

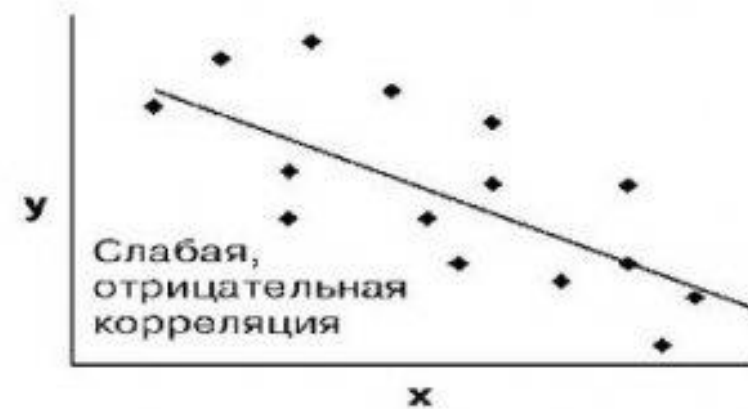
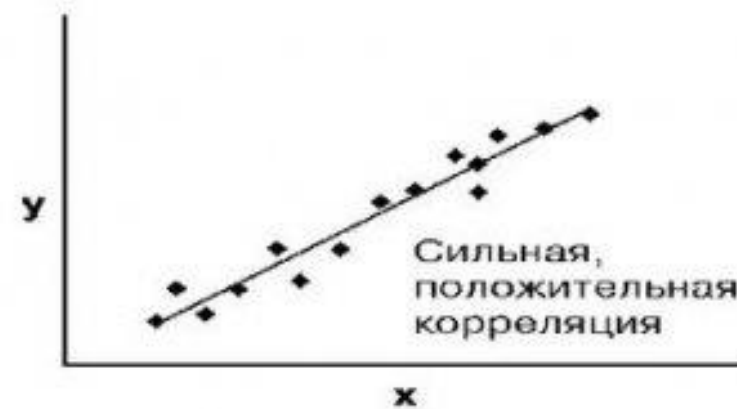
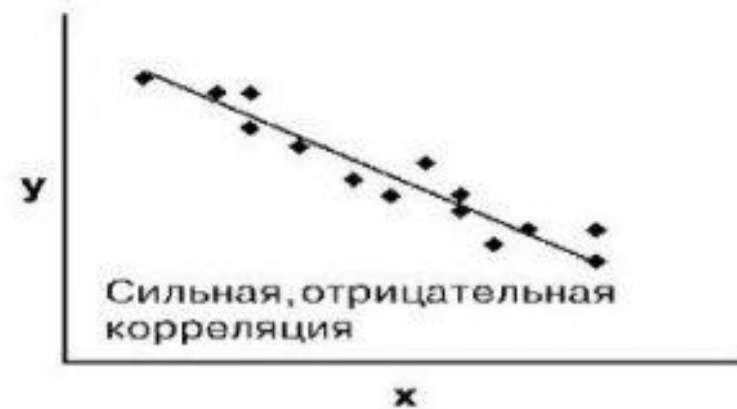
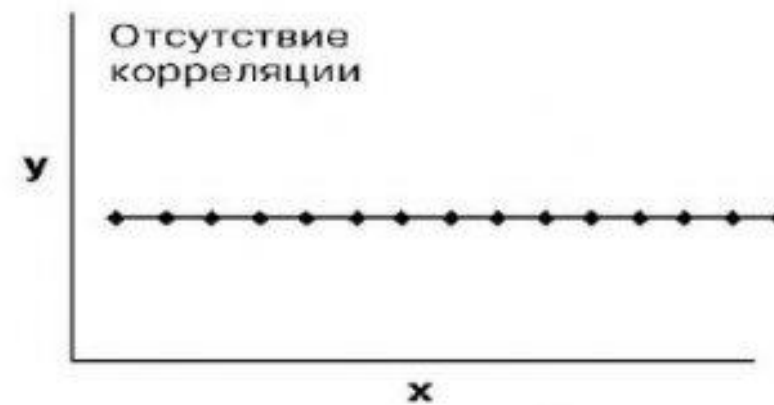


