ФКН

Департамент больших данных и информационного поиска

Москва 2025

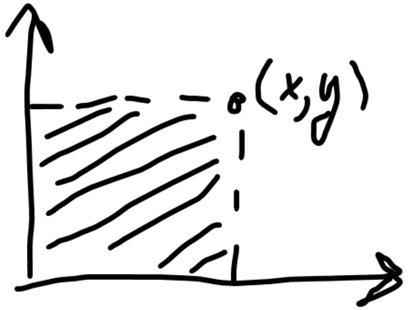
Лекция 5
Линейная регрессия 1

Машинное обучение в цифровом продукте

Полякова И.Ю.

Многомерные СВ

Двумерные СВ



- Совместная ф-я распределения:

$$\boldsymbol{F_{X,Y}(x,y) = P(X \leq x; Y \leq y)}$$

- Совместная ф-я плотности – ф-я такая, что:

$$\boldsymbol{F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(x,y)dxdy}$$

Свойства:

1. $f_{X,Y}(x,y)dxdy \geq 0$
2. $f_{X,Y}(x,y)dxdy$ не убывает по своим параметрам
3. $\int\int f_{X,Y}(x,y)dxdy = 1$ (условие нормировки)
4. $P((X,Y) \in D) = \iint_D \ f_{X,Y}(x,y)dxdy$

Двумерные СВ

- Частные (маргинальные) плотности распределения:

$$\boldsymbol{f_x(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)dy}$$

$$\boldsymbol{f_y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)dx}$$

Двумерные СВ

- Частные (маргинальные) плотности распределения:

$$\boldsymbol{f_x(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x, y)dy}$$

$$\boldsymbol{f_y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x, y)dx}$$

Двумерные СВ

- Условные математические ожидания распределения:

$$\boldsymbol{E[x|y] = \int x\, f_x(x|y) dx}$$

$$\boldsymbol{E[y|x] = \int y\, f_y(y|x) dy}$$

Двумерные СВ

- Теоретические центральные моменты k1, k2 порядка:

$$m_{k_1 k_2} = \iint (X - E(X))^{k_1} (Y - E(Y))^{k_2} f_{xy}(x, y)$$

- **Ковариация** (центральный момент порядка 1,1)

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))] = \mathrm{E(XY)} - \mathrm{E(X)E(Y)}$$

- **Корреляция Пирсона**

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sigma_x \sigma_y} \quad \in [0, 1]$$

Двумерные СВ

- Условные плотности распределения:

$$\boldsymbol{f_x(x|y) = \frac{f_{xy}(x,y)}{f_y(y)}}$$

$$\boldsymbol{f_y(y|x) = \frac{f_{xy}(x,y)}{f_x(x)}}$$

- Необходимое и достаточное условие независимости СВ:

