НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

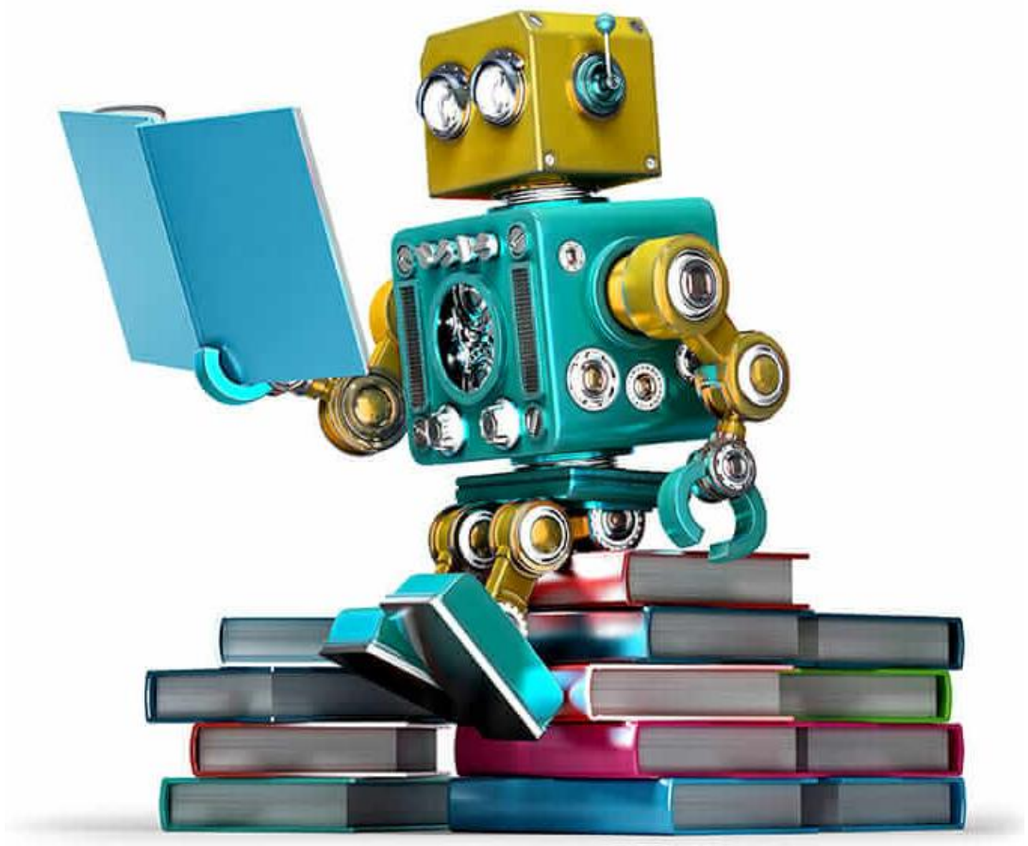
ОСНОВЫ АНАЛИЗА ДАННЫХ В МЕЖДУНАРОДНЫХ ОТНОШЕНИЯХ

Лекция 2

Маргарита Бурова

Москва, 2018

ЗАДАЧИ МАШИННОГО ОБУЧЕНИЯ

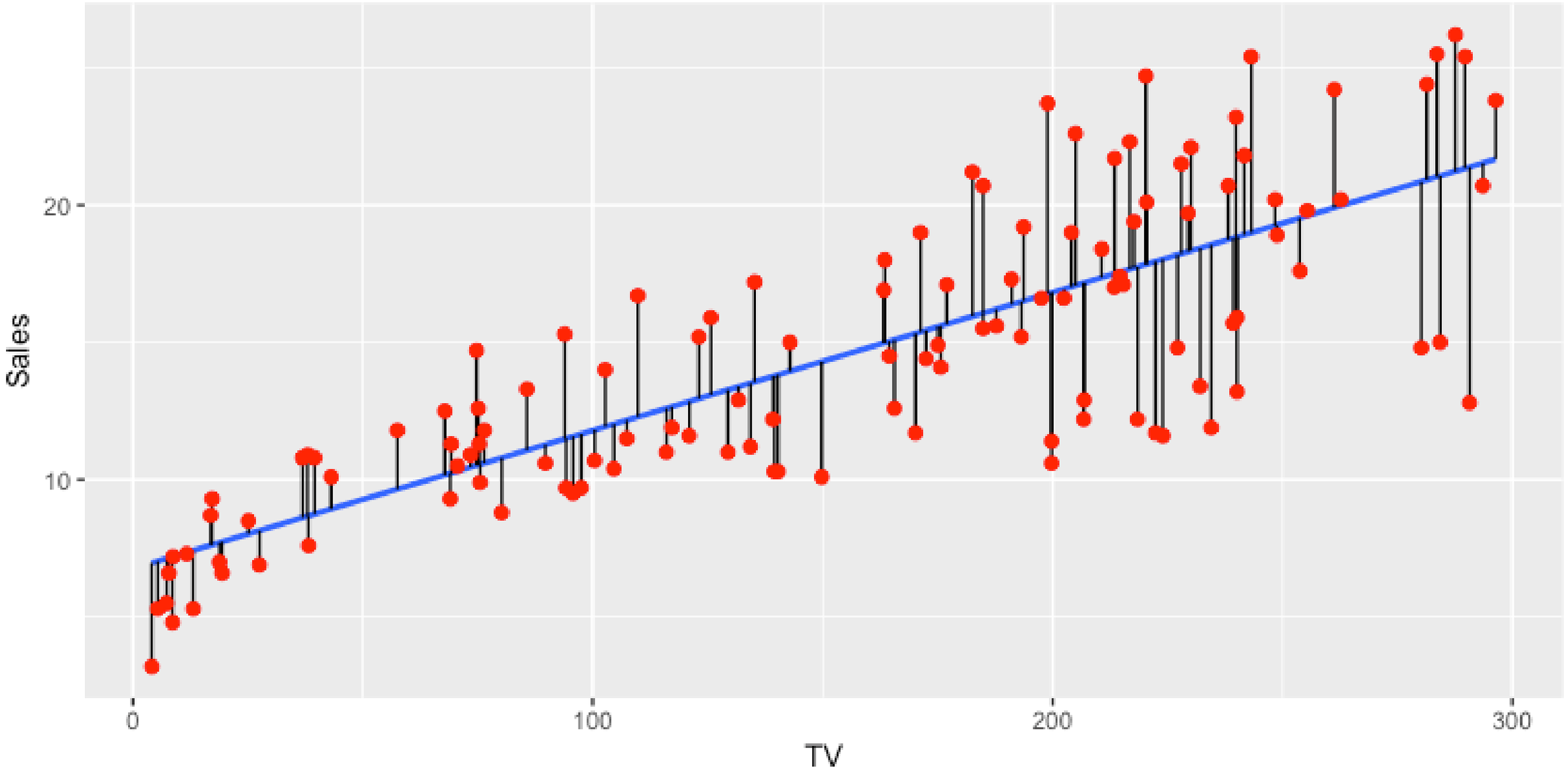


Классификация

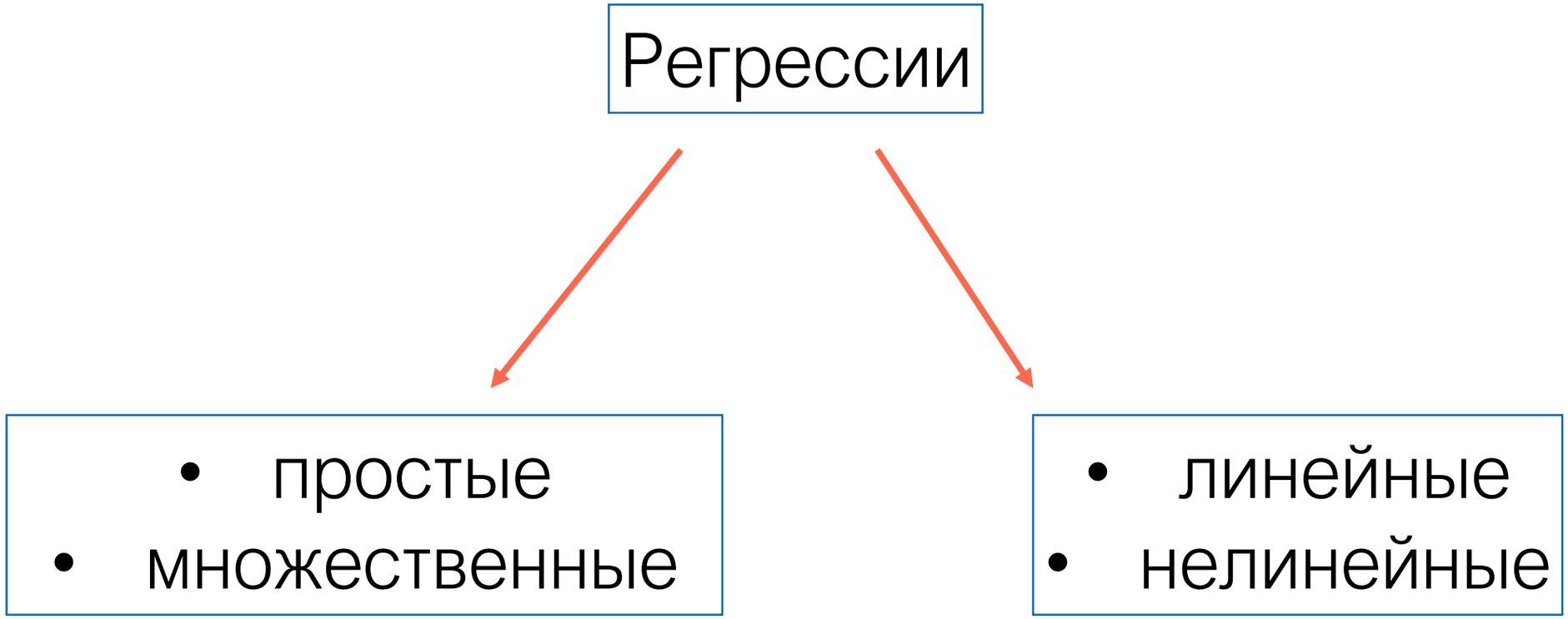
Регрессия

Кластеризация

РЕГРЕССИЯ: ПРИМЕР



РЕГРЕССИИ



ОПРЕДЕЛЕНИЕ РЕГРЕССИИ

Регрессия – это способ объяснить зависимость переменной через одну или набор других.

Y – это переменная, которую планируют объяснять. Ее называют зависимой.

$X_1,\ldots, X_n$ – это переменные, через которую планируют объяснить Y. Их называют независимыми/регрессорами/предикторами/факторами

РЕГРЕССИЯ ПОМОГАЕТ:

1. Объяснить разброс/неоднородность зависимой переменной через независимые

2. Предсказать значения зависимой переменной через независимые.

3. Определить вклад каждой из независимых переменных

РЕГРЕССИИ: НЕМНОГО МАТЕМАТИКИ

Объяснить одну переменную набором других = Восстановить функцию.

Т.е. описать их взаимосвязь уравнением:

$$Y = f(X)$$

В простейшем случае взаимосвязь описывается линейным уравнением:

$$Y = a_1 X_1 + a_2 X_2 \ldots a_M X_M + b$$

M – число предикторов.

