <- sapply(1:1000, function(x){cor.test(m[, x], m[, 1001])$p.value })
sum(pvals < 0.05)

## [1] 47
```

# Задача регрессии

- 1) Есть матрица признаков объектов X (design matrix), каждой строке соответствует объект, каждому столбцу - переменная.

Obs	HospitalID	Sex	Cholesterol
1	2	Male	194
2	3	Female	200
3	0	Male	233
4	1	Female	192
5	2	Female	209
6	3	Female	200
7	0	Female	184
8	1	Female	228
9	2	Female	150
10	3	Male	221

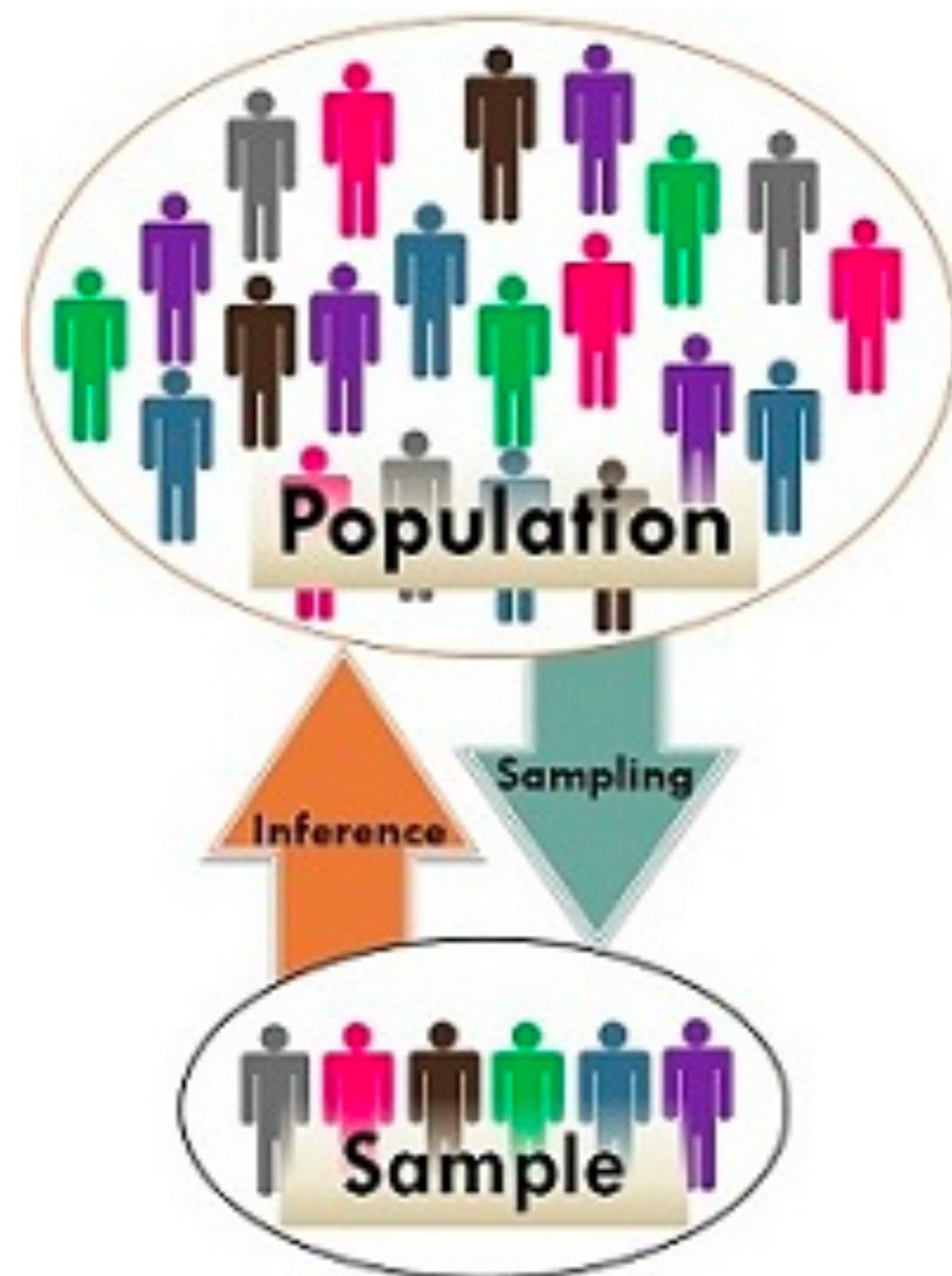
# Задача регрессии

Для каждого объекта имеем некую величину  $y$ , которая зависит от признаков этого объекта. Предполагаем, что существует некая функция  $y = f(x)$ , которая предсказывает  $y$  идеально.

Obs	HospitalID	Sex	Cholesterol	BP_Status
1	2	Male	194	Normal
2	3	Female	200	High
3	0	Male	233	High
4	1	Female	192	Optimal
5	2	Female	209	Normal
6	3	Female	200	High
7	0	Female	184	Normal
8	1	Female	228	High
9	2	Female	150	Normal
10	3	Male	221	Normal

# Задача регрессии

- 3) Мы не можем выбрать все существующие объекты (взять всю генеральную совокупность), взяли лишь выборку

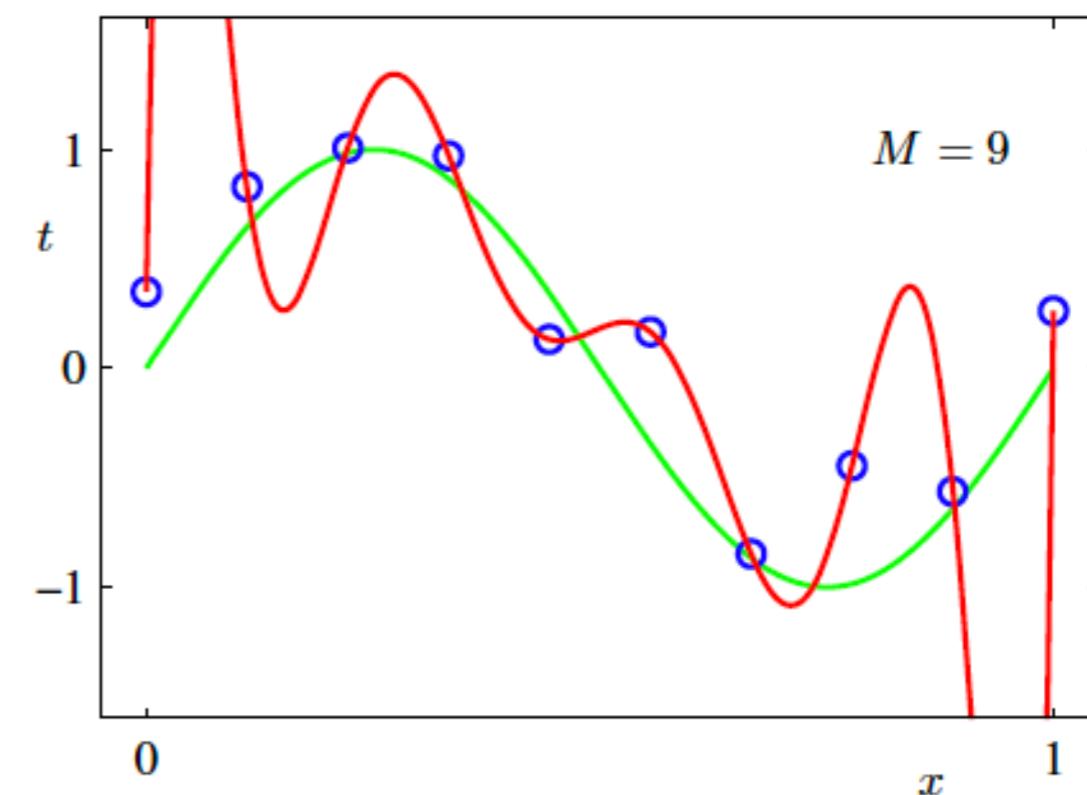
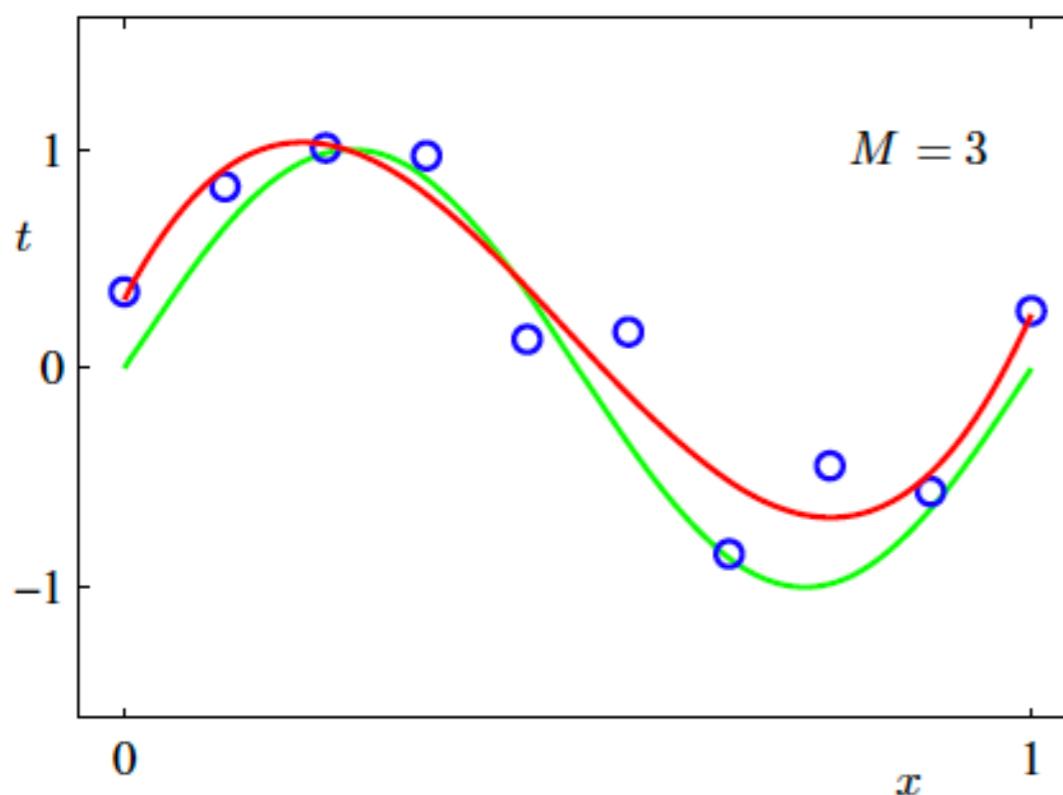
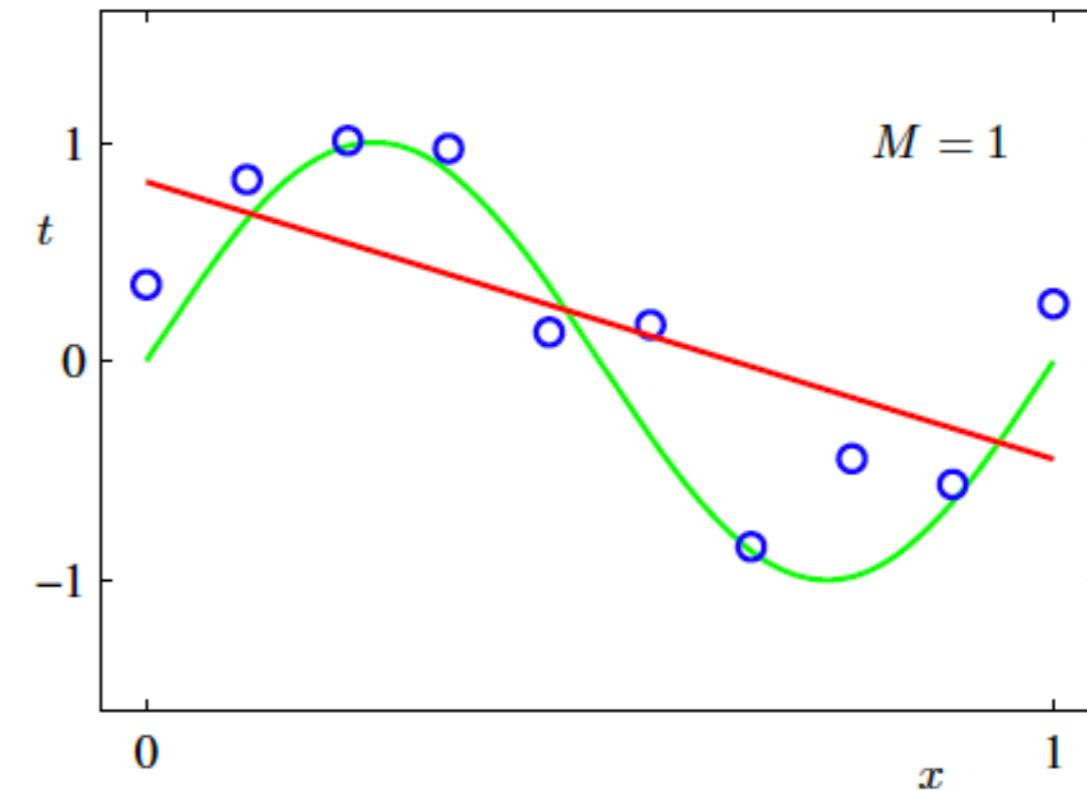
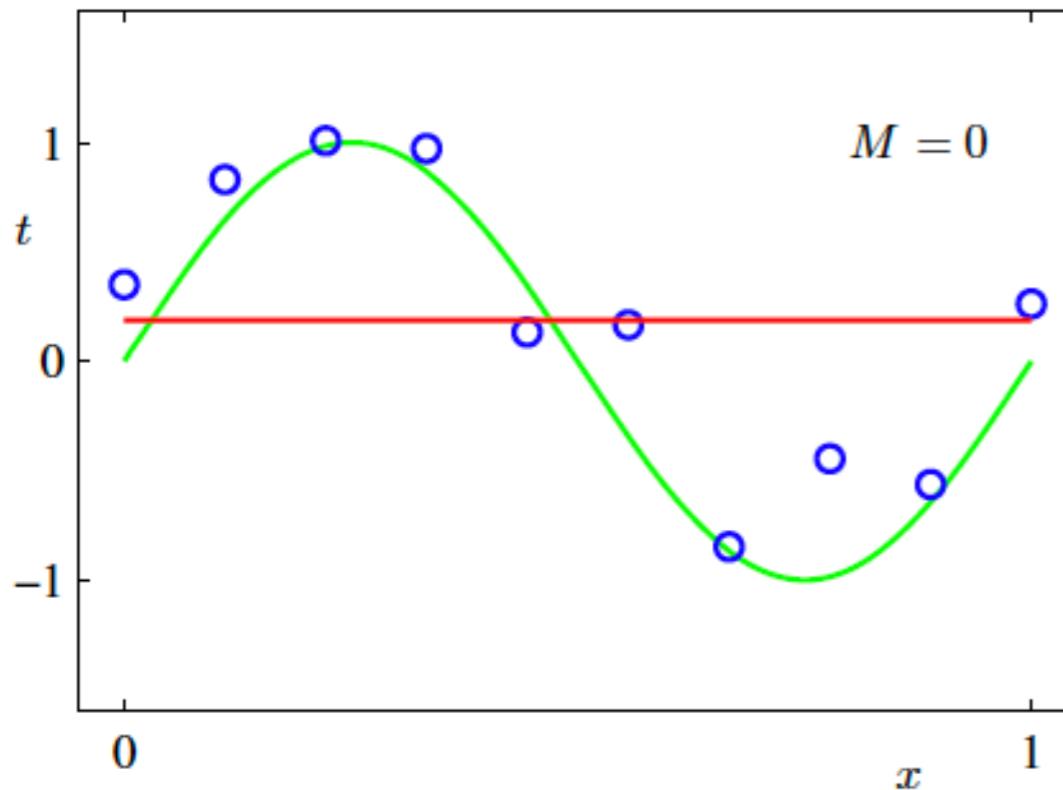


# Задача регрессии

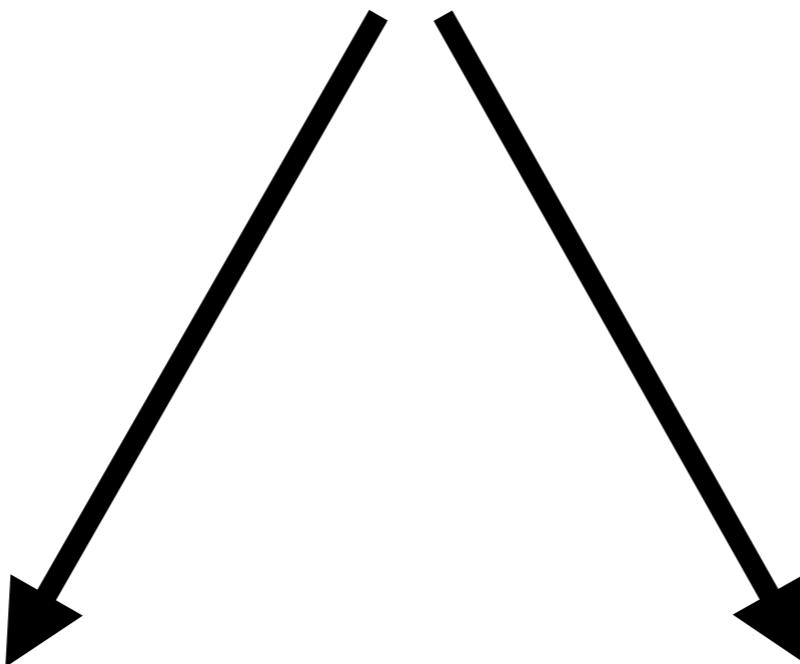
- 4) Мы измеряем  $y$  с некоторой ошибкой

$$y^* = y + \text{noise}$$

5) Потому мы не можем восстановить функцию  $f$ . Но мы можем пытаться ее аппроксимировать функцией из какого-то набора, которому, возможно,  $f$  принадлежит. Это и будем делать