$$\boldsymbol{f_{xy}(x,y) = f_x(x) * f_y(y)}$$

Многомерное нормальное распределение

$$X \sim N(\mu, \Sigma) \qquad \Sigma \text{ - cov.matrix}_{n\times n}$$

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)},$$

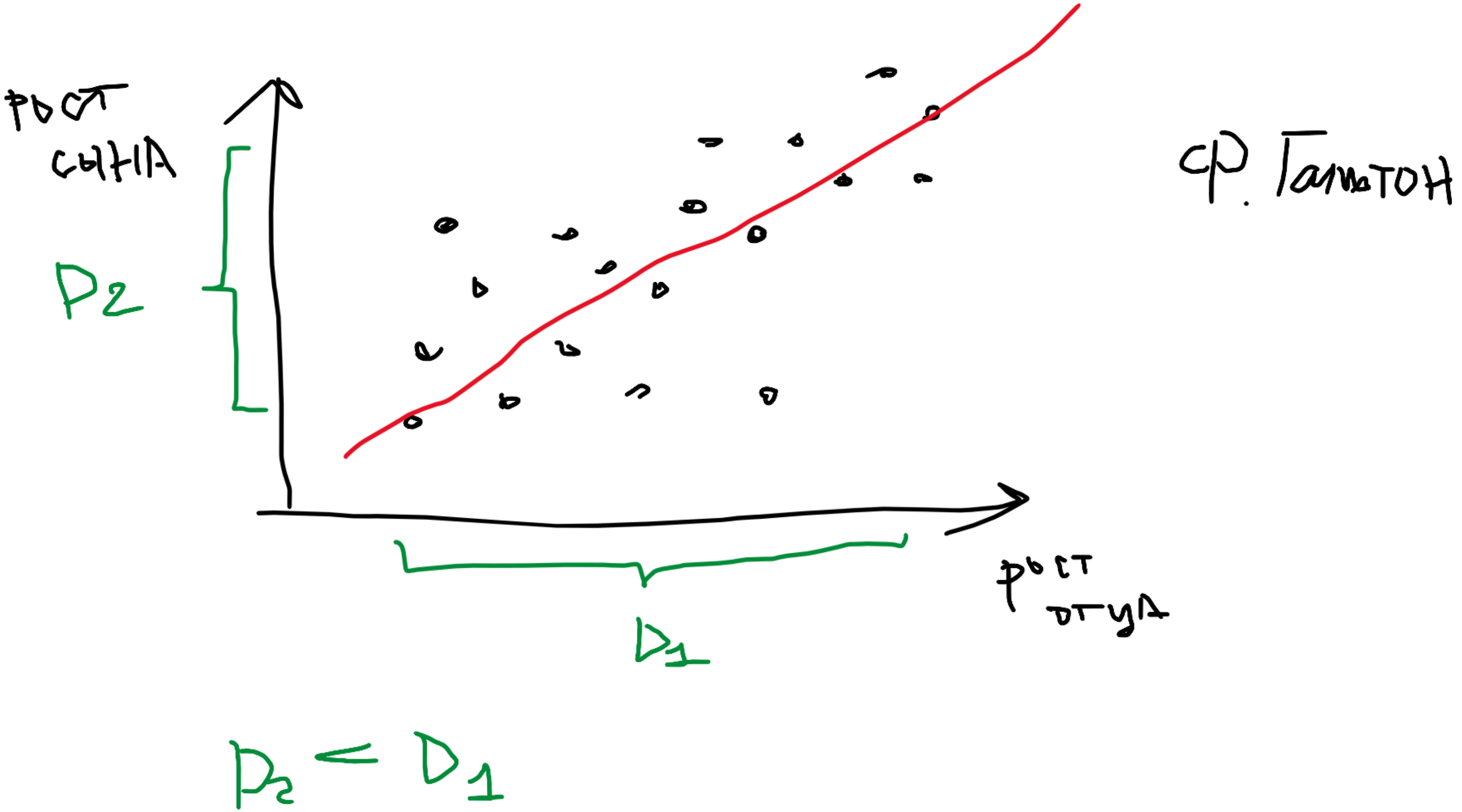
$$\begin{pmatrix} \sigma_1^2 & cov(X_1, X_2) & \dots & cov(X_1, X_n) \\ cov(X_2, X_1) & \sigma_2^2 & \dots & cov(X_2, X_n) \\ \dots & \dots & \dots & \dots \\ cov(X_n, X_1) & cov(X_n, X_2) & \dots & \sigma_n^2 \end{pmatrix}$$

где $\det \Sigma$ - определитель положительно определенной матрицы Σ;

$x = (x_1, \dots, x_n)$; $\mu = (\mu_1, \dots, \mu_n)$

Линейная регрессия

Почему «регрессия»?



Со временем зависимая переменная «регрессирует» к среднему

Постановка

$$y = X\theta + \varepsilon$$

$$y_i = 1 \cdot \theta_0 + X_{i1} \cdot \theta_1 + X_{i2} \cdot \theta_2 + \dots + X_{ik} \cdot \theta_k + \varepsilon_i$$

Постановка

шум

$$y = X\theta + \varepsilon$$

параметры

объясняемая/
целевая
переменная

признаки/
предикторы

Постановка

$$y \in R^n \quad \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad X \in R^{n \times k} \quad \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & \ddots & & \vdots \\ \vdots & & & \\ x_{n1} & & \cdots & x_{nk} \end{pmatrix}$$

$$\theta \in R^k \quad \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_k \end{pmatrix} \qquad \varepsilon \in R^n \quad \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

К концепции оценки параметров в классической линейной регрессии можно прийти по-разному:

- Можно думать об этом как о **минимизации MSE** в машинном обучении
- Можно как об **оценке ММП**
- Можно с точки зрения **«здравого смысла»**

Все это, тем не менее, не противоречит друг другу, а хорошо дополняет понимание модели

Линейная регрессия: ММП

fixed fixed СВ

$$y = X\theta + \varepsilon$$

$\hat{\theta}$ - ?

$$\varepsilon = y - X\theta$$

$$\ln L(\theta) = \ln p(\underbrace{y - X\theta}_{y-\hat{y}})$$

распр-е шума?