РЕГРЕССИИ:ОДНОМЕРНАЯ МОДЕЛЬ

$\hat{Y} = \hat{a}X_1 + \hat{b}$ - эмпирические значения по модели

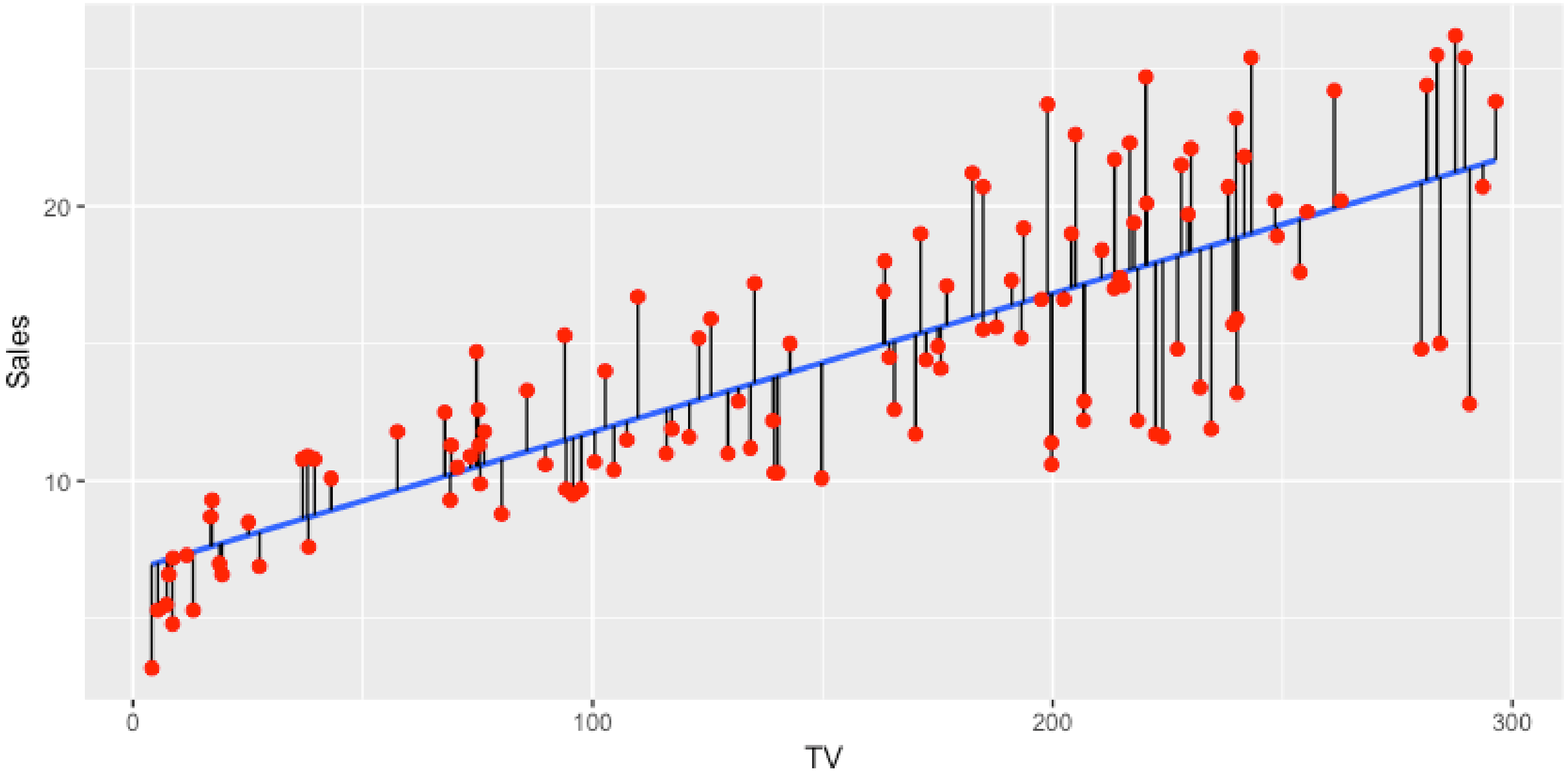
Ошибкой называется разница между действительными значениями и эмпирическими:

$$\mathrm{E} = Y - \hat{Y} = \sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n} e$$

Наша цель: минимизировать суммарную квадратичную ошибку

$$\sum_{i=1}^{n} e^2 = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2 = \sum_{i=1}^{n}(aX + b - y_i)^2 \to \min_{a,b}$$

