

# Метод наименьших квадратов

Эконометрика. Coursera. Лекция 1

# Эконометрика на одном слайде :)

## Вопросы:

- Как устроен мир? Как переменная  $x$  влияет на переменную  $y$ ?
- Что будет завтра? Как спрогнозировать переменную  $y$ ?

## Ответ:

Модель — формула для объясняемой переменной

## Например:

- $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$

# Основные типы данных:

- Временные ряды
- Перекрёстные данные
- Панельные данные

Есть много-много других!

Данные по России:

Год	Население	Безработица
2010	142962	7.4
2011	142914	6.5
2012	143103	5.5
2013	143395	5.5

Результаты зимних Олимпийских игр 2014:

Страна	Золото	Серебро	Бронза
Россия	13	11	9
Норвегия	11	5	10
Канада	10	10	5
США	9	7	12

Сочетание первых двух: данные по нескольким переменным для множества объектов в разные моменты времени

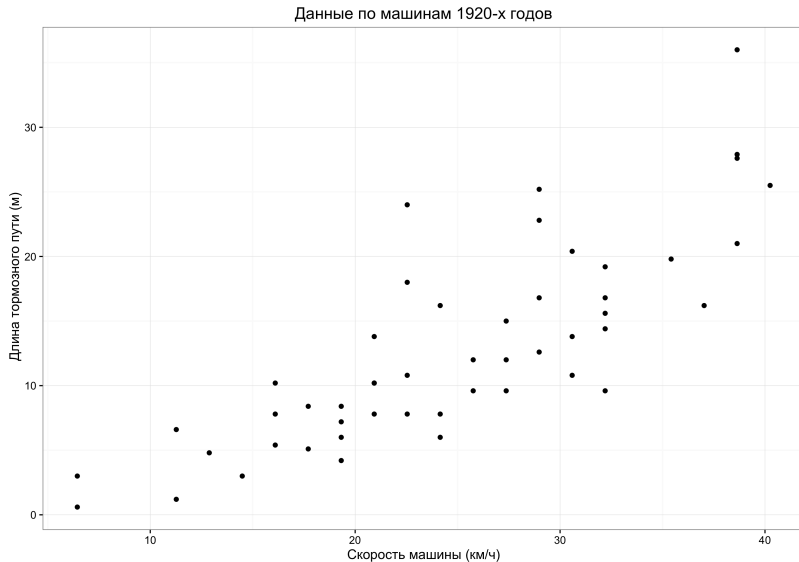
- Одна зависимая, объясняемая, переменная:  $y$
- Несколько регрессоров, объясняющих, переменных:  $x, z, \dots$
- По каждой переменной  $n$  наблюдений:  $y_1, y_2, \dots, y_n$

Исторические данные 1920-х годов :)

Длина тормозного пути (м), $y_i$	Скорость машины (км/ч), $x_i$
0.6	6.44
3.0	6.44
1.2	11.27
...	...



# Всегда изображайте данные!



Пример:  $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$

- Наблюдаемые переменные:  $y$ ,  $x$
- Неизвестные параметры:  $\beta_1$ ,  $\beta_2$
- Случайная составляющая, ошибка:  $\varepsilon$

## План действий

- придумать адекватную модель
- получить оценки неизвестных параметров:  $\hat{\beta}_1$ ,  $\hat{\beta}_2$
- прогнозировать, заменив неизвестные параметры на оценки:

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$$

- Способ получить оценки неизвестных параметров модели исходя из реальных данных.

Ошибка прогноза:  $\hat{\varepsilon}_i = y_i - \hat{y}_i$ .

Сумма квадратов ошибок прогноза:

$$Q(\hat{\beta}_1, \hat{\beta}_2) = \sum_{i=1}^n \hat{\varepsilon}_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Суть МНК: В качестве оценок взять такие  $\hat{\beta}_1, \hat{\beta}_2$ , при которых сумма квадратов ошибок прогноза,  $Q$ , минимальна.

## Пример с машинами:

Фактические данные:

$$x_1 = 6.68, x_2 = 6.68, \dots,$$

$$y_1 = 0.6, y_2 = 3, \dots$$

Модель:  $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ . Формула для прогнозов:  $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$

Сумма квадратов ошибок прогнозов:  $Q = \sum_{i=1}^n (y_i - \hat{y}_i)^2$

$$Q = (0.6 - \hat{\beta}_1 - \hat{\beta}_2 6.68)^2 + (3 - \hat{\beta}_1 - \hat{\beta}_2 6.68)^2 + \dots$$

Точка минимума, найдена в R:  $\hat{\beta}_1 = -5.3, \hat{\beta}_2 = 0.7$ :

Формула для прогнозов:  $\hat{y}_i = -5.3 + 0.7 x_i$

## Простой пример [у доски]

Имя	Вес (кг), $y_i$	Рост (см), $x_i$
Вася	60	170
Коля	70	170
Петя	80	181

Оцените модели:

$$y_i = \beta + \varepsilon_i,$$

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

Маленькая подготовка:  $n\bar{x} = \sum_i x_i = \sum_i \bar{x}$ ,  $\sum_i (x_i - \bar{x}) = 0$ .