При выводах формул и рассуждениях о лин. регрессии в контексте работы с пришедшей выборкой, X – все-таки рассматривается как фиксированная матрица наблюдений, а не случайная величина

Линейная регрессия: ММП и МНК

Пусть: $\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$, тогда

$$f(\varepsilon_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \cdot \exp^{-\frac{\varepsilon_i^2}{2\sigma_i^2}}$$

$$L = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \cdot \exp^{-\frac{1}{2\sigma^2}\cdot\sum_i \varepsilon_i^2}$$

$$\ln L = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_i (y_i - \hat{y}_i(\theta))^2 \to \max_\theta$$

$$\Updownarrow$$

$$\sum_i (y_i - \hat{y}_i)^2 \to \min_\theta \to \text{МНК}$$

Линейная регрессия: оценки ММП (МНК)

Теорема:

$\varepsilon \sim N(0, \sigma^2)$, тогда

$$\boxed{\hat{\theta} = (X^T X)^{-1} X^T y}$$

Сложность такой операции: $O(K^2 N + K^3)$, поэтому зачастую в ML используют численные методы

Док-во

$$(y - X\theta)^T (y - X\theta) \to \min_{\theta}$$

$$\frac{\partial (y - X\theta)^T (y - X\theta)}{\partial \theta} = \frac{\partial (y^T y + \theta^T X^T X \theta - \theta^T X^T y - y^T X \theta)}{\partial \theta} =$$