РЕГРЕССИИ:ОДНОМЕРНАЯ МОДЕЛЬ



РЕГРЕССИИ:КАЧЕСТВО МОДЕЛИ

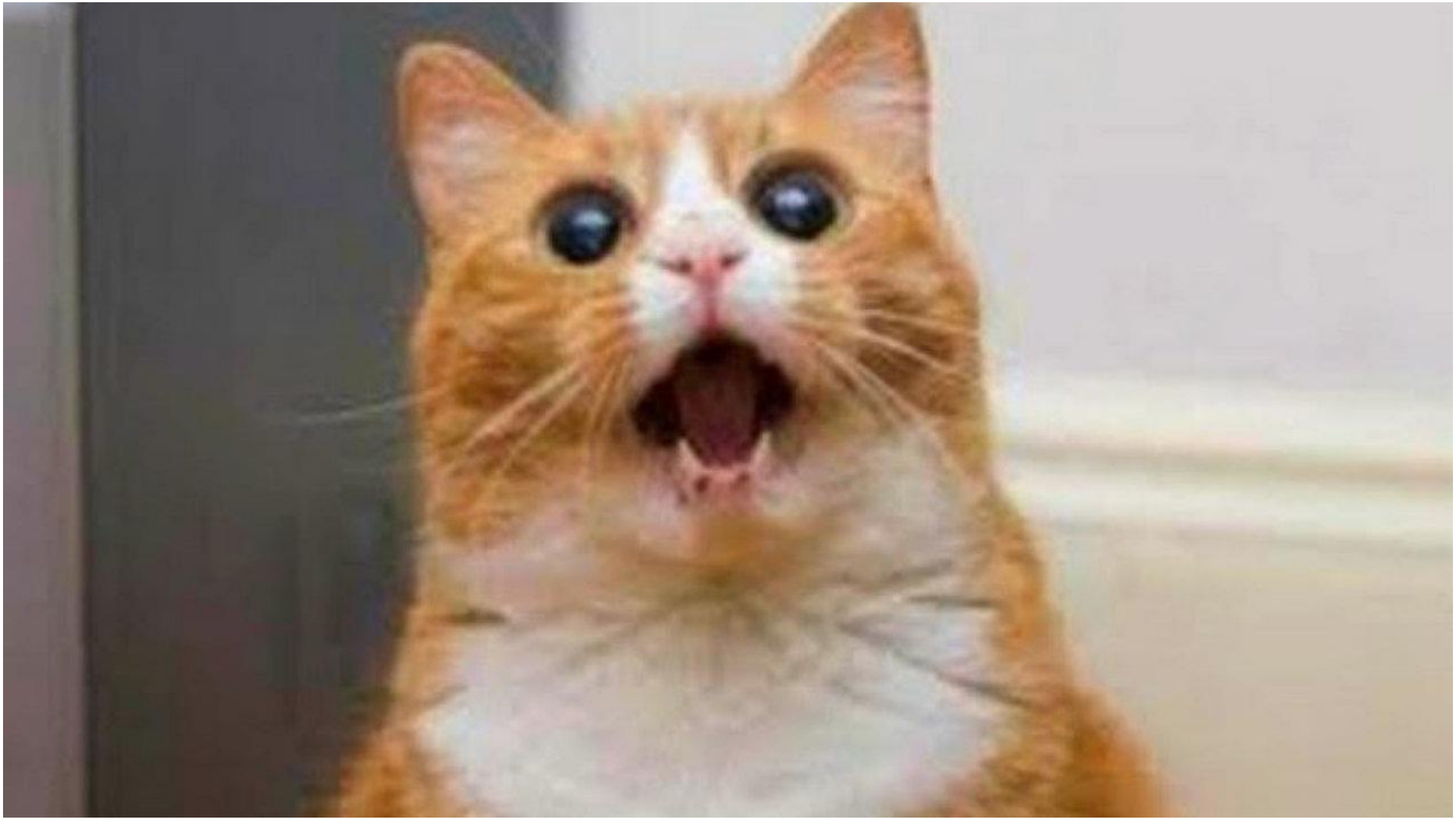
Goodness-of-fit – оценка того, как хорошо мы смогли описать восстановленной функцией данные.

Для оценки качества регрессии используется статистика, называемая коэффициент детерминации - R^2 .

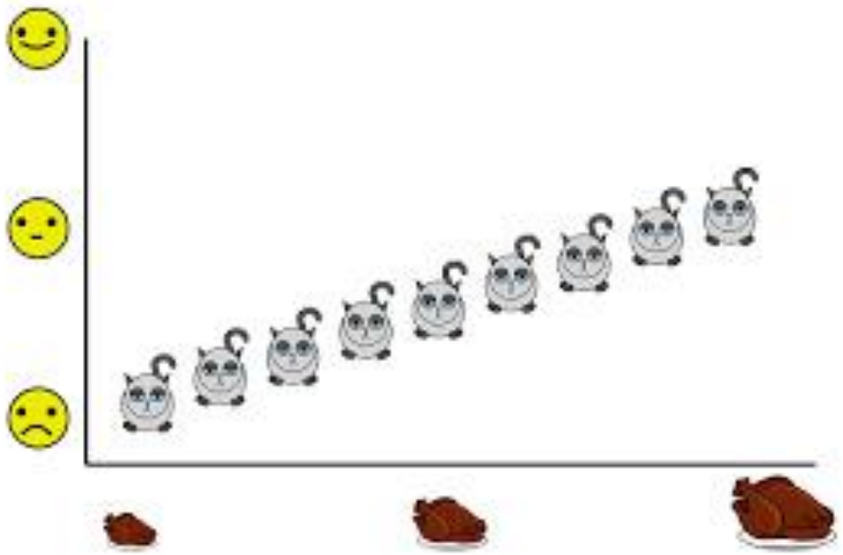
R^2 показывает, какую долю разброса данных мы объяснили построенной регрессией.

$R^2 > 0.7$ – приемлемое качество;
$R^2 < 0.5$ – неудачное моделирование.