Корреляция и регрессия



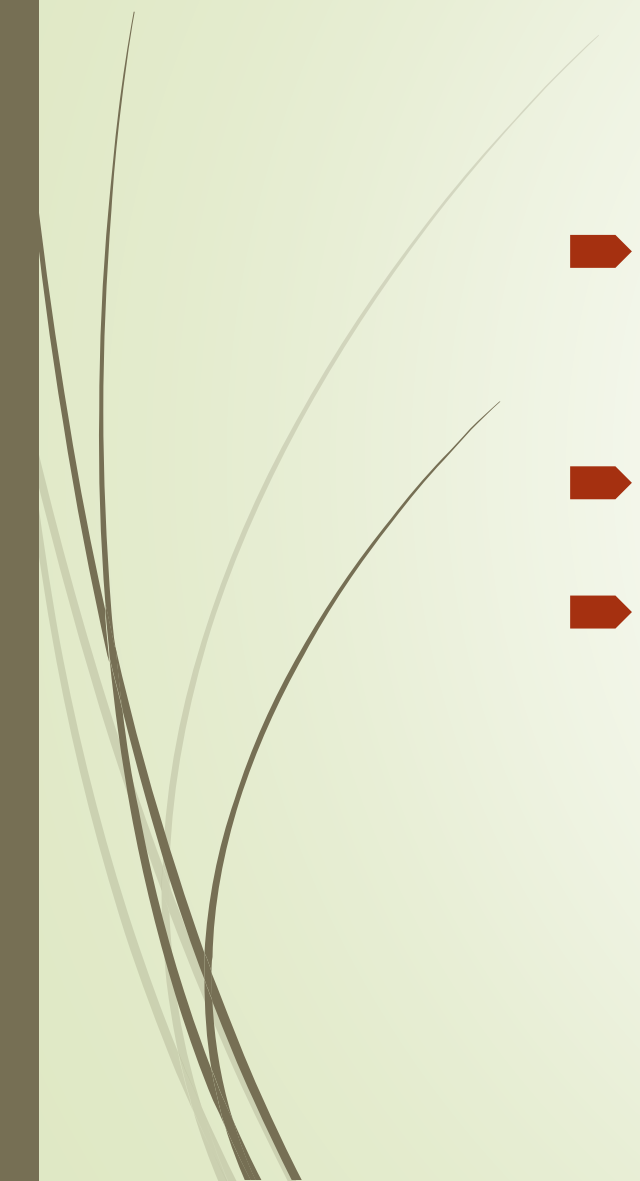
Как связаны 2 переменные?

- Масса тела и рост
- Экспрессия двух генов
- Концентрация антибиотика и количество погибших клеток



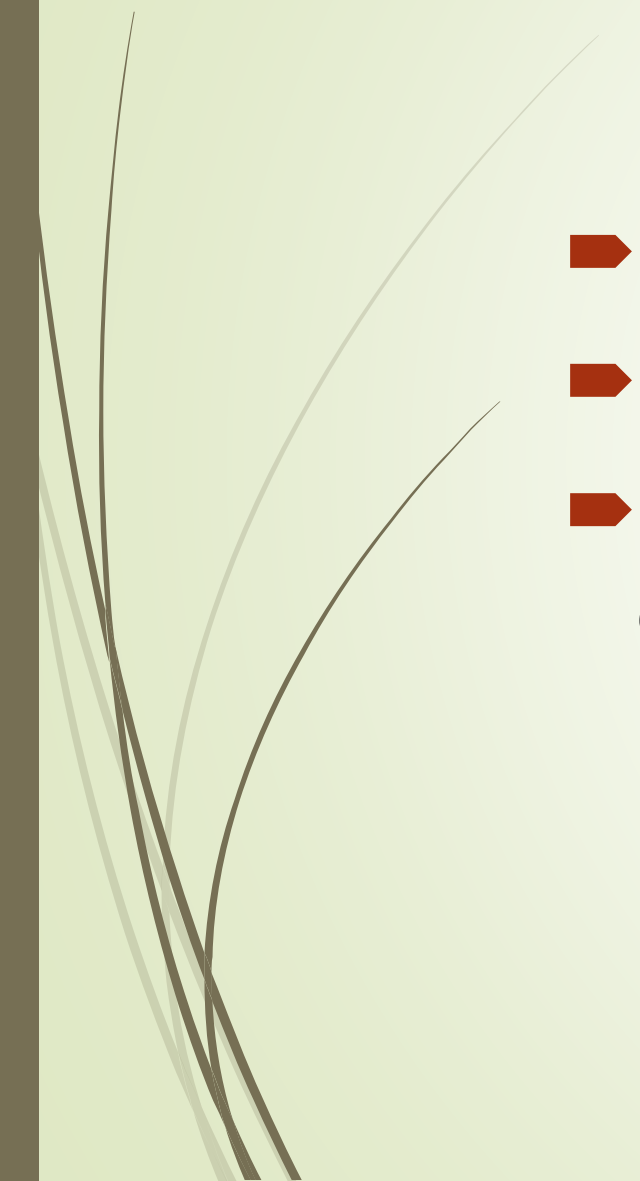


Свойства корреляции

- Форма – прямолинейная или криволинейная
 - Направление – прямая или обратная
 - Степень – сильная, средняя, слабая
- 