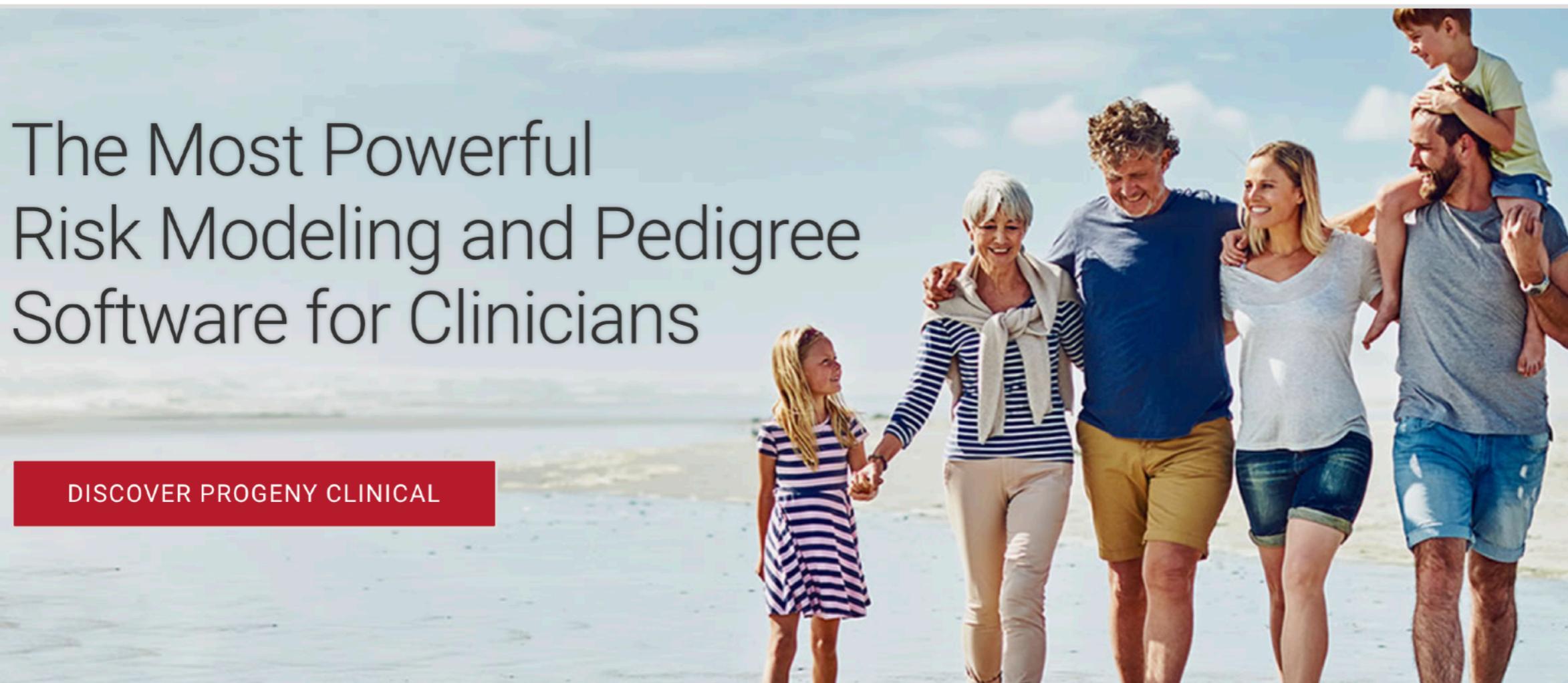
6) В зависимости от того, что мы хотим делать с полученной аппроксимацией, есть два направления:



**Предсказательные модели**, чем лучше предсказывает, тем лучше.  
Умеет предсказание объяснять -  
хорошо, но не критично

**Inference модели**. Они должны  
предсказывать хорошо (иначе  
смысла в них никакого нет), но  
между двумя моделями с разумным  
качеством отдается приоритет той  
модели, которая лучше объясняет  
свои предсказания.

# Предсказание пациентов с опухолями

[PRODUCTS](#)[SERVICES](#)[SUPPORT](#)[COMPANY](#)[CONTACT](#)[DISCOVER PROGENY CLINICAL](#)

## New High Risk Triage Screening Tools

See our quick screening tools to identify high risk patients for breast, colorectal and other cancers...

[See Features](#)

# Линейная регрессия

Будем искать аппроксимирующую функцию в виде.

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

Так как в измерении у есть ошибки, то

$$y = a \cdot x + b + N(0, 1)$$

# Линейная регрессия

Будем искать аппроксимирующую функцию в виде

Одномерный случай

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

Многомерный случай

$$y = a_1 \cdot x_1 + \dots + a_n \cdot x_n + b$$

# Mean Squared Loss

Остаток, residual

$$r_i = y_i - \hat{y}_i = y_i - h(x) = y_i - bx - a$$

Хотим минимизировать функцию

$$MSE = \frac{1}{N} \sum r_i^2$$

# Построение модели в R