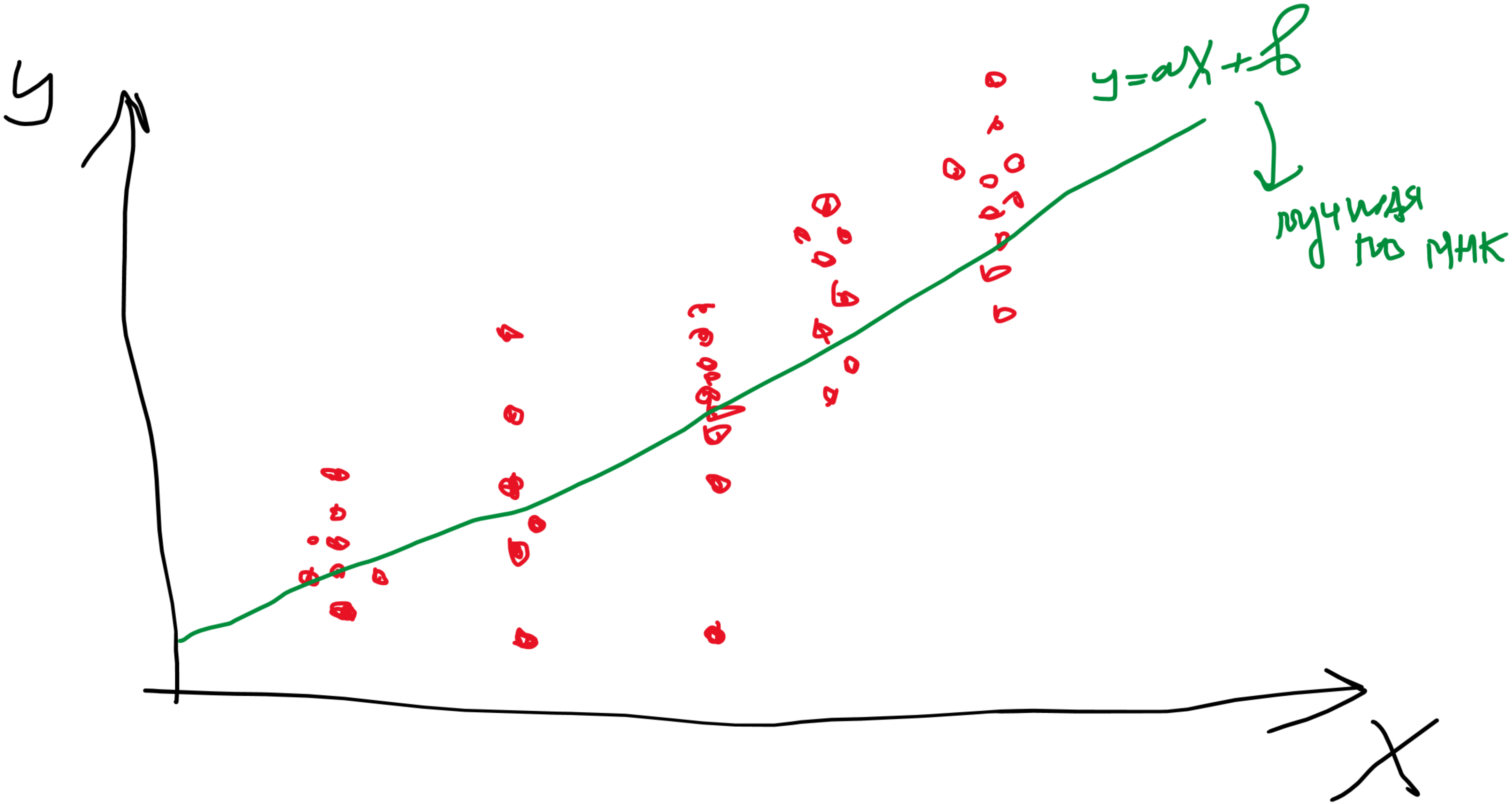
$$= 2X^T X \theta - 2X^T y = 0$$

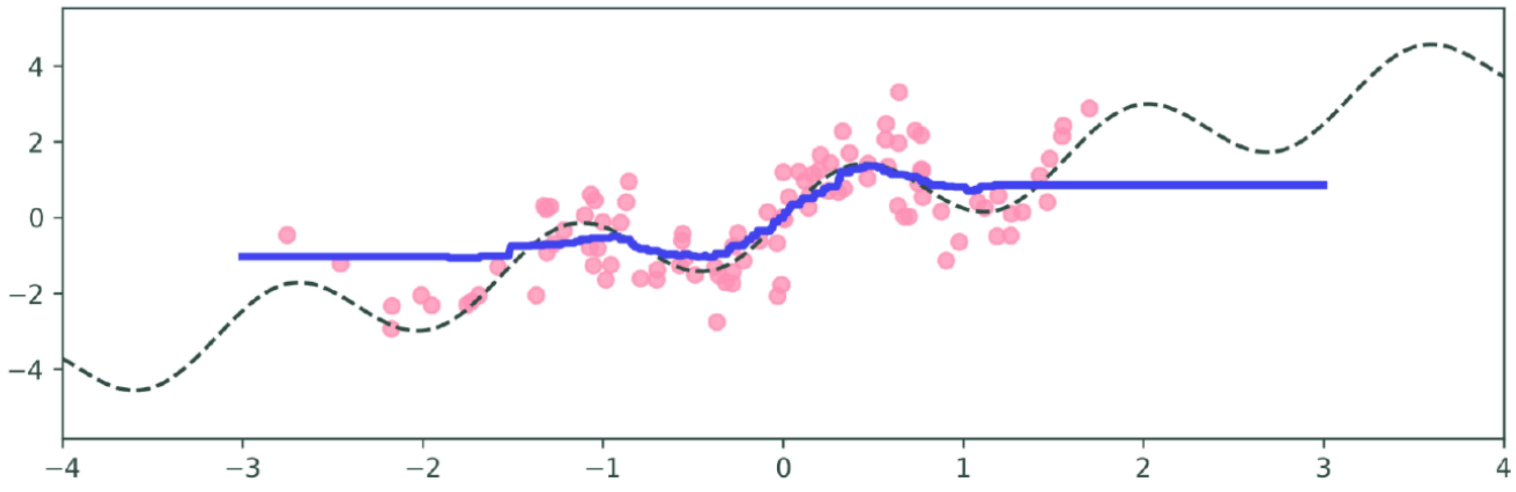
$$\hat{\theta} = (X^T X)^{-1} X^T y$$

Минимум MSE и условное мат.ожидание

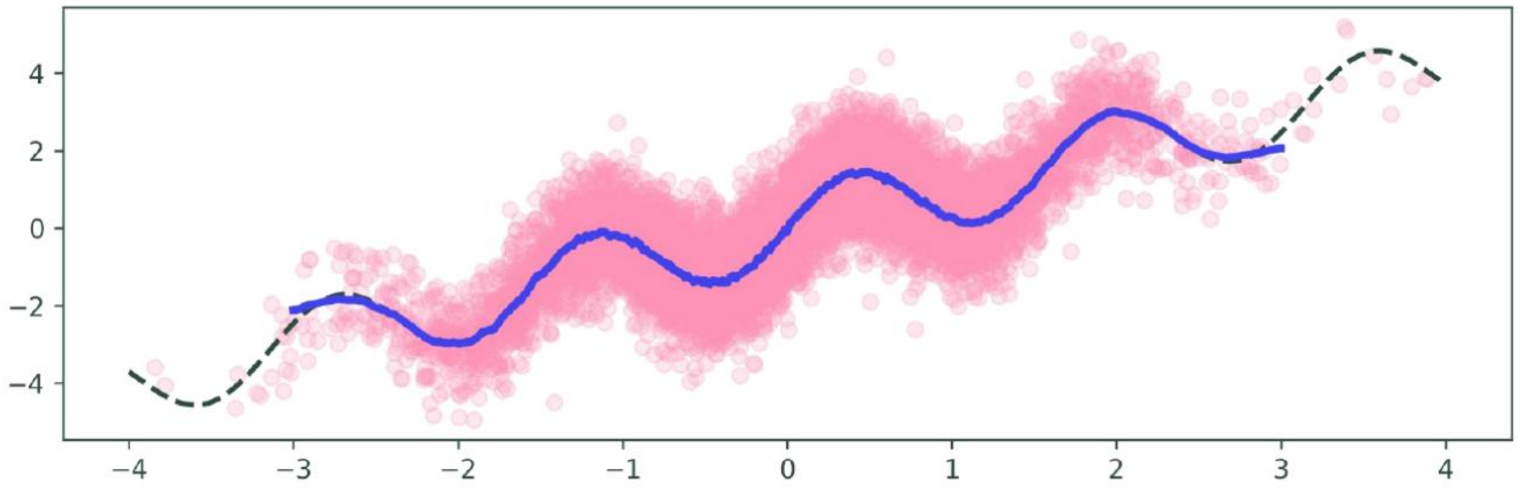
- **Point:** классическая линейная регрессия предсказывает условное среднее зависимой переменной