Коэффициент корреляции

- Может принимать значения от -1 до 1
 - Знак определяет направление связи
 - Абсолютная величина определяет степень корреляции между двумя переменными
- 

Как выбрать статистический тест

Сравнение с теоретическим средним	НР: Т-тест	
	ННР: тест Уилкоксона	
Сравнение 2 групп	Не парное сравнение	НР: непарный Т-тест ННР: тест Манна-Уитни
	Парное сравнение	НР: парный Т-тест ННР: тест Уилкоксона
Сравнение 3 и более групп	Не повторные измерения	НР: однофакторный дисперсионный анализ +post-hoc ННР: тест Крускала-Уоллиса +post-hoc
	Повторные измерения	НР: дисперсионный анализ с повторными измерениями +post-hoc ННР: тест Фридмана +post-hoc
Корреляция	НР: корреляция Пирсона	
	ННР: корреляция Спирмана	
Предсказание	На основе 1 переменной	НР: линейная регрессия ННР: непараметрическая регрессия
	На основе нескольких переменных	Множественная линейная и нелинейная регрессии



Коэффициент корреляции Пирсона

```
corr1 <- data %>%  
  group_by(Group) %>%  
  cor_test(Gene1_expression, Gene2_expression,  
method = "pearson")
```




Коэффициент корреляции Спирмана

```
corr2 <- data %>%  
  filter(Group == "Control") %>%  
  cor_test(Gene1_expression, Gene2_expression,  
method = "spearman")
```



Корреляционные матрицы

rowname	BRCA1	BRCA2	ATM	TP53	NBN
BRCA1	1	0.38	0.15	-0.044	0.35
BRCA2	0.38	1	0.43	0.015	0.48
ATM	0.15	0.43	1	0.037	0.32
TP53	-0.044	0.015	0.037	1	-0.0072
NBN	0.35	0.48	0.32	-0.0072	1

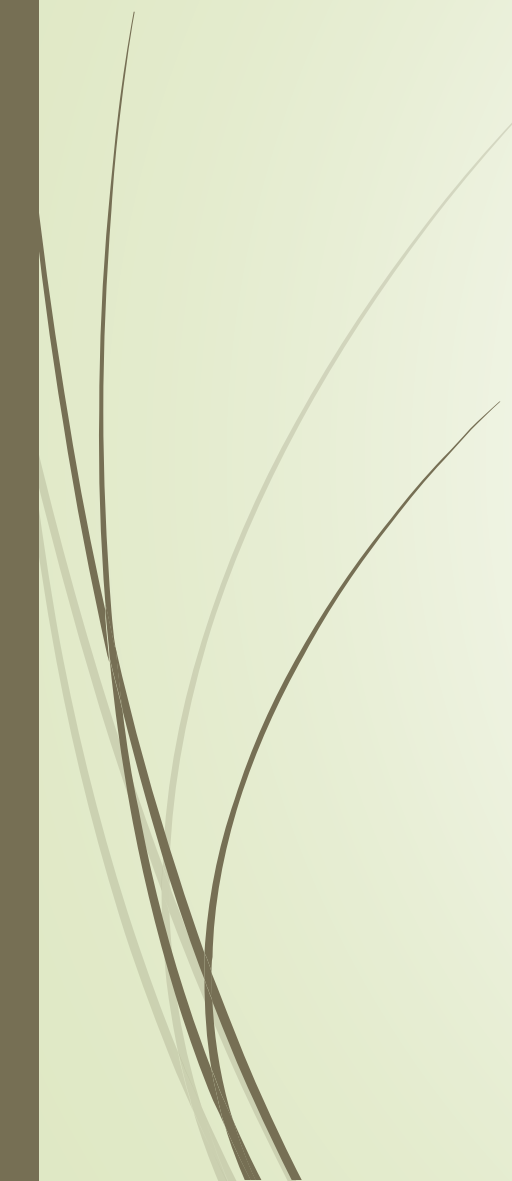
rowname	BRCA1	BRCA2	ATM	TP53	NBN
BRCA1	*	.			.
BRCA2	.	*	.		.
ATM		.	*		.
TP53				*	
NBN	.	.	.		*

rowname	BRCA1	BRCA2	ATM	TP53	NBN
BRCA1	0	2.37e-35	1.93e-06	0.17	1.35e-29
BRCA2	2.37e-35	0	7.12e-45	0.639	5.24e-58
ATM	1.93e-06	7.12e-45	0	0.244	1.45e-24
TP53	0.17	0.639	0.244	0	0.821
NBN	1.35e-29	5.24e-58	1.45e-24	0.821	0