```
# install.packages('datarium')
data("marketing", package = "datarium")  
Независимая переменная  
model <- lm(sales ~ youtube, data = marketing)  
model  
↑  
Предсказываемая величина  
↑  
Датасет
```

```
##  
## Call:  
## lm(formula = sales ~ youtube, data = marketing)  
##  
## Coefficients:  
## (Intercept)      youtube  
##             8.43911        0.04754  
Свободный коэффициент (a)
```

Коэффициент при переменной (b)

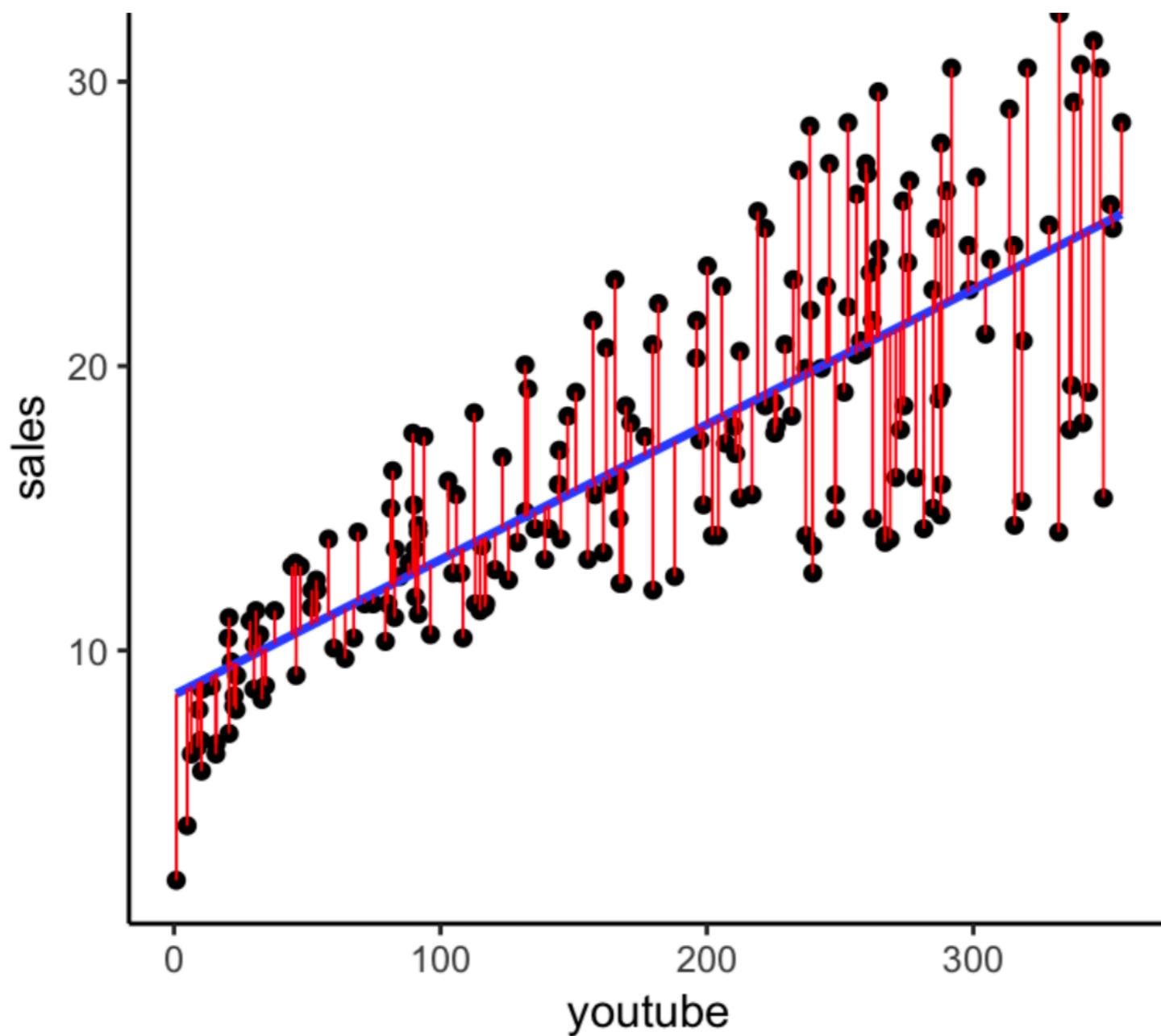
# Данные о модели

```
library(broom)
model.diag.metrics <- augment(model)
head(model.diag.metrics)
```

```
## # A tibble: 6 x 9
##   sales youtube .fitted .se.fit .resid     .hat .sigma .cooksdi .std.resid
##   <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
## 1 26.5     276.     21.6     0.385    4.96    0.00970   3.90   0.00794    1.27
## 2 12.5      53.4     11.0     0.431    1.50    0.0122    3.92   0.000920   0.387
## 3 11.2      20.6     9.42     0.502    1.74    0.0165    3.92   0.00169    0.449
## 4 22.2      182.     17.1     0.277    5.12    0.00501   3.90   0.00434    1.31
## 5 15.5      217.     18.8     0.297   -3.27    0.00578   3.91   0.00205   -0.839
## 6  8.64     10.4     8.94     0.525   -0.295   0.0180    3.92   0.0000534   -0.0762
```

# Residuals

```
ggplot(model.diag.metrics, aes(youtube, sales)) +  
  geom_point() +  
  stat_smooth(method = lm, se = FALSE) +  
  geom_segment(aes(xend = youtube, yend = .fitted), color = "red", size = 0.3)
```



# Много переменных

```
# install.packages('datarium')
data("marketing", package = "datarium")

model <- lm(sales ~ youtube + facebook + newspaper, data = marketing)
model
```

```
##
## Call:
## lm(formula = sales ~ youtube + facebook + newspaper, data = marketing)
##
## Coefficients:
## (Intercept)      youtube      facebook      newspaper
##       3.526667       0.045765      0.188530     -0.001037
```

# Статистическая значимость переменных

```
summary(model)
```

```
##  
## Call:  
## lm(formula = sales ~ youtube + facebook + newspaper, data = marketing)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -10.5932  -1.0690   0.2902   1.4272   3.3951  
##  
## Coefficients:  
##                 Estimate Std. Error t value Pr(>|t| )  
## (Intercept) 3.526667  0.374290  9.422 <2e-16 ***  
## youtube     0.045765  0.001395 32.809 <2e-16 ***  
## facebook    0.188530  0.008611 21.893 <2e-16 ***  
## newspaper   -0.001037  0.005871 -0.177    0.86  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 2.023 on 196 degrees of freedom  
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956  
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

**Значимость коэффициента  
(p-value гипотезы о том, что коэффициент равен 0)**



# R-squared

$$SS_{tot} = \sum_i (y_i - \bar{y})^2$$

$$SS_{reg} = \sum_i (y_i - \hat{y}_i)^2 = \sum_i r_i^2$$

$$R^2 = 1 - \frac{SS_{reg}}{SS_{tot}}$$

Коэффициент детерминации, в случае выполнения некоторых предположений, доля объясняемой **дисперсии**

Равен квадрату коэффициента корреляции Спирмена

# R-squared

Какие проблемы вы видите?