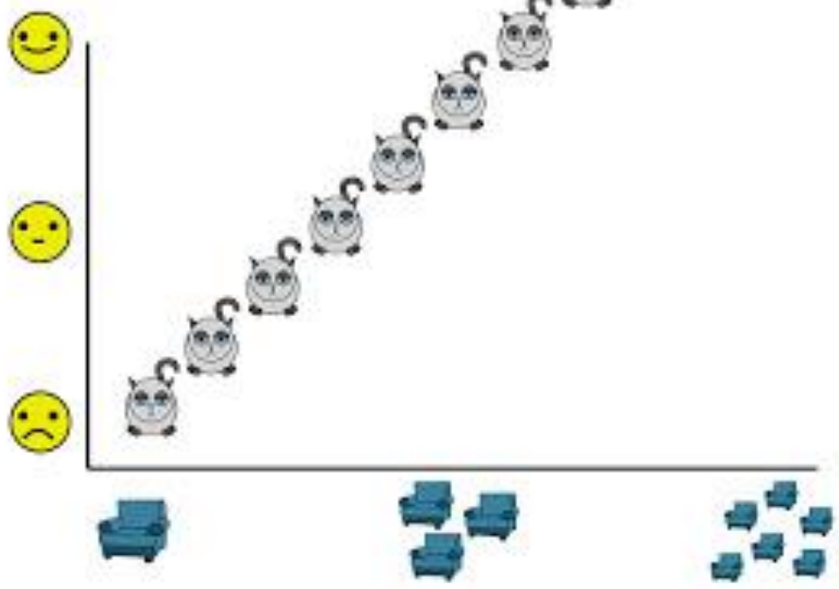
ТЕПЕРЬ ПОВТОРИМ НА КОТИКАХ



ТЕПЕРЬ ПОВТОРИМ НА КОТИКАХ



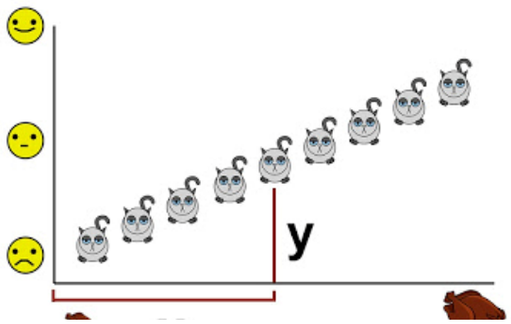
r = 1



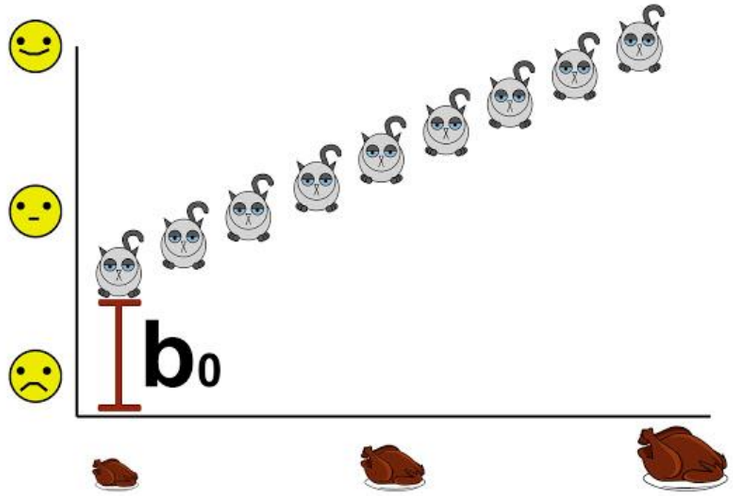
r = 1

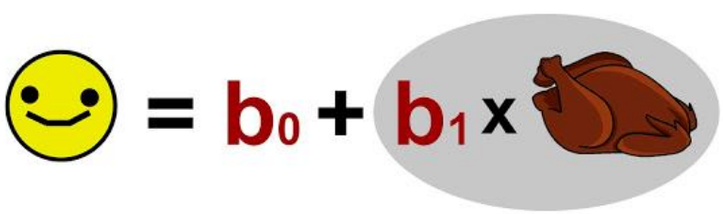
*Иллюстрации взяты из книги «Статистика и котики»