В модели  $y_i = \beta + \varepsilon_i$

$$\hat{\beta} = \bar{y}$$

Интерпретация:

В модели без объясняющих переменных наилучший прогноз — это среднее значение зависимой переменной

В модели  $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$

$$\hat{\beta}_2 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$$

Интерпретация:

Точка  $(\bar{x}, \bar{y})$  лежит на линии регрессии  $\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x$

# Терминология и обозначения:

$y_i$  — зависимая, объясняемая, переменная

$x_i$  — регрессор, объясняющая переменная

$\varepsilon_i$  — ошибка, ошибка модели, случайная составляющая

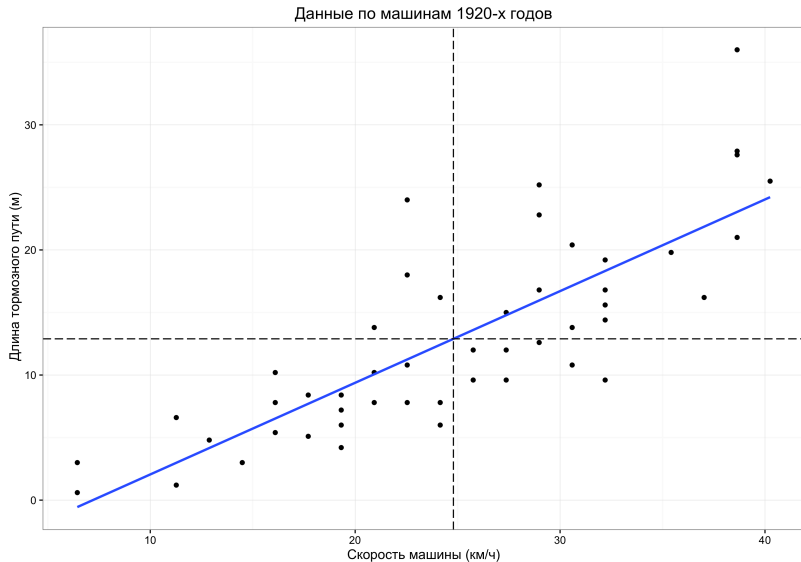
$\hat{y}_i$  — прогноз, прогнозное значение

$\hat{\varepsilon}_i = y_i - \hat{y}_i$  — остаток, ошибка прогноза

$RSS = \sum_{i=1}^n \hat{\varepsilon}_i^2$  — сумма квадратов остатков



# Регрессия проходит через среднюю точку [у доски]



# Много объясняющих переменных [у доски]

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$$

Выпишем систему уравнений для оценок  $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$

$$\begin{cases} \sum \hat{\varepsilon}_i \cdot 1 = 0 \\ \sum \hat{\varepsilon}_i \cdot x_i = 0 \\ \sum \hat{\varepsilon}_i \cdot z_i = 0 \end{cases}$$

- Сумма квадратов остатков

$$RSS = \sum \hat{\varepsilon}_i^2$$

- Общая сумма квадратов

$$TSS = \sum (y_i - \bar{y})^2$$

- Объясненная сумма квадратов

$$ESS = \sum (\hat{y}_i - \bar{y})^2$$

# Абсолютный ликбез по линейной алгебре

Вектора:  $y$ ,  $x$ ,  $\hat{y}$ ,  $\varepsilon$ ,  $\dots$

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad \hat{\varepsilon} = \begin{pmatrix} \hat{\varepsilon}_1 \\ \hat{\varepsilon}_2 \\ \vdots \\ \hat{\varepsilon}_n \end{pmatrix} \quad \vec{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

В нашей модели:  $\hat{y} = \hat{\beta}_1 \cdot \vec{1} + \hat{\beta}_2 \cdot x + \hat{\beta}_3 \cdot z$

# Матрица всех регрессоров

$$X = \begin{pmatrix} 1 & x_1 & z_1 \\ 1 & x_2 & z_2 \\ \vdots & & \\ 1 & x_n & z_n \end{pmatrix}$$

# Длина вектора

Длина вектора,  $|y| = \sqrt{y_1^2 + y_2^2 + \dots + y_n^2}$

Квадрат длины вектора,  $|y|^2 = y_1^2 + y_2^2 + \dots + y_n^2 = \sum_i y_i^2$

Примеры:

$RSS = \sum \hat{\varepsilon}_i^2$  — квадрат длины вектора  $\hat{\varepsilon}$

$TSS = \sum (y_i - \bar{y})^2$  — квадрат длины вектора  $(y - \bar{y} \cdot \vec{1})$

$$\begin{pmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} - \bar{y} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = y - \bar{y} \cdot \vec{1}$$