< 0.25 = “ “
 $0.25-0.5$ = “.”
 $0.5-0.75$ = “+”
 > 0.75 = “*”



Корреляционные матрицы

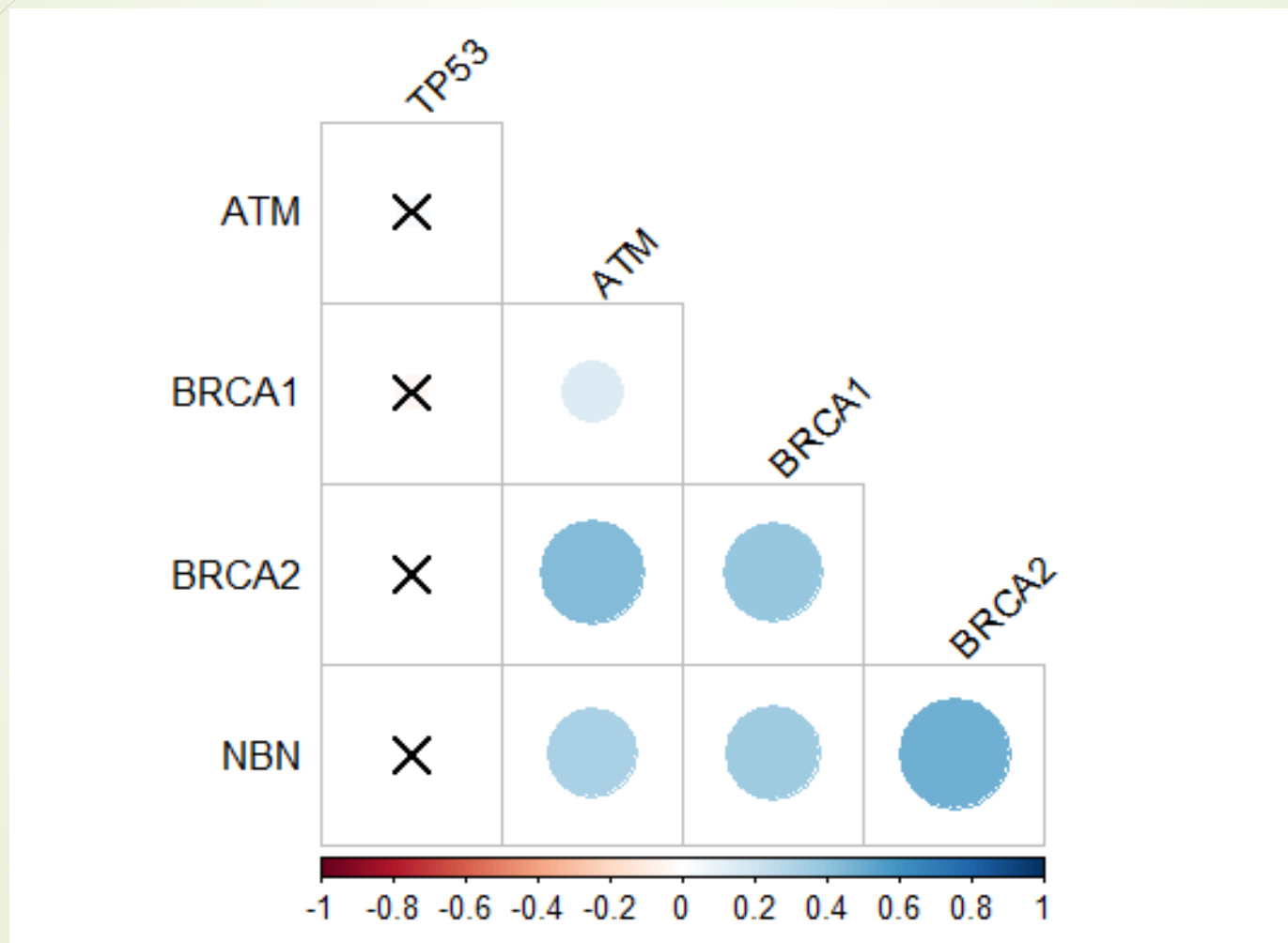


```
cor.mat <- data %>%  
  select(BRCA1, BRCA2, ATM, TP53, NBN) %>%  
  cor_mat()
```

```
cor.mat %>%  
  cor_get_pval()
```

```
cor.mat %>%  
  cor_as_symbols() %>%  
  pull_lower_triangle()
```