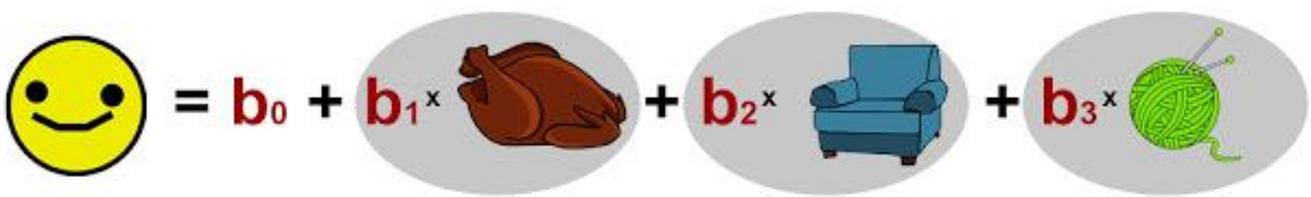
ТЕПЕРЬ ПОВТОРИМ НА КОТИКАХ



$$b_1 = \frac{y}{x}$$





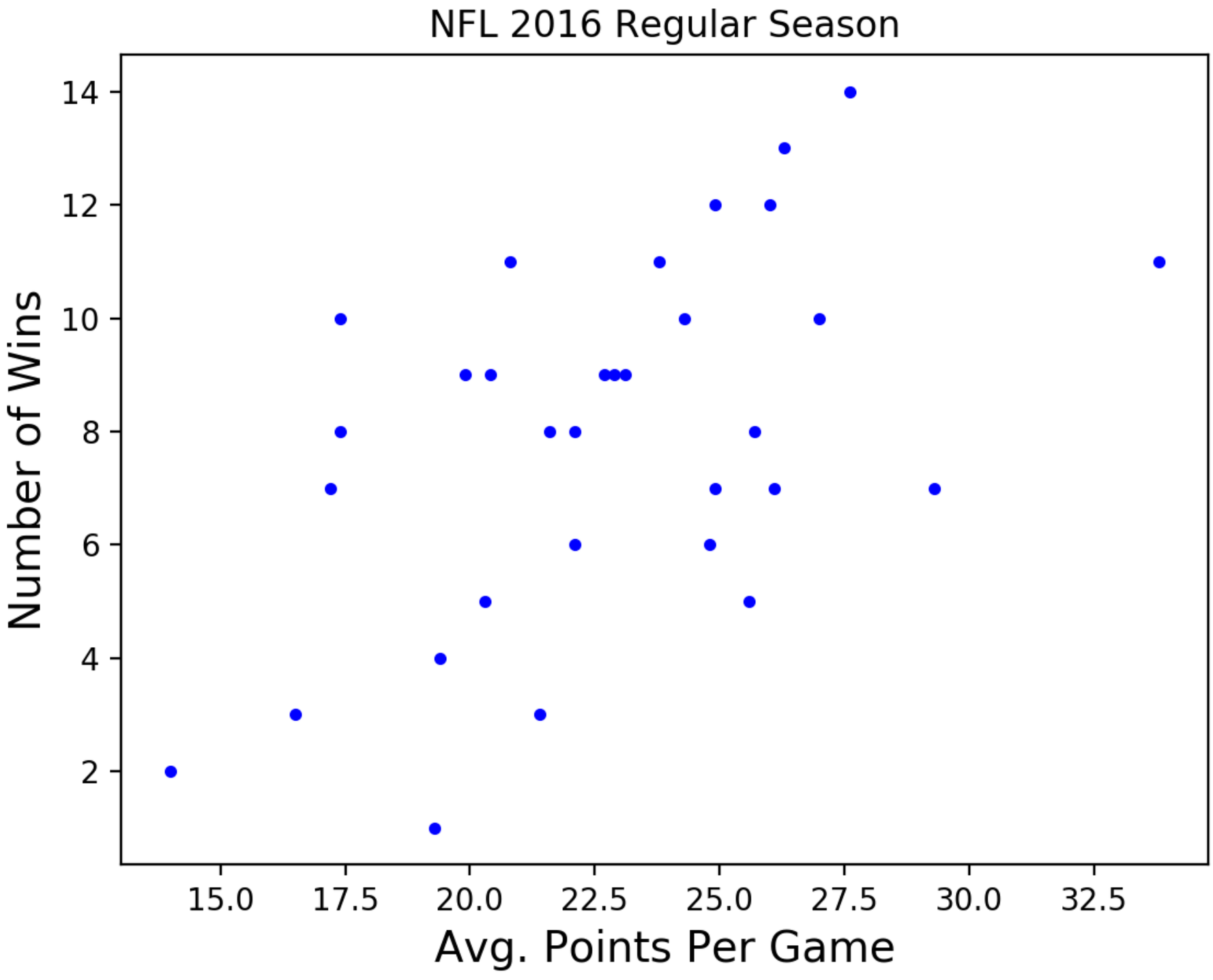


*Иллюстрации взяты из книги «Статистика и котики»

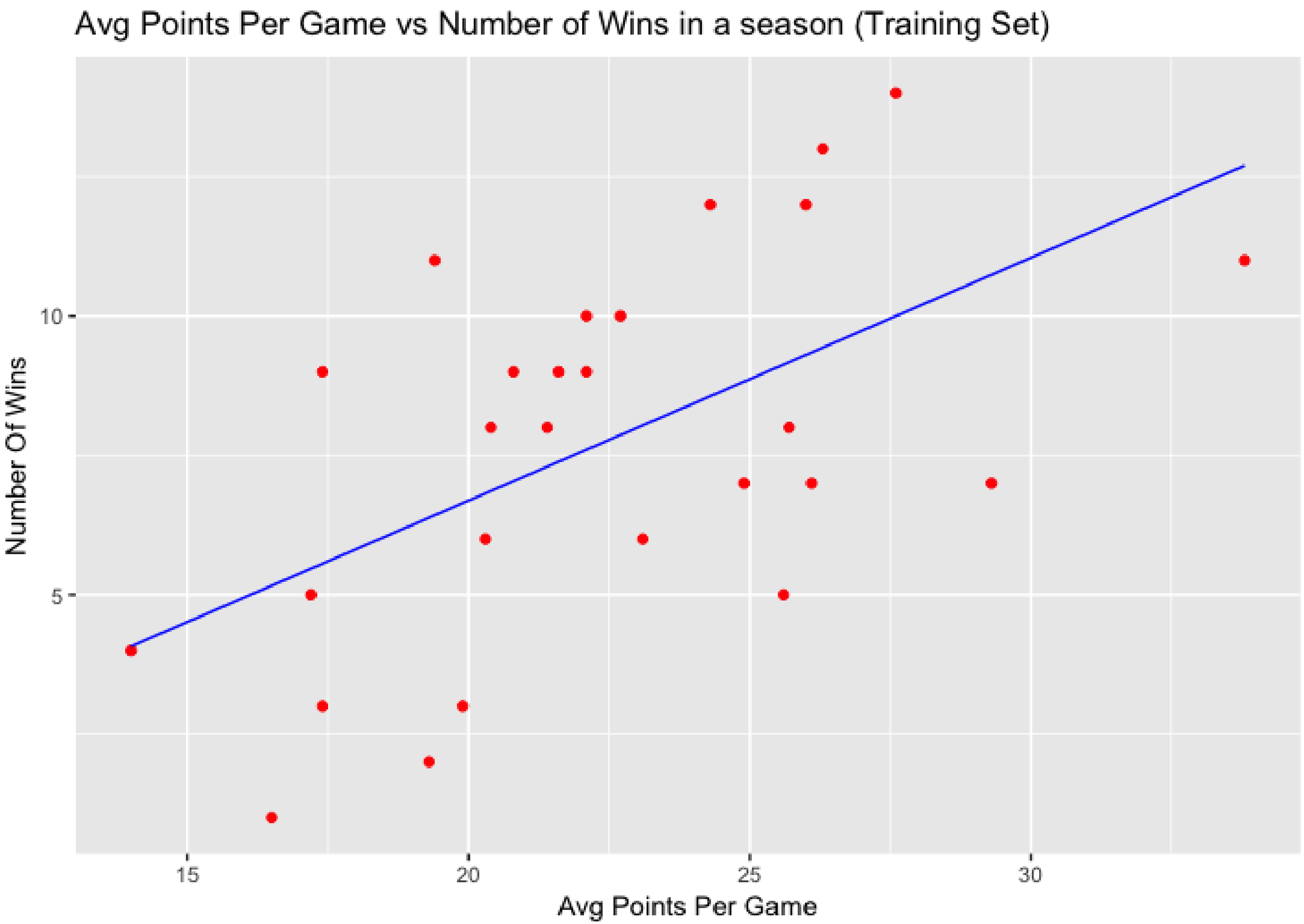
ЛИНЕЙНАЯ РЕГРЕССИЯ: ПРИМЕР

Team	Avg. Points Per Game	Number of Wins In Season
Atlanta Falcons	33.8	11
New Orleans Saints	29.3	7
New England Patriots	27.6	14
Green Bay Packers	27	10
Dallas Cowboys	26.3	13
Arizona Cardinals	26.1	7
Oakland Raiders	26	12
Indianapolis Colts	25.7	8
San Diego Chargers	25.6	5
Buffalo Bills	24.9	7
Pittsburgh Steelers	24.9	11
Washington Redskins	24.8	8
Kansas City Cheifs	24.3	12
Tennesee Titans	23.8	9
Carolina Panthers	23.1	6
Philadelphia Eagles	22.9	7
Miami Dolphins	22.7	10
Seattle Seahawks	22.1	10
Tampa Bay Buccaneers	22.1	9
Detroit Lions	21.6	9