- Почему?


y
$y = ax + b$
лучшая по МНК
x



Предсказание метода k ближайших соседей при $k = 25$ и $n = 100$



Предсказание метода k ближайших соседей при $k = 500$ и $n = 100000$

Источник: лекция Максима Николаева, ШАД: https://disk.yandex.ru/d/b2rOMFWfsqqBLA

Док-во

$$L(y, \hat{y}) = (y - \hat{y})^2$$

Теоретическая ошибка:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

На практике ошибка зависит от пришедшей выборки:

$$E[(y-\hat{y})^2 | x] =$$

$$= \int (y-\hat{y})^2 f(y|x) dy \to \min_{\hat{y}}$$

Док-во

FOC:

$$\frac{\partial E[(y-\hat{y})^2 \mid x]}{\partial \hat{y}} = -2\int (y-\hat{y})\, f(y|x)\, dy = 0$$

$$\underbrace{\int y\, f(y|x)\, dy}_{E(y|x)} - \int \overset{\text{const}}{\hat{y}}\, f(y|x) = 0$$

$$E(y|x) - \hat{y}\cdot 1 = 0 \implies \boxed{\hat{y} = E(y|x)}$$

по опр-ю

- Методом KNN можно было бы решить подавляющее большинство задач в мире…

- Если бы не «проклятье размерности»!

Теорема Гаусса-Маркова

Если соблюдены условия Гаусса-Маркова, то МНК оценка линейной регрессии является **BLUE**

BLUE

Best Linear Unbiased Estimation

Теорема Гаусса-Маркова

Условия Г-М:

1. Модель правильно специфицирована;
2. Объясняющие переменные линейно независимы;
3. Ошибки независимы друг от друга;
4. Дисперсия ошибок одинакова;
5. Ошибки не зависят от наблюдений;
6. Мат. ожидание ошибок равно нулю.

Теорема Гаусса-Маркова

1. Модель **правильно специфицирована**

- Истинная зависимость от параметров правда линейная;
- Учтены все важные признаки

Иначе: проблема эндогенности

Теорема Гаусса-Маркова

2. Объясняющие переменные детерминированы и **линейно независимы**

- В данных нет мультиколлинеарности;
- Матрица X полного ранга

Иначе: оценки невозможно вычислить (нельзя взять обратную матрицу)

Теорема Гаусса-Маркова

2. Объясняющие переменные **детерминированы** и линейно независимы

- Предпосылка про детерминированность вводится для упрощения технических расчетов (от нее можно отказаться);
- Вводить ее или не вводить зависит от того, как Вы собираете выборку

Пример: Вы собрали з/п Data Scientist с определенными навыками. Затем захотели обновить выборку. Можно взять Data Scientist с такими же навыками, как в исходной выборке и посмотреть их з/п (X – детерминирован), а можно выбрать людей случайным образом (X – СВ)