Коррелограммы





Коррелограммы

`cor.mat %>%`

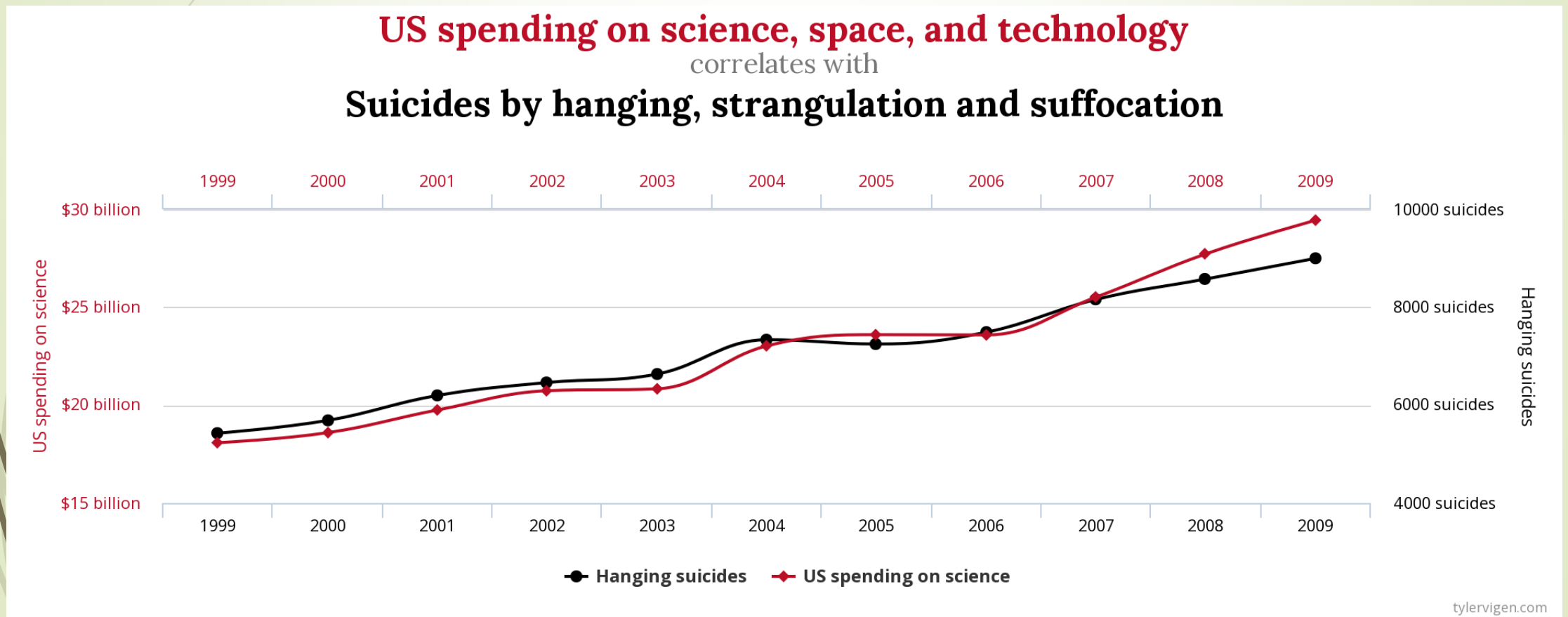
`cor_reorder() %>%`

`pull_lower_triangle() %>%`

`cor_plot()`

Корреляция совершенно не подразумевает
наличие причинно-следственной связи!