ЛИНЕЙНАЯ РЕГРЕССИЯ: ПРИМЕР



ЛИНЕЙНАЯ РЕГРЕССИЯ: ПРИМЕР



ПРОВЕРКА ГИПОТЕЗ

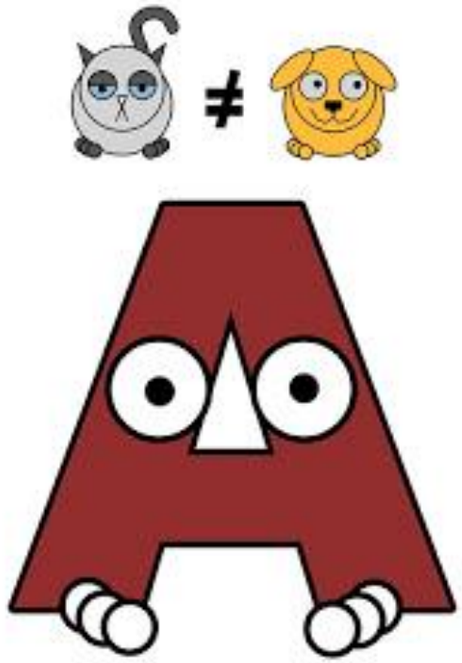
Гипотеза- некоторое утверждение относительно изучаемого набора данных

Существует нулевая гипотеза и альтернативная гипотеза

ПРОВЕРКА ГИПОТЕЗ

Нулевая гипотеза

p<0,05



Альтернативная гипотеза

*Иллюстрации взяты из книги «Статистика и котики»

ПРОВЕРКА ГИПОТЕЗ

*Иллюстрации взяты из книги «Статистика и котики»

СТАТИСТИЧЕСКИЕ ТЕСТЫ

- Одновыборочные
- Двухвыборочные
- Парные

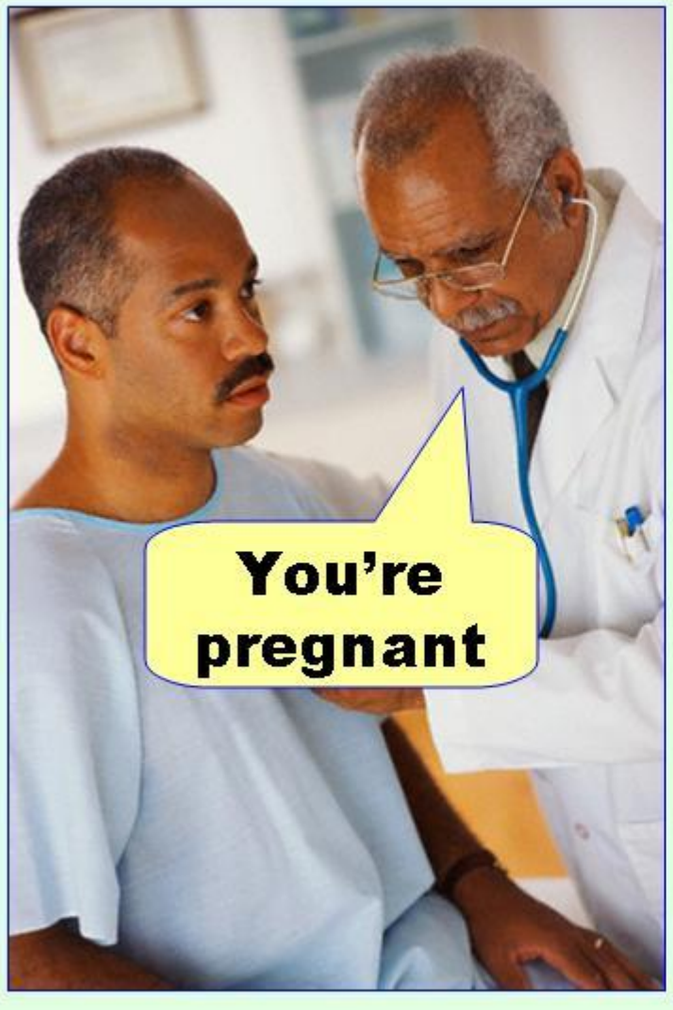
ОШИБКИ 1 И 2 РОДА

Decision based on sample	Truth about the population: H_0 true	Truth about the population: H_a true
Reject H_0	Type I error	Correct decision
Accept H_0	Correct decision	Type II error