# R-squared

Какие проблемы вы видите?

$$SS_{reg} = \sum_i (y_i - \hat{y}_i)^2 = \sum_i r_i^2$$

**Чем лучше описываем наблюдения, тем лучше будет SSreg.**

# R-squared

Какие проблемы вы видите?

$$SS_{reg} = \sum_i (y_i - \hat{y}_i)^2 = \sum_i r_i^2$$

**Чем лучше описываем наблюдения, тем лучше будет SSreg.**

**Чем больше переменных, тем лучше описываем наблюдения**

# R-squared

Какие проблемы вы видите?

$$SS_{reg} = \sum_i (y_i - \hat{y}_i)^2 = \sum_i r_i^2$$

**Чем лучше описываем наблюдения, тем лучше будет SSreg.**

**Чем больше переменных, тем лучше описываем наблюдения**

**Чем больше переменных - тем лучше R-squared**

# Adjusted R-squared

$$R^2 = 1 - (1 - R^2) \frac{n - 1}{n - p - 1}$$

n - число наблюдений, p - число независимых переменных

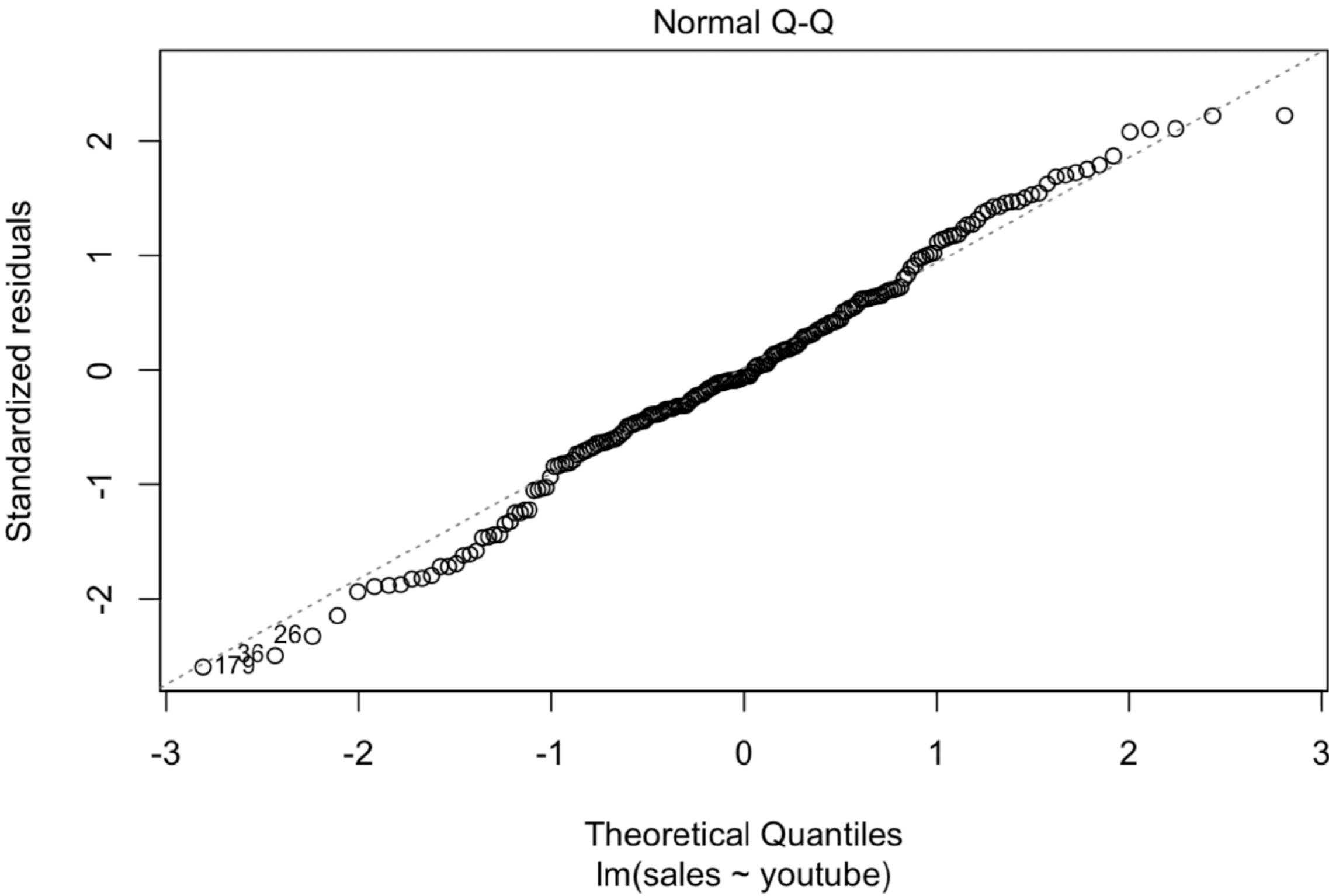
# Допущение линейной регрессии

- 1) Остатки распределены нормально
- 2) Дисперсия остатков не зависит от x (гомоскедастичность)
- 3) Остатки независимы (= наблюдения независимы)
- 4) Отсутствует мультиколлинеарность (независимые переменные линейно друг от друга не зависят)
- 5) Предсказываемая переменная зависит от независимых линейно

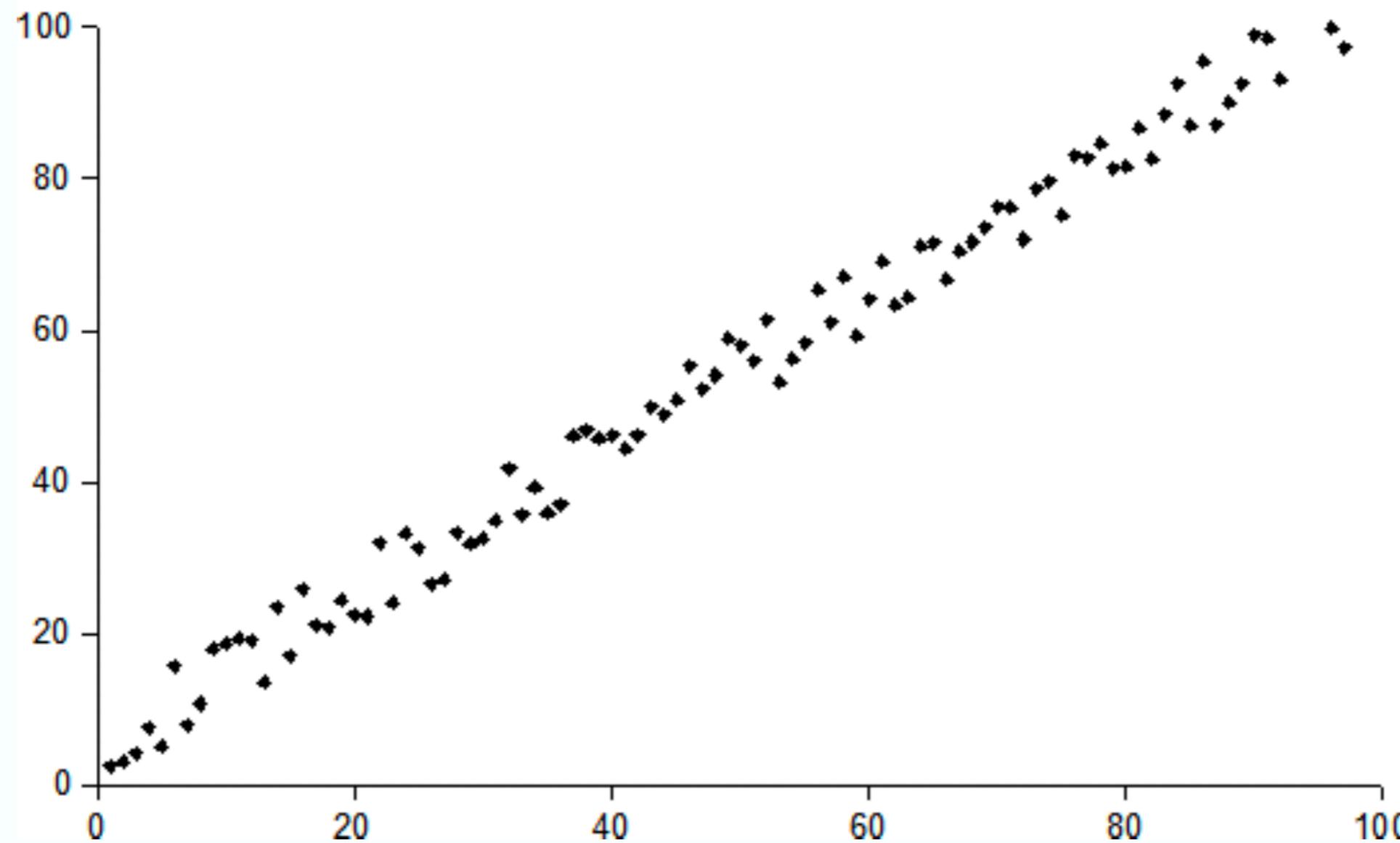
**Отчасти допущения помогает проверять функция `plot(model, <номер графика>)`,  
номер графика легче угадывать**