<http://tylervigen.com/spurious-correlations>



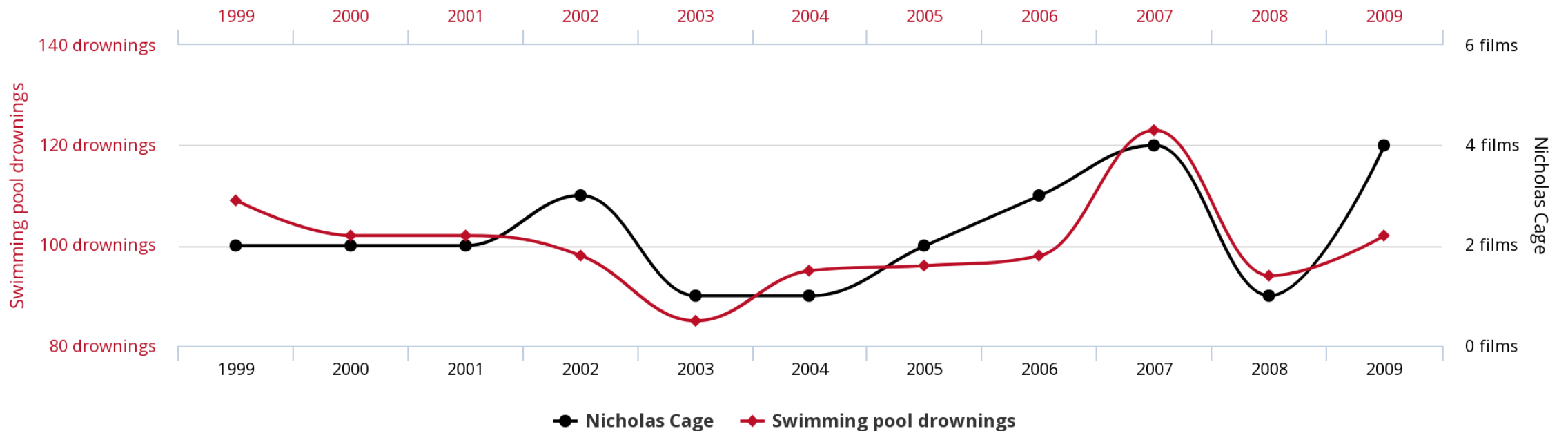
Корреляция совершенно не подразумевает наличие причинно-следственной связи!

<http://tylervigen.com/spurious-correlations>

Number of people who drowned by falling into a pool

correlates with

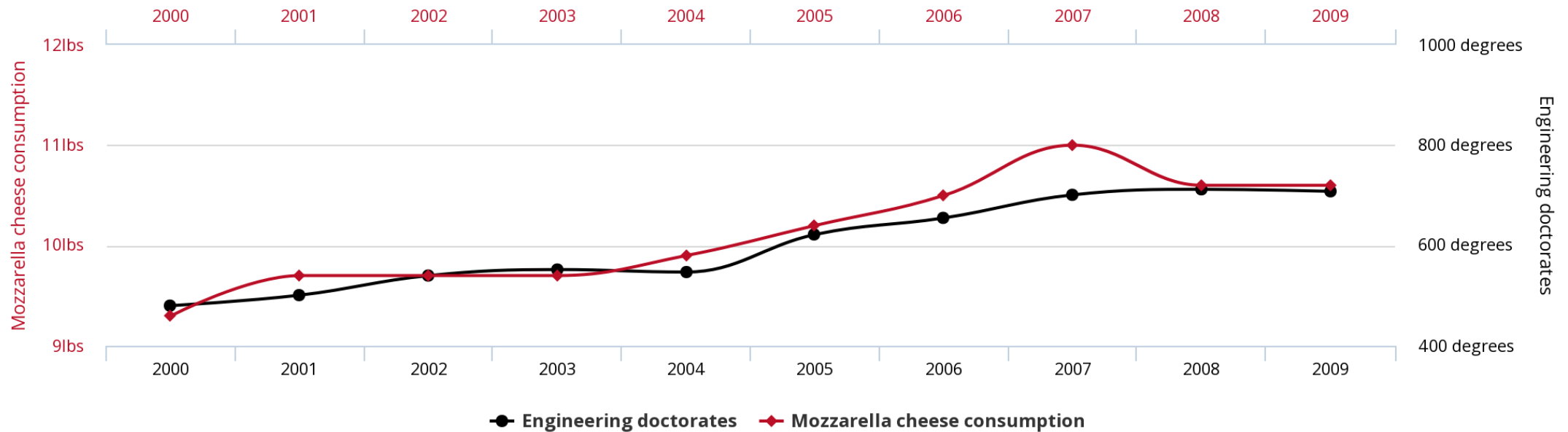
Films Nicolas Cage appeared in



Корреляция совершенно не подразумевает наличие причинно-следственной связи!

<http://tylervigen.com/spurious-correlations>

Per capita consumption of mozzarella cheese
correlates with
Civil engineering doctorates awarded



Предсказание?

Регрессионный анализ – инструмент для количественного предсказания значения одной переменной на основании другой.

Даёт нам правила, определяющие линию регрессии, которая ЛУЧШЕ ДРУГИХ предсказывает одну переменную на основании другой.