Теорема Гаусса-Маркова

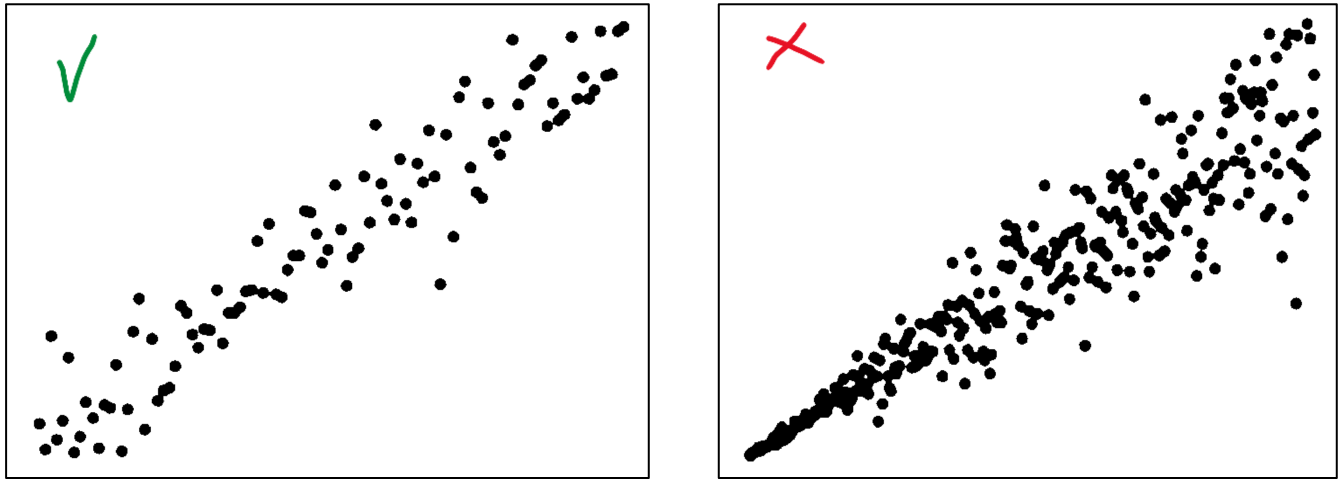
3. Случайные **ошибки не зависят друг от друга**

- Иначе говоря, отсутствует проблема автокорреляции
- Нарушается в панельных данных и временных рядах

Теорема Гаусса-Маркова

4. **Дисперсия** ошибок **одинакова**

- Иначе говоря, данные гомоскедастичны;
- Разброс ошибок в среднем постоянен



Теорема Гаусса-Маркова

5. Ошибки **не зависят от наблюдений**

- Наблюдается экзогенность;
- Предпосылка может нарушаться, если есть пропущенная переменная, одновременность, ошибки измерения…

Теорема Гаусса-Маркова

6. **Мат. ожидание** ошибок равно **нулю**

- Прочие факторы могут приводит к отклонениям в ту или другую сторону, но в среднем это влияние компенсируется

Unbiased estimation

<u>Утв</u>: $\hat{\theta}$ - несмещ. $\forall \varepsilon$ с $E(\varepsilon) = 0$

$$E(\hat{\theta}) = (X^T X)^{-1} X^T (X\theta + \overset{0}{\overset{\shortparallel}{E(\varepsilon)}}) =$$

$$= \underbrace{\underbrace{(X^T X)^{-1}}\ \underbrace{X^T X}}_{1}\theta + 0 = \theta$$

Best estimation

Утв= $\mathrm{Var}(\theta) = \sigma (X^T X)^{-1}$

$$\hat\theta = (X^T X)^{-1} X y \Rightarrow \hat\theta = (X^T X)^{-1} X (X\theta + \varepsilon) \Rightarrow \hat\theta = \underbrace{(X^T X)^{-1} X^T X}_{I}\, \theta + (X^T X)^{-1} X^T \varepsilon \Rightarrow$$

$$\Rightarrow \hat\theta = \theta + (X^T X)^{-1} X^T \varepsilon \Rightarrow \hat\theta - \theta = (X^T X)^{-1} X^T \varepsilon$$

$$\mathrm{Var}(\hat\theta - \theta) = E\left[\left((X^T X)^{-1} X^T \varepsilon\right)\left((X^T X)^{-1} X^T \varepsilon\right)^T\right] =$$

$$= E\left[(X^T X)^{-1} X^T \varepsilon \varepsilon^T X (X^T X)^{-1}\right] =$$

$$= (X^T X)^{-1} X^T \cdot E[\varepsilon \varepsilon^T] \cdot X (X^T X)^{-1} =$$

$$= \sigma \cdot \underbrace{(X^T X)^{-1} \cdot X^T X}_{I} \cdot (X^T X)^{-1} = \sigma (X^T X)^{-1}$$

ГОМОСКЕДАСТИЧНОСТЬ=

$$E[\varepsilon \varepsilon^T] = \begin{pmatrix} \sigma & 0 & \dots & 0 \\ 0 & \ddots & & 0 \\ \vdots & & & \vdots \\ 0 & \dots & & \sigma \end{pmatrix}$$