# 1) Остатки распределены нормально

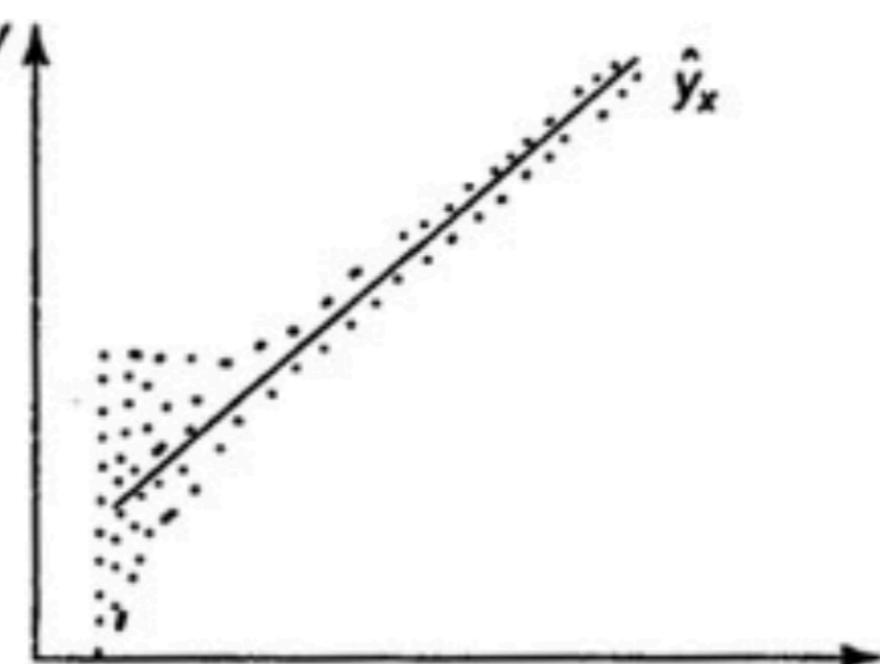
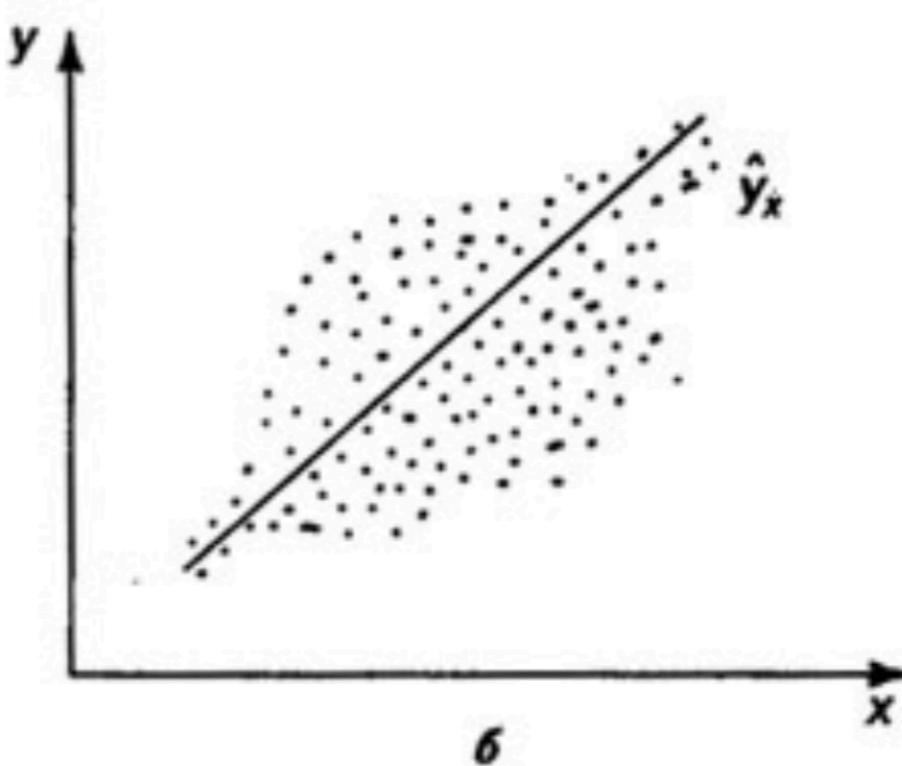
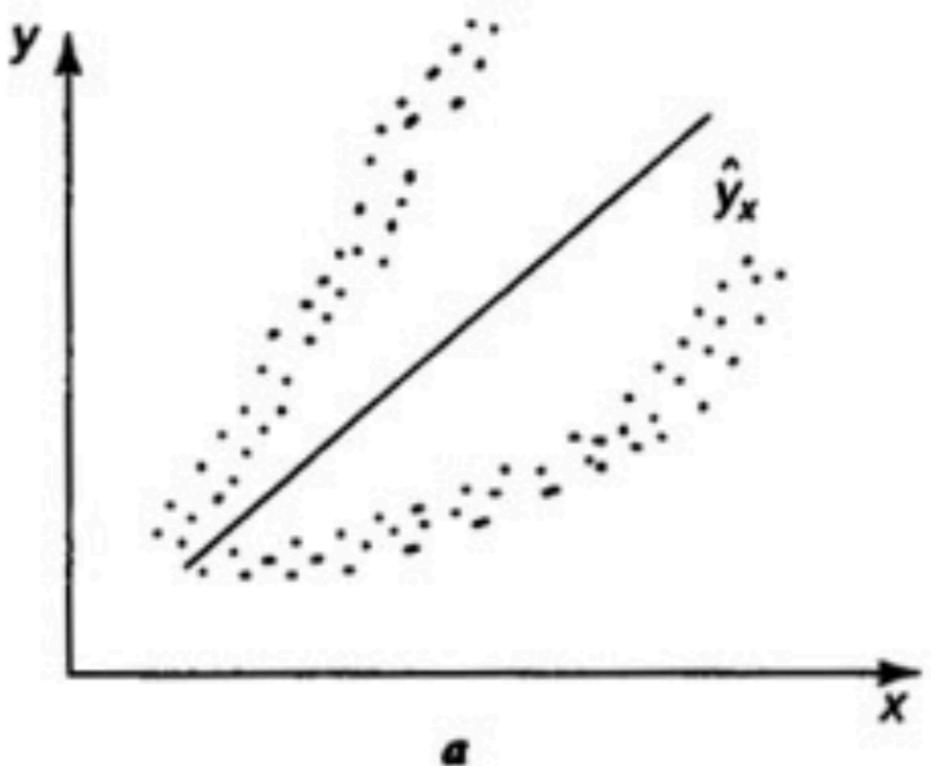
`plot(model, 2)`



## 2) Дисперсия остатков не зависит от x (гомоскедастичность)

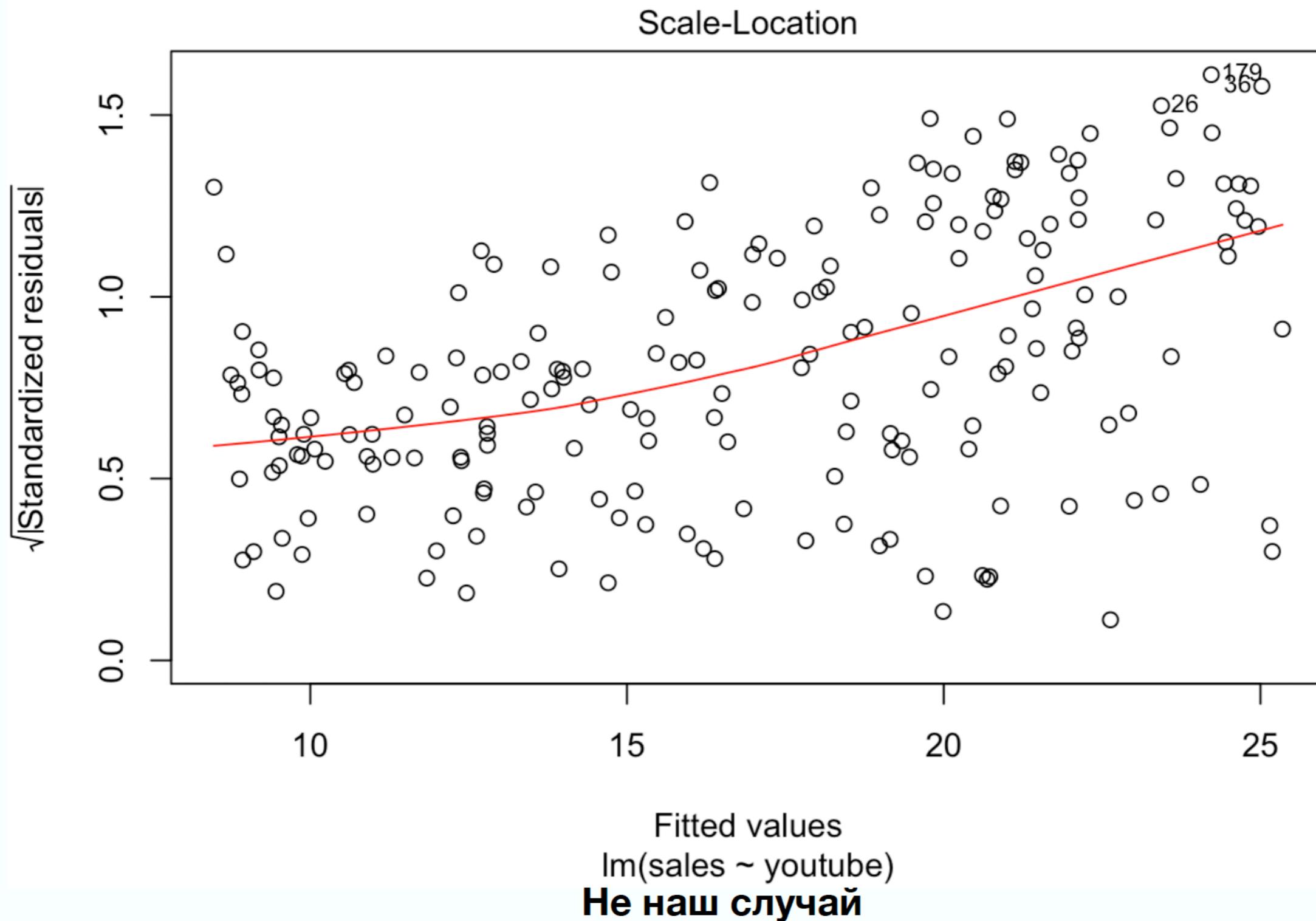


# Гетероскедастиность (heterogeneity of variance)



# Homogeneity of variance (гомоскедастичность)

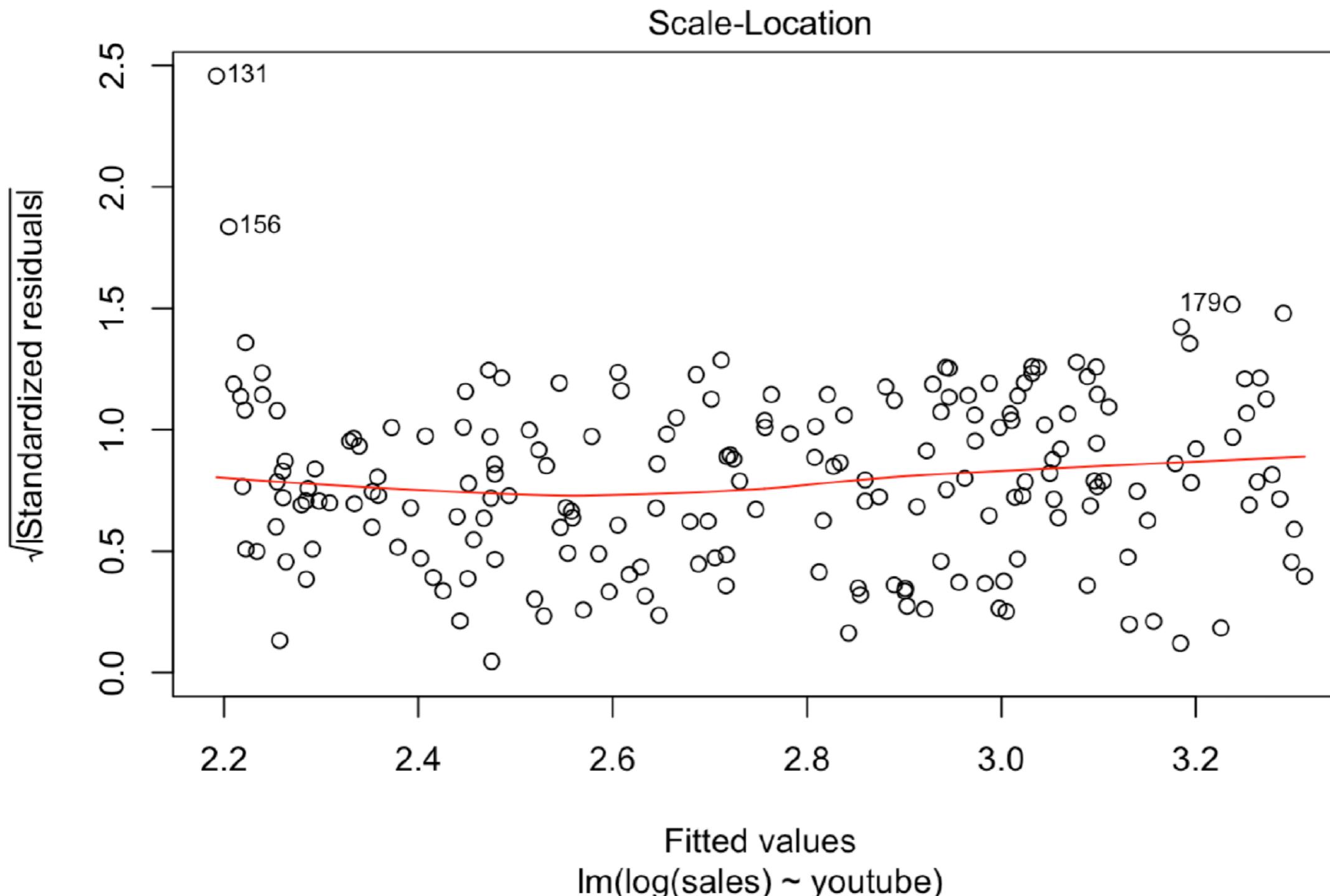
```
plot(model, 3)
```



# Homogeneity of variance (гомоскедастичность)

Иногда помогают преобразования (например, логарифмирование)

```
model2 <- lm(log(sales) ~ youtube, data = marketing)
plot(model2, 3)
```



- 4) Отсутствует мультиколлинеарность (независимые переменные линейно друг от друга не зависят)

## Мультиколлинеарность (строгая)

$$x_1 = 2x_2$$

$$y = a_1 \cdot x_1 + a_2 \cdot x_2 + b$$

$$y = (a_1 + 1) \cdot x_1 + (a_2 - 2) \cdot x_2 + b$$

**Равноправны и сводятся к**

$$y = (2 \cdot a_1 + a_2) \cdot x_2 + b$$

# Мультиколлинеарность (строгая)

$$x_1 = 2x_2$$

$$y = a_1 \cdot x_1 + a_2 \cdot x_2 + b$$

$$y = (a_1 + 1) \cdot x_1 + (a_2 - 2) \cdot x_2 + b$$

Равноправны и сводятся к

$$y = (2 \cdot a_1 + a_2) \cdot x_2 + b$$

Обычно - просто ошибка экспериментатора

# Мультиколлинеарность (нестрогая)

- 1) Переменные, измеряющие примерно одно и то же:  
(давление в начале и в конце дня )
- 2) Переменные, естественно связанные друг с другом

**Приводит к высоким стандартным ошибкам коэффициентов при переменных:**

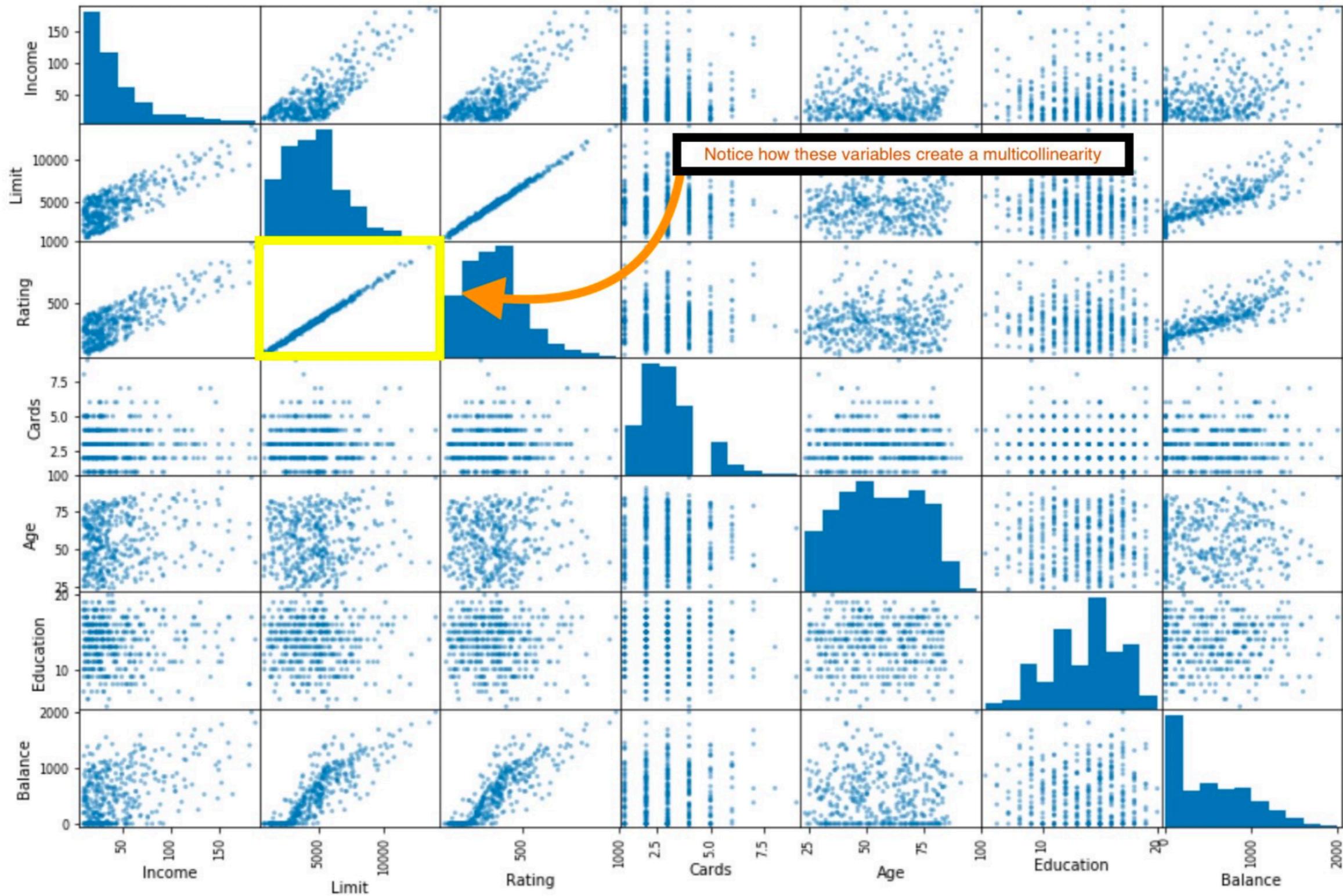
- 1) неустойчивость модели (несколько добавленных объектов резко меняют оценки)**
- 2) часть коэффициентов оценивается как незначимая**
- 3) На предсказательную силу модели почти не влияет**

# Мультиколлинеарность (нестрогая)

Типичное проявление - по отдельности группа коэффициентов незначима, но если выкинуть все переменные из этой группы, то качество модели падает

# Мультиколлинеарность

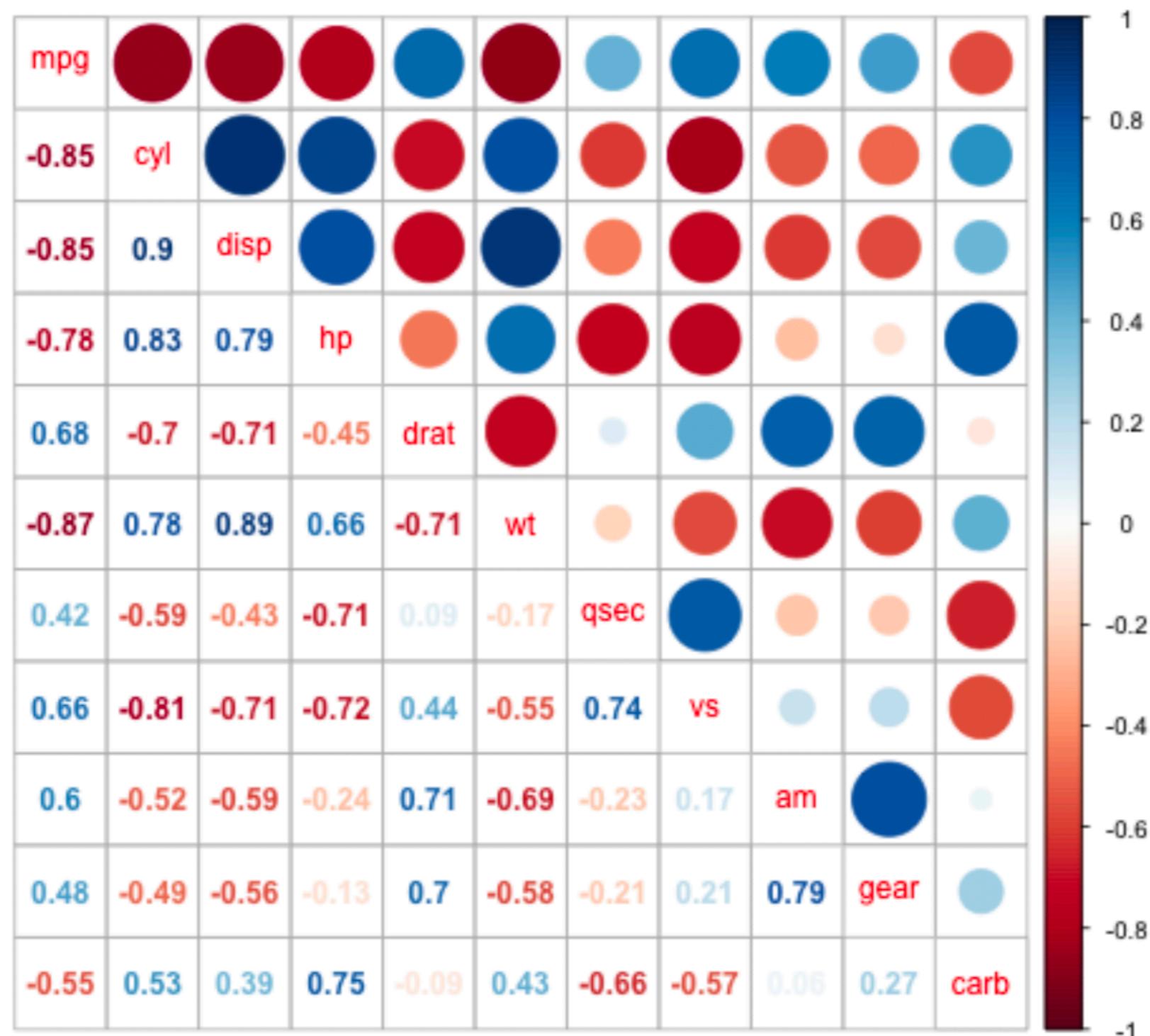
Построить зависимости между переменными



# Мультиколлинеарность

Посчитать корреляцию между переменными

<https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html>



# Мультиколлинеарность

Сделать вспомогательные регрессии - пытаемся предсказать каждую переменную і через другие. У каждой модели смотрим R\_squared.