По оси Y располагают переменную, которую мы хотим предсказать (зависимую, dependent), а по оси X – переменную, на основе которой будем предсказывать (независимую, independent).

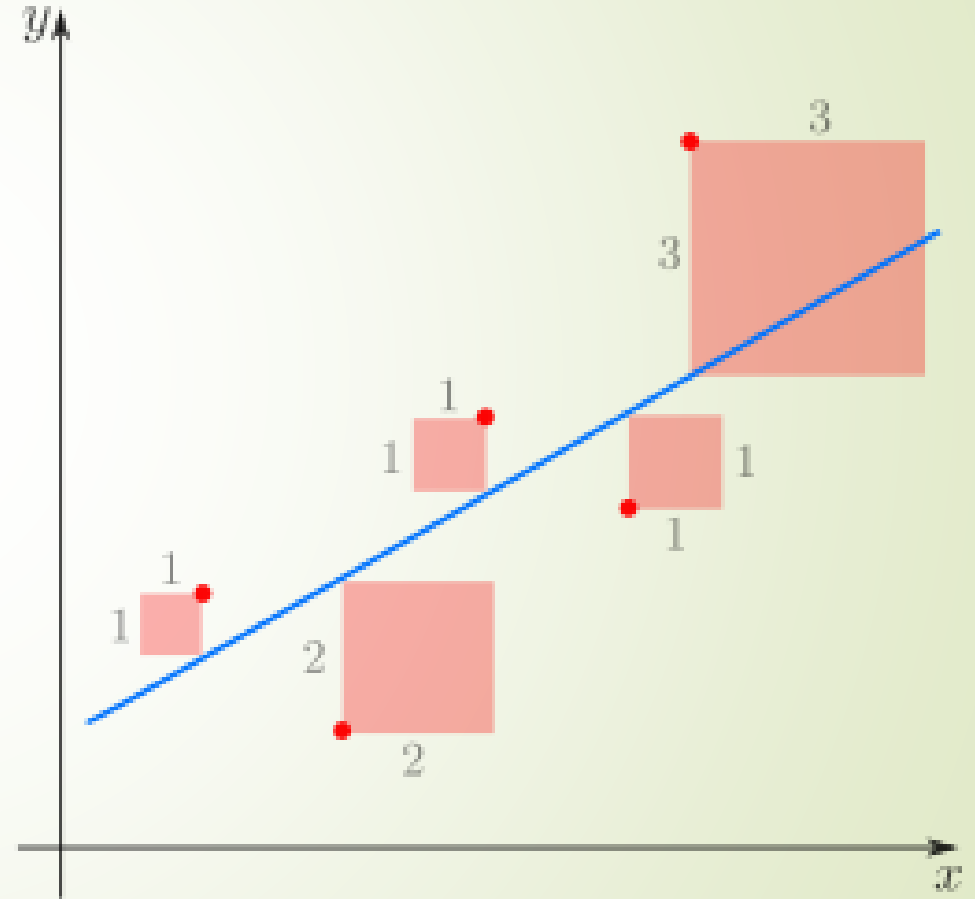
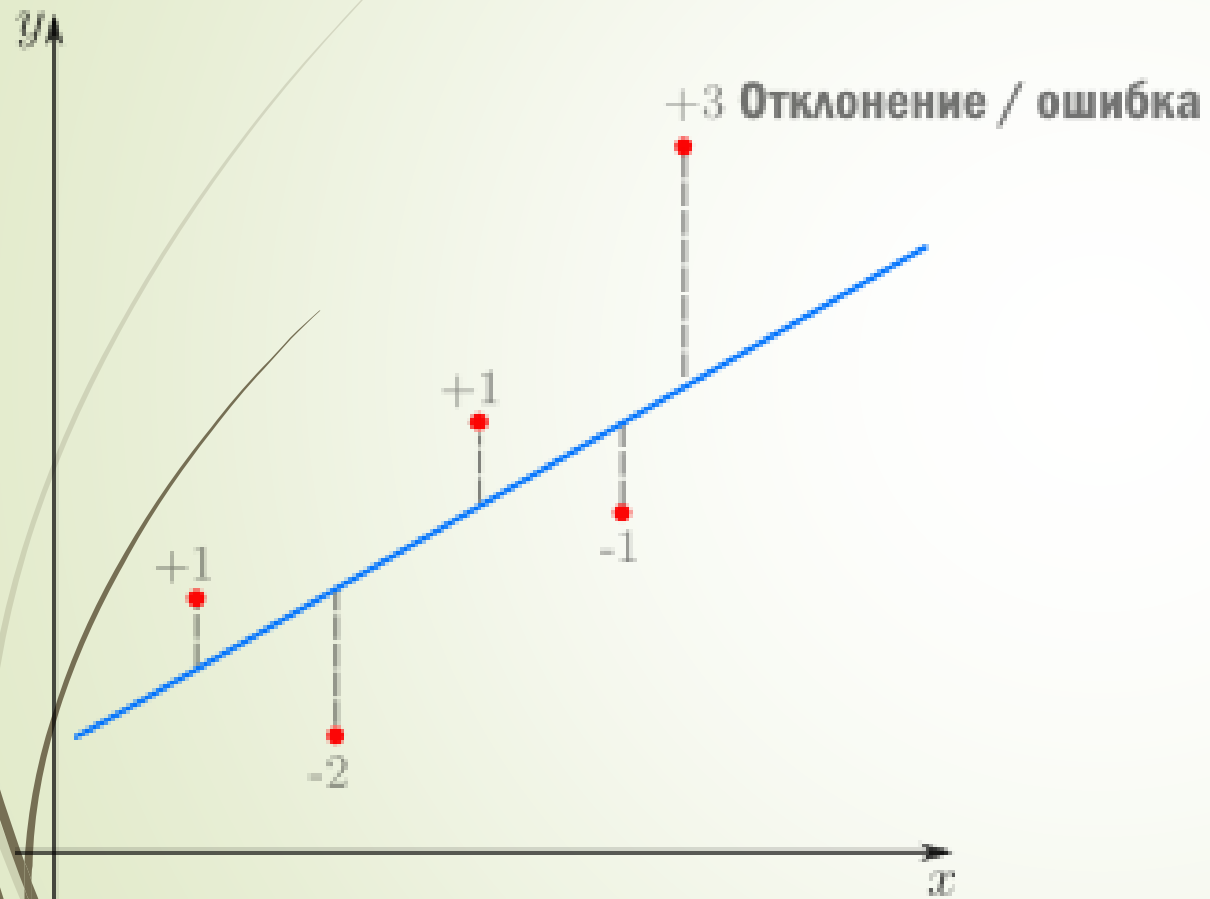