Если $\varepsilon \sim N(0, \sigma^2)$, то

$$\hat{\theta} \sim (\theta, \sigma^2 (X^T X)^{-1})$$

Проверка базовых гипотез в линейной регрессии

Гипотеза о наличии связи

$$y_i = \beta_0 + \underbrace{\beta_1}_{} X_1$$

$$H_0 = \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

Гипотеза о наличии связи

$$\text{МНК}: \sum (y_i-\hat{y}_i)^2 = \sum (y_i-\beta_0-\beta_1 X_i)^2 \xrightarrow[\beta_0,\beta_1]{} \min$$

$$\begin{cases} \dfrac{\partial}{\partial \beta_0} = -2\sum (y_i-\beta_0-\beta_1 X_i) = 0 \\ \dfrac{\partial}{\partial \beta_1} = -2X_i\sum (y_i-\beta_0-\beta_1 X_i) = 0 \end{cases} \Rightarrow \begin{cases} \sum y - n\beta_0 - \beta_1 \sum x = 0 \\ \sum xy - \beta_0 \sum x - \beta_1 \sum x^2 = 0 \end{cases} \Rightarrow$$

$$\beta_0 = \frac{\sum y}{n} - \beta_1 \frac{\sum x}{n} = \bar{y} - \beta_1 \bar{x}$$

$$\Rightarrow \sum xy - \bar{y}\sum x + \beta_1 \bar{x} \sum x - \beta_1 \sum x^2 \Rightarrow \beta_1 = \frac{\sum (y_i-\bar{y}) X_i}{\sum (X_i-\bar{X}) X_i}$$

Гипотеза о наличии связи

$$\beta_1 = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

cov_{xy}

$$cov_{xy} = 0 \longleftrightarrow \beta_1 = 0$$

Гипотеза о наличии связи

$$T = \frac{\hat{\beta}_1}{\hat{\sigma}(\hat{\beta})} \sim t_{n-2}$$

Квадраты остатков

$$RSS = \sum (y_i - \hat{y}_i)^2$$

residual

Суммарная квадратичная оцененная ошибка прогноза полной модели лин. регрессии

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

total

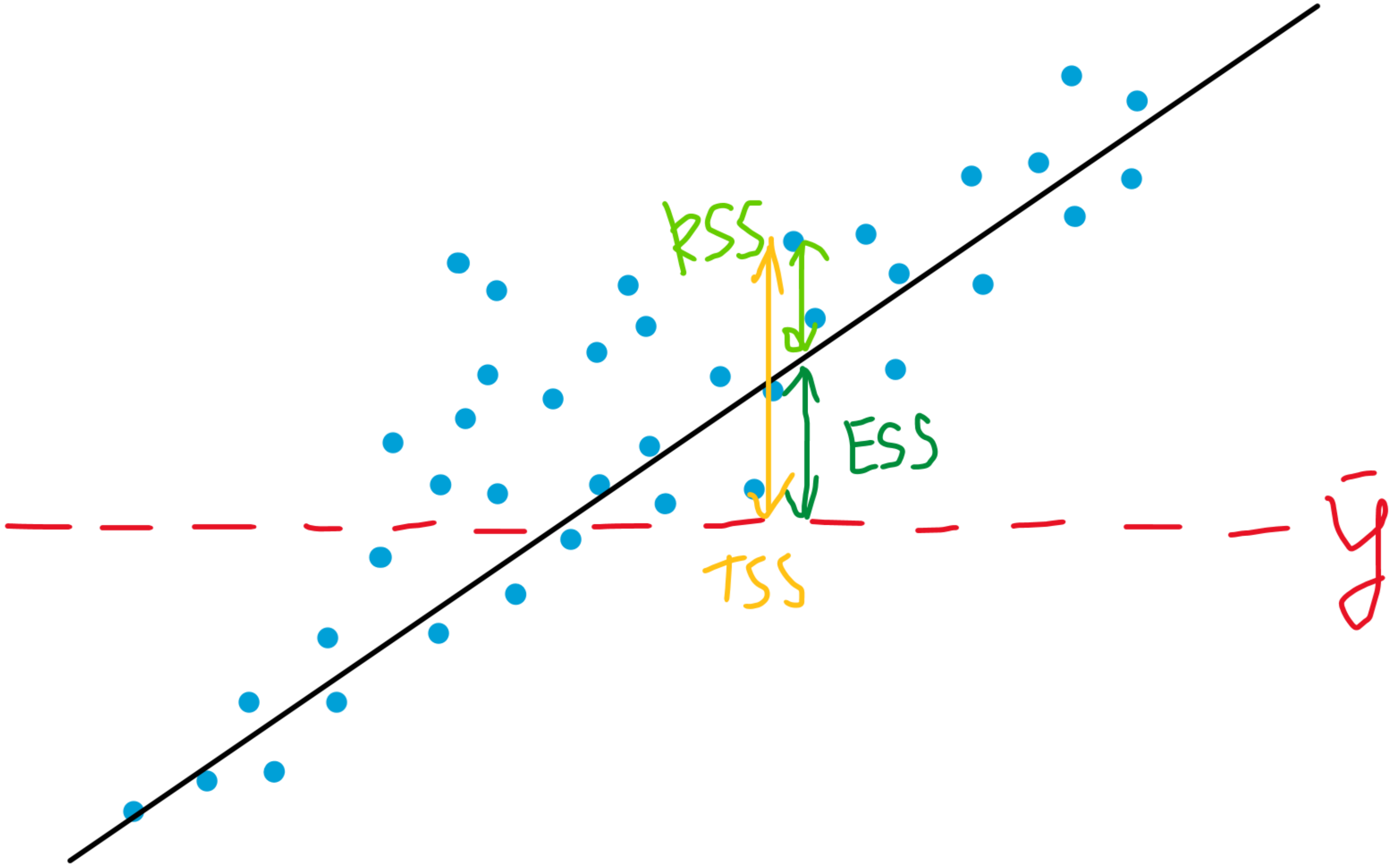
Суммарная квадратичная ошибка тривиального бейзлайна (просто выборочное среднее Y)