## Скалярное произведение двух векторов:

$$(x, y) = |x| \cdot |y| \cdot \cos(\angle(x, y))$$

$$(x, y) = x_1y_1 + x_2y_2 + \dots + x_ny_n = \sum_i x_iy_i$$

Условие перпендикулярности:

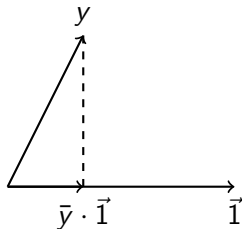
$$x \perp y \Leftrightarrow \sum_i x_iy_i = 0$$

т.к.  $\cos(90^\circ) = 0$ .

# Иллюстрация для регрессии на константу [у доски]

Модель:  $y_i = \beta + \varepsilon_i$

Прогнозы:  $\hat{y}_i = \hat{\beta} = \bar{y}$

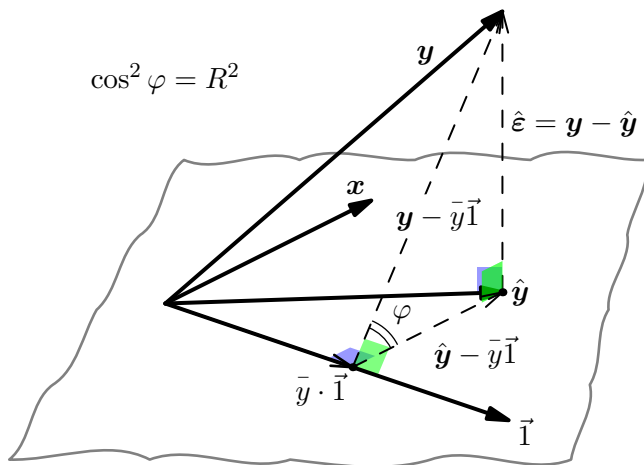




# Геометрическая интерпретация условий первого порядка

$$\begin{cases} \sum \hat{\varepsilon}_i \cdot 1 = 0 \\ \sum \hat{\varepsilon}_i \cdot x_i = 0 \\ \sum \hat{\varepsilon}_i \cdot z_i = 0 \end{cases} \Leftrightarrow \begin{cases} \hat{\varepsilon} \perp \vec{1} \\ \hat{\varepsilon} \perp x \\ \hat{\varepsilon} \perp z \end{cases}$$

## Иллюстрация для множественной регрессии [у доски]



# Если в регрессию включён свободный член $\beta_1$

Если в регрессию включён свободный член,  $y_i = \beta_1 + \dots$ , и оценки МНК единственны, то:

- $\sum \hat{\varepsilon}_i = 0$
- $\sum y_i = \sum \hat{y}_i$
- $\bar{y} = \bar{\hat{y}}$
- $TSS = RSS + ESS$

# Коэффициент детерминации — простой показатель качества

В моделях со свободным членом  $R^2 = ESS/TSS$

$TSS$  — общий разброс  $y$

$ESS$  — объясненный регрессорами разброс

$R^2$  — доля объясненного разброса в общем разбросе

Теорема. Если в регрессию включён свободный член,  $y_i = \beta_1 + \dots$ , и оценки МНК единственны, то  $R^2$  равен выборочной корреляции между  $y$  и  $\hat{y}$ , т.е.

$$R^2 = (sCorr(y, \hat{y}))^2 = \left( \frac{\sum (y_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{\sum (y_i - \bar{y})^2} \sqrt{\sum (\hat{y}_i - \bar{y})^2}} \right)^2$$

# Явная формула для оценок коэффициентов

Модель:  $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_1 & z_1 \\ 1 & x_2 & z_2 \\ \vdots & & \\ 1 & x_n & z_n \end{pmatrix}$$

Линейная алгебра позволяет получить явные формулы:

$$\hat{\beta} = (X'X)^{-1}X'y$$

УРА!!! МНК позволяет оценивать модели!!!

Предположив  $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$

Получаем  $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$

- Как выбрать форму модели?
- А будет ли решение задачи минимизации единственным?
- А будет ли решение задачи минимизации вообще существовать?
- А почему сумма квадратов остатков, а не, скажем, модулей?
- А насколько точны полученные оценки?
- ...

# Источники мудрости:

- Артамонов Н.В., Введение в эконометрику: главы 1.1, 1.2, 2.1
- Борzych Д.А., Демешев Б.Б. Эконометрика в задачах и упражнениях: глава 1
- Катышев П.К., Пересецкий А. А. Эконометрика. Начальный курс: главы 2.1, 2.2, 3.1, 3.2
- Себер Дж., Линейный регрессионный анализ: главы 1.0, 1.1, 1.2, 2.1, 3.1