Чем отличаются?

РЕГРЕССИЯ (regression) – предсказание одной переменной на основании другой. Одна переменная – независимая (independent), а другая – зависимая (dependent).

КОРРЕЛЯЦИЯ (correlation) – показывает, в какой степени две переменные **СОВМЕСТНО ИЗМЕНЯЮТСЯ**. Нет зависимой и независимой переменных, они **ЭКВИВАЛЕНТНЫ**.

Метод наименьших квадратов





Как предсказывать?

$$Y_i = \beta_0 + \beta_1 x_i$$

β_0 — сдвиг (пересечение с осью Y)

β_1 — наклон прямой Y

x_i — значение переменной X в i -м наблюдении

Функция регрессии lm()

```
model <- lm(BRCA2 ~ BRCA1, data = data)
summary(model)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.77843	0.09398	8.283	3.87e-16 ***
BRCA2	0.66767	0.05172	12.909	< 2e-16 ***

β_0

β_1

$$Y_i = \beta_0 + \beta_1 x_i \Rightarrow Y_i = \mathbf{0.78} + \mathbf{0.67}x_i$$

Оценка модели

Multiple R-squared: **0.1438**, Adjusted R-squared: **0.143** 😞

F-statistic: 166.6 on 1 and 992 DF, p-value: < **2.2e-16** 😊

Чем ближе R^2 к единице, тем лучше модель описывает выборку!

F-statistic - насколько предсказываемая величина зависит от предиктора. Для этого выдвигается нулевая гипотеза, что предсказываемая величина вообще не зависит от предикторов. Для этой гипотезы определяется р-значение

Оценка модели

1. Зависимость ошибок от предсказанных значений. Симметричность точек относительно линии

2. Q-Q plot. Насколько точки близки к диагонали?

3. Scale-Location plot. Симметричность точек относительно линии

4. Residuals-Leverage plot. Здесь по оси x - расстояние Кука, а по оси y - стандартизированный размер выбросов. Расстояние Кука показывает high-leverage points - точки, которые имеют экстремальные предсказанные значения, то есть очень большие или очень маленькие значения по предикторам. Для линейной регрессии такие значения имеют большее значение, чем экстремальные точки по предсказываемой переменной. Особенно сильное влияние имеют точки, которые имеют экстремальные значения и по предикторам, и по предсказываемой переменной. Одна такая точка может поменять направление регрессионной прямой!