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

explained

Различие прогнозов полной модели и тривиального бейзлайна

Квадраты остатков



Коэффициент детерминации

$$R^2 = \frac{ESS}{TSS} \in [0, 1]$$

Доля объясненной более сложной моделью ошибки в ошибке тривиального бейзлайна

- С осторожностью стоит использовать R^2 для рассуждений о «качестве» модели. R^2 скорее показывает полноту спецификации;
- R^2, как он есть, также нельзя использовать для сравнения моделей между собой.

Гипотеза о многих параметрах

А модель в целом «адекватная»?
Или её надо полностью поменять?

Идея: чем больше R^2, тем полнее модель. На этом сконструируем наш тест

$$H_0 = \theta_i = \theta_j = 0, \; i \neq j$$

Все коэффициенты модели равны нулю: модель «бесполезная»

$$H_1 = \forall \theta_i \neq 0$$

Хотя бы один коэффициент в модели отличен от нуля

Если отклоняем H0, то хотя бы один «адекватный» параметр в нашей модели есть. Конкретнее о каждом параметре можно узнать, проведя на нем уже известный нам z-тест или t-тест в зависимости от кол-ва наблюдений

Гипотеза о многих параметрах: F-тест (критерий Фишера)

$$F = \frac{R^2/(K-1)}{(1-R^2)/(n-K)} \sim F_{K-1,\, n-K}$$

гаусс. ошибки

K - кол-во параметров

$$(K-1)F \xrightarrow[n\to\infty]{as} \chi^2_{K-1}$$

негаусс. ошибки

Скорректированный коэффициент детерминации

$$R^2_{adj} = 1 - \frac{n-1}{n-k}(1 - R^2)$$

- Штрафует за добавление дополнительных переменных;
- По скорректированному коэффициенту детерминации можно сравнивать модели между собой

Зачем нужна линейная регрессия?

Задача описания
(как устроен мир?)

VS

Задача предсказания
(что будет дальше?)

Эконометрика

Основное: интерпретируемость, отсюда идет борьба за предпосылки
Есть «арсенал» для обоснования адекватности полученной модели

Machine Learning

Основное: обобщающая способность, хорошее качество прогноза на новых данных

На стыке:
интерпретируемый ML

Дополнительно

- Гайд по интерпретируемому ML: https://christophm.github.io/interpretable-ml-book/;
- Примеры в Python: https://towardsdatascience.com/explainable-artificial-intelligence-part-3-hands-on-machine-learning-model-interpretation-e8ebe5afc608/
- ВШЭ, Прикладная статистика (week 13) https://www.youtube.com/watch?v=OhKVEDPvtPw&list=PLCf-cQCe1FRw6XWyfIfL84-W-BVz9Q8js&index=10
- Подробно о линреге в ML и проблемах с аналитическим решением: https://education.yandex.ru/handbook/ml/article/linear-models

ВЫСШАЯ • ШКОЛА • ЭКОНОМИКИ •