ОШИБКИ 1 И 2 РОДА

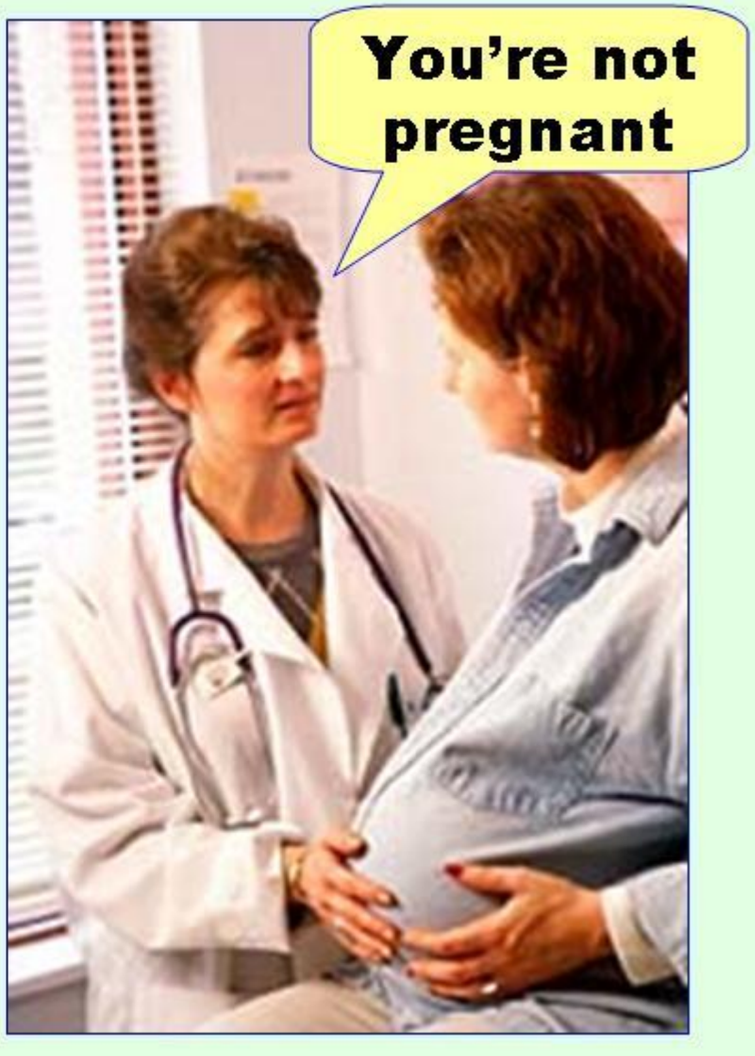
Type I error

(false positive)



Type II error

(false negative)

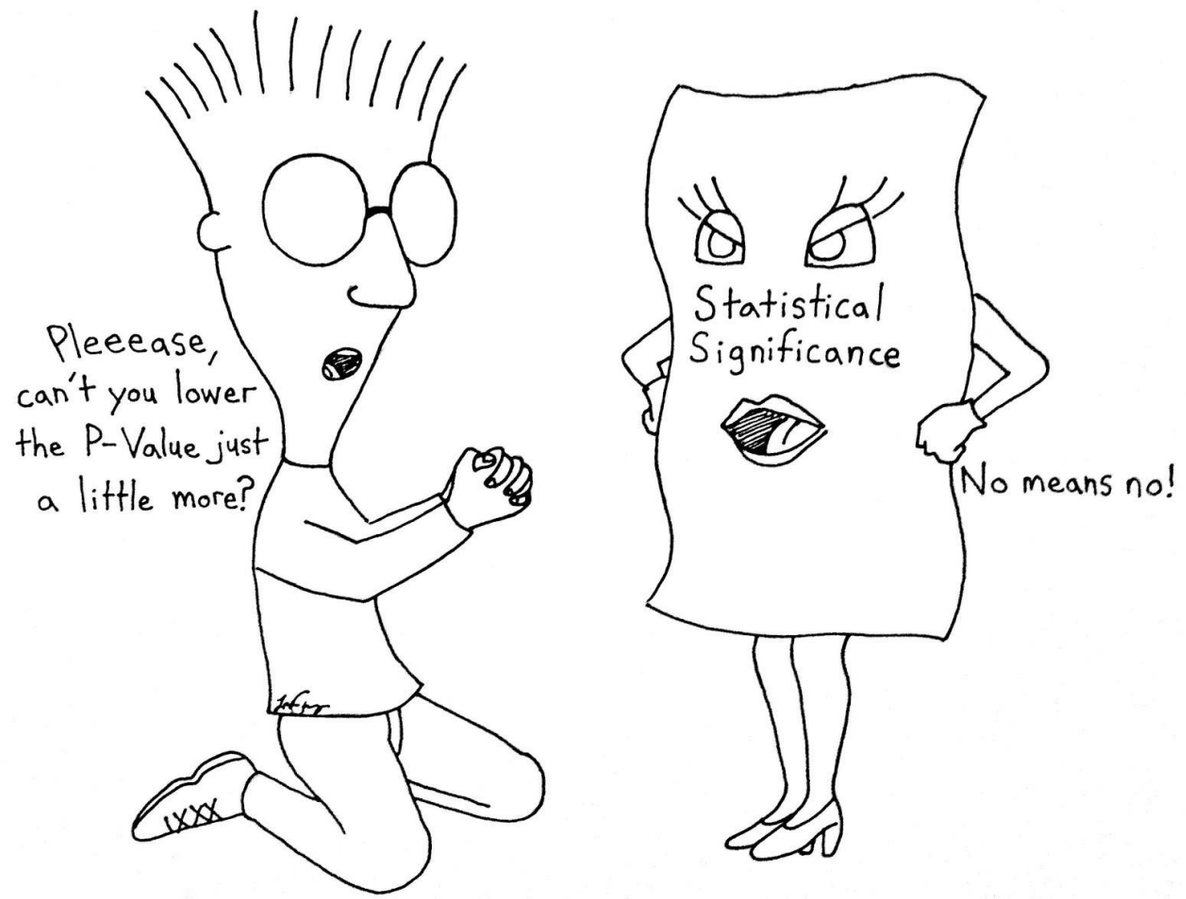


P - VALUE

P-значение (англ. **P-value**) — величина, используемая при тестировании статистических гипотез. Фактически это вероятность ошибки при отклонении нулевой гипотезы (ошибки первого рода).

P - VALUE

- Если добавить новые данные, то p-value может измениться
- Не стоит зацикливаться на значении 0,05



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