Машинное обучение в гидрологии

Регрессия

Содержание лекции

1. Регрессия
2. Линейная регрессия
   - Постановка задачи
   - Оценка коэффициентов
     - МНК
     - Метод максимального правдоподобия
3. Оценка качества модели
4. Регуляризация
   - Гребневая регрессия
   - Лассо регрессия
5. Отбор признаков для модели

Как ставится задача регрессии

$\mathrm{X} = \mathbb{R}^n$ – множество объектов

$Y = \mathbb{R}$ – множество ответов

y : X → Y – неизвестная зависимость (target function)

Дано:

$X_{obs} = \{x_1, \dots x_N\} \subset \mathrm{X}$ – уже наблюдаемые объекты

$y_i = y(x_i),\ i = \{1, \dots N\}$ – известные ответы

Найти:

$a: X \to Y$ – алгоритм, решающую функцию (decision function) наилучшим образом приближающую **y** на всем множестве **X**

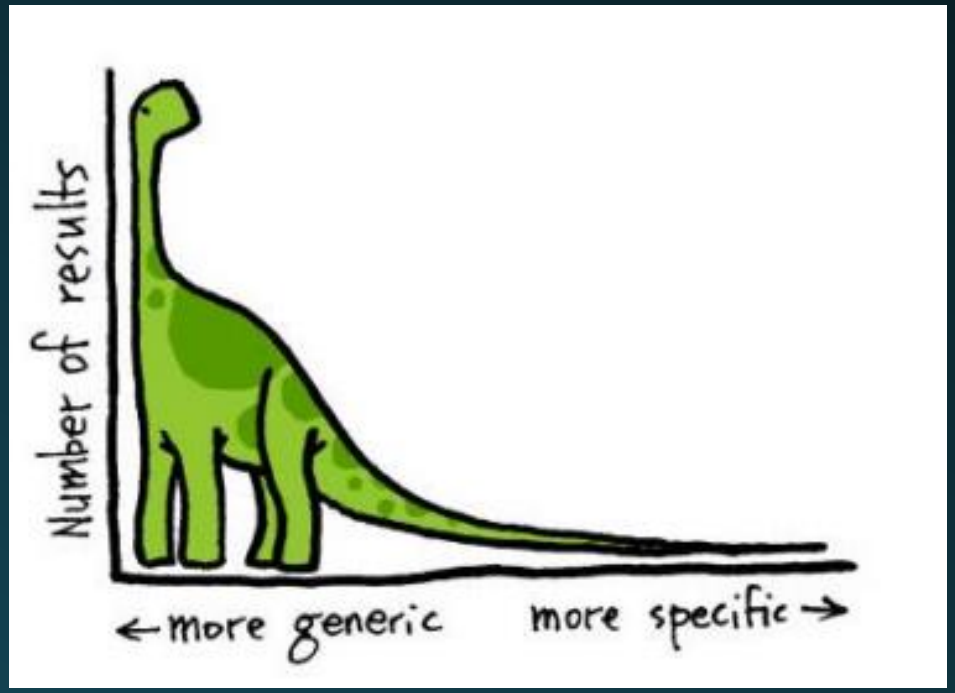
Задача предсказания расхода

Выборка: Наблюдения по которым достоверно известен расход