$$V_i = \frac{1}{1 - R_i^2}$$

## Коэффициент вздутия

Если он большой, например, больше 10, значит, соответствующая переменная линейно связана с другими

# Мультиколлинеарность

$$V_i = \frac{1}{1 - R_i^2}$$

```
library(car)
```

```
## Loading required package: carData
```

```
model <- lm(mpg ~ disp + hp + wt + drat, data = mtcars)
vif(model)
```

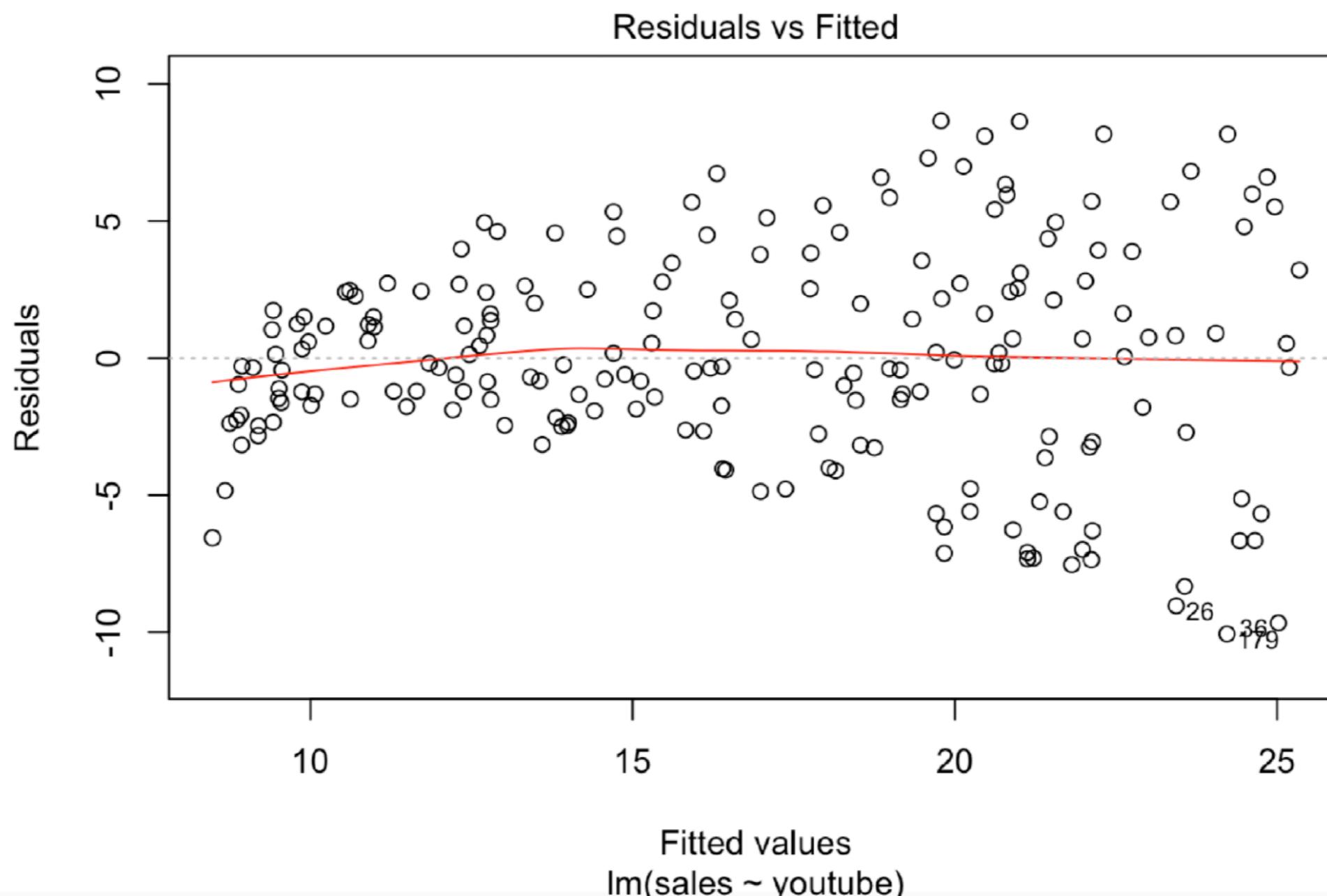
```
##      disp          hp          wt          drat
## 8.209402  2.894373  5.096601  2.279547
```

5)

Предсказываемая переменная зависит от независимых  
линейно

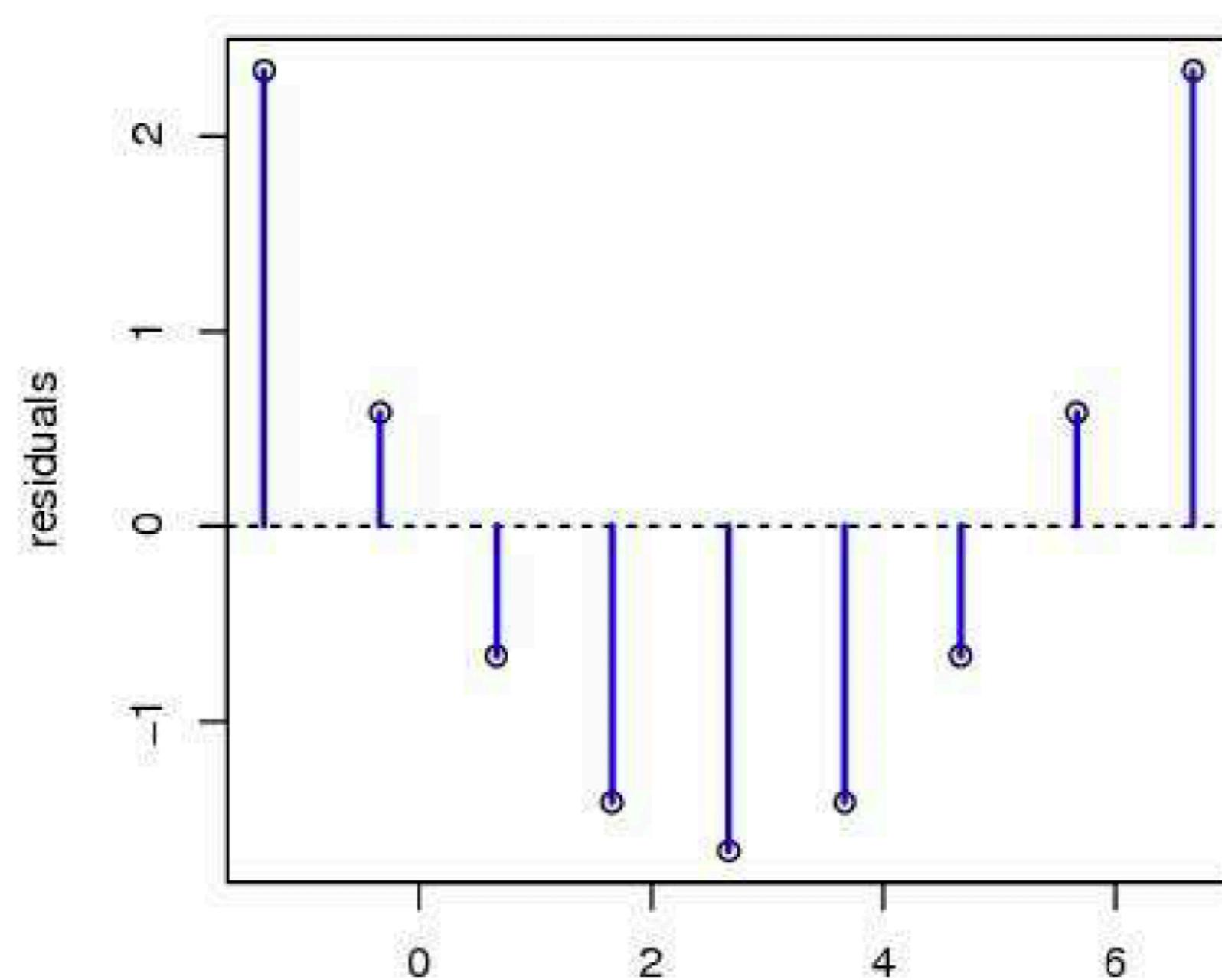
# Проверка на линейность

```
model <- lm(sales ~ youtube, data = marketing)
plot(model, 1)
```

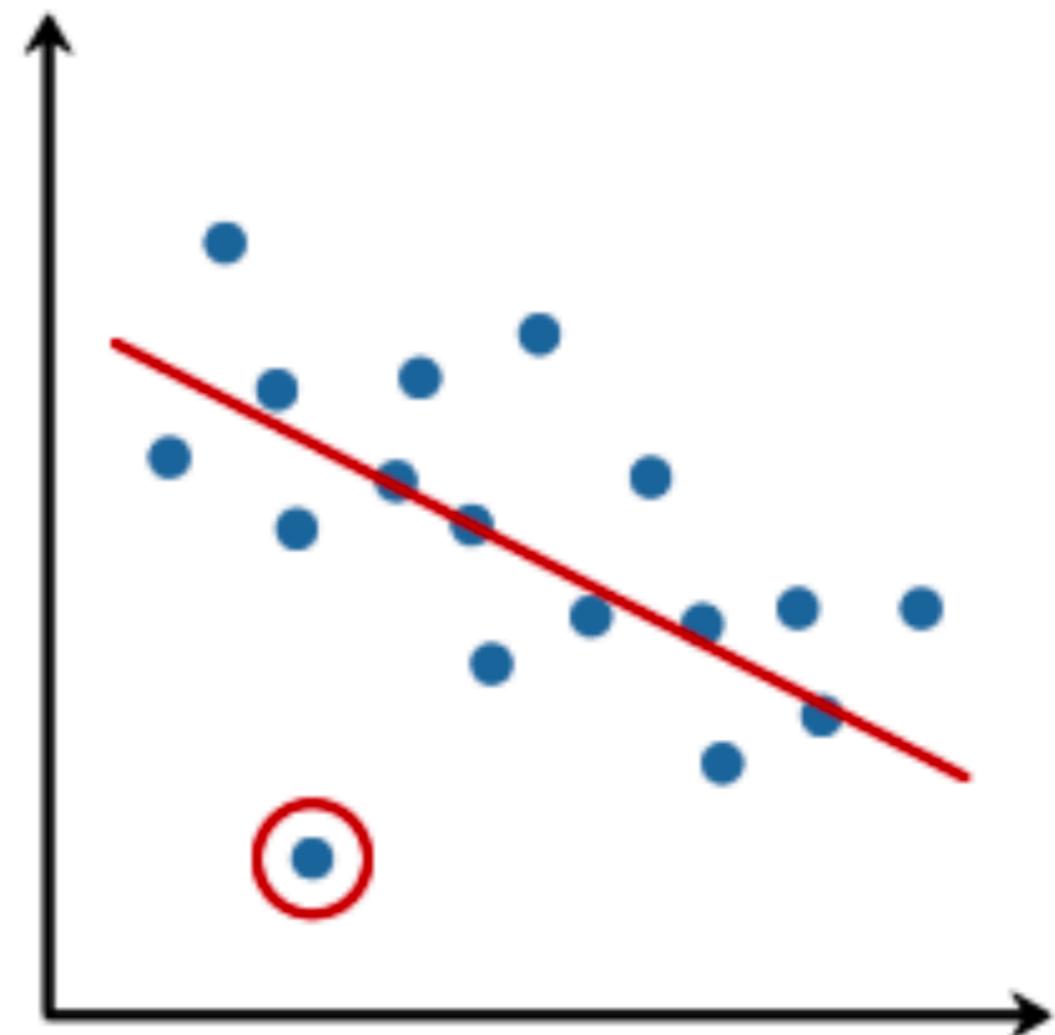
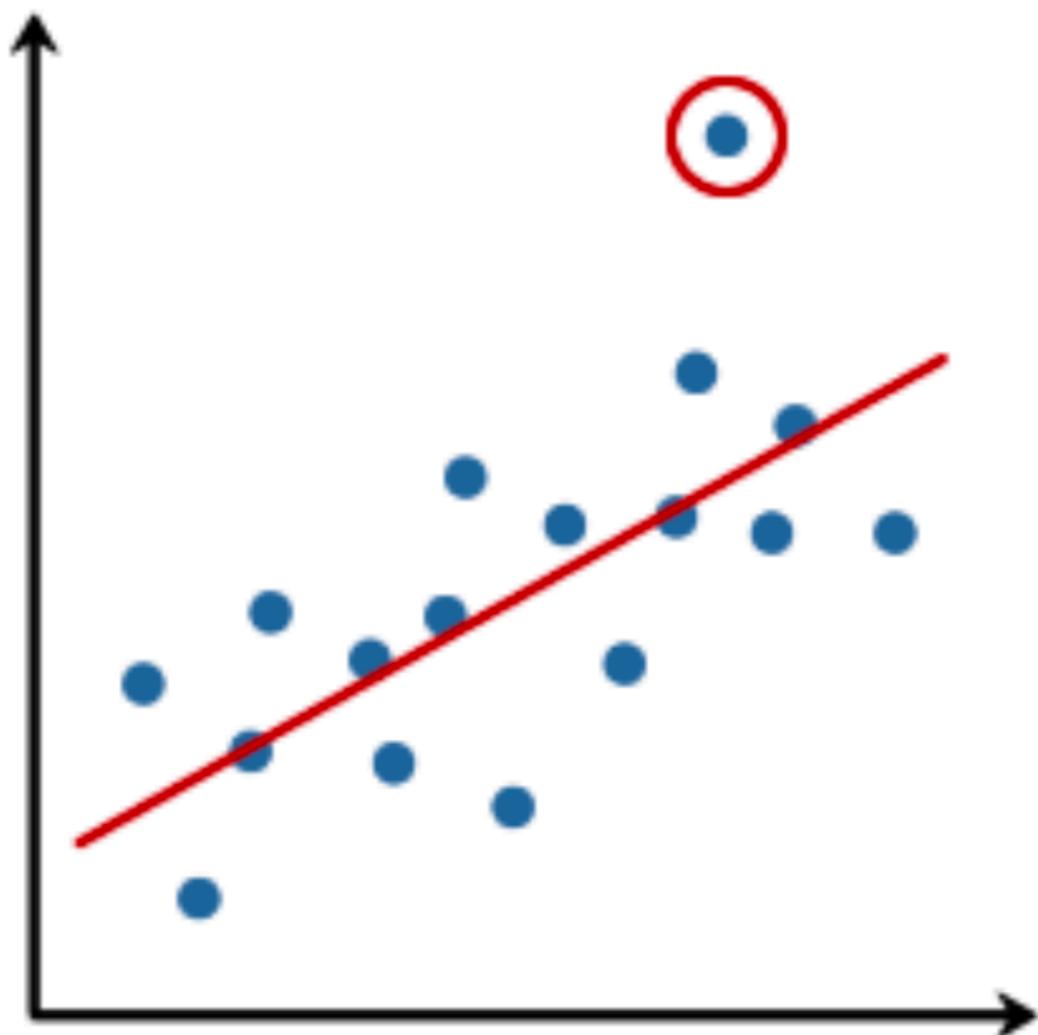


5) Предсказываемая переменная зависит от независимых линейно

## Пример нелинейности

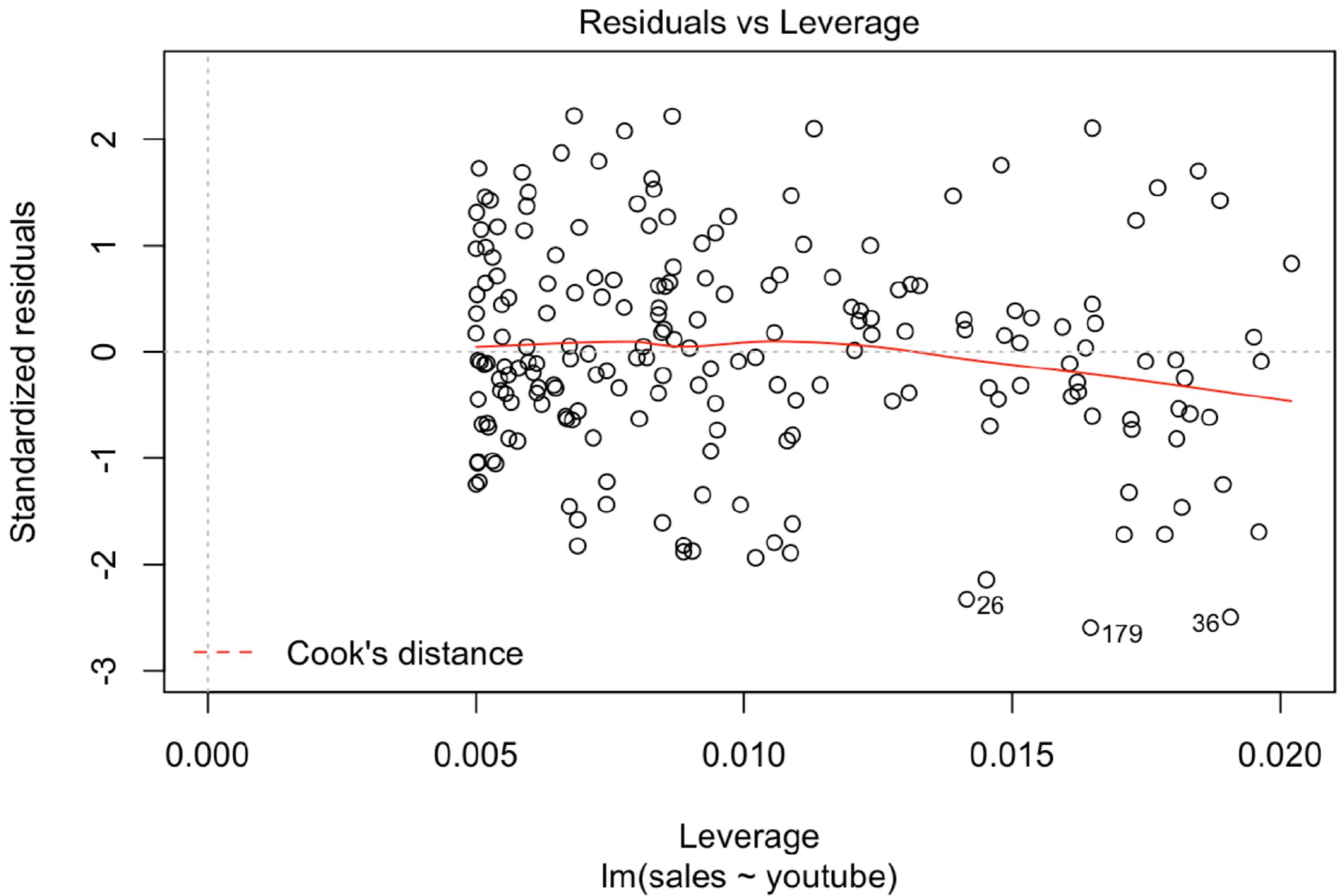


# Выбросы



```
plot(model, 5)
```

**Если отклонение больше 3, то имеем основание подозревать, что точка - outlier**



У меня есть набор данных с целевой переменной  $y$  и 1000 признаков. Предсказываю  $y$  по признакам. Далее отбираю самые 5 самых значимых. Есть ли подвох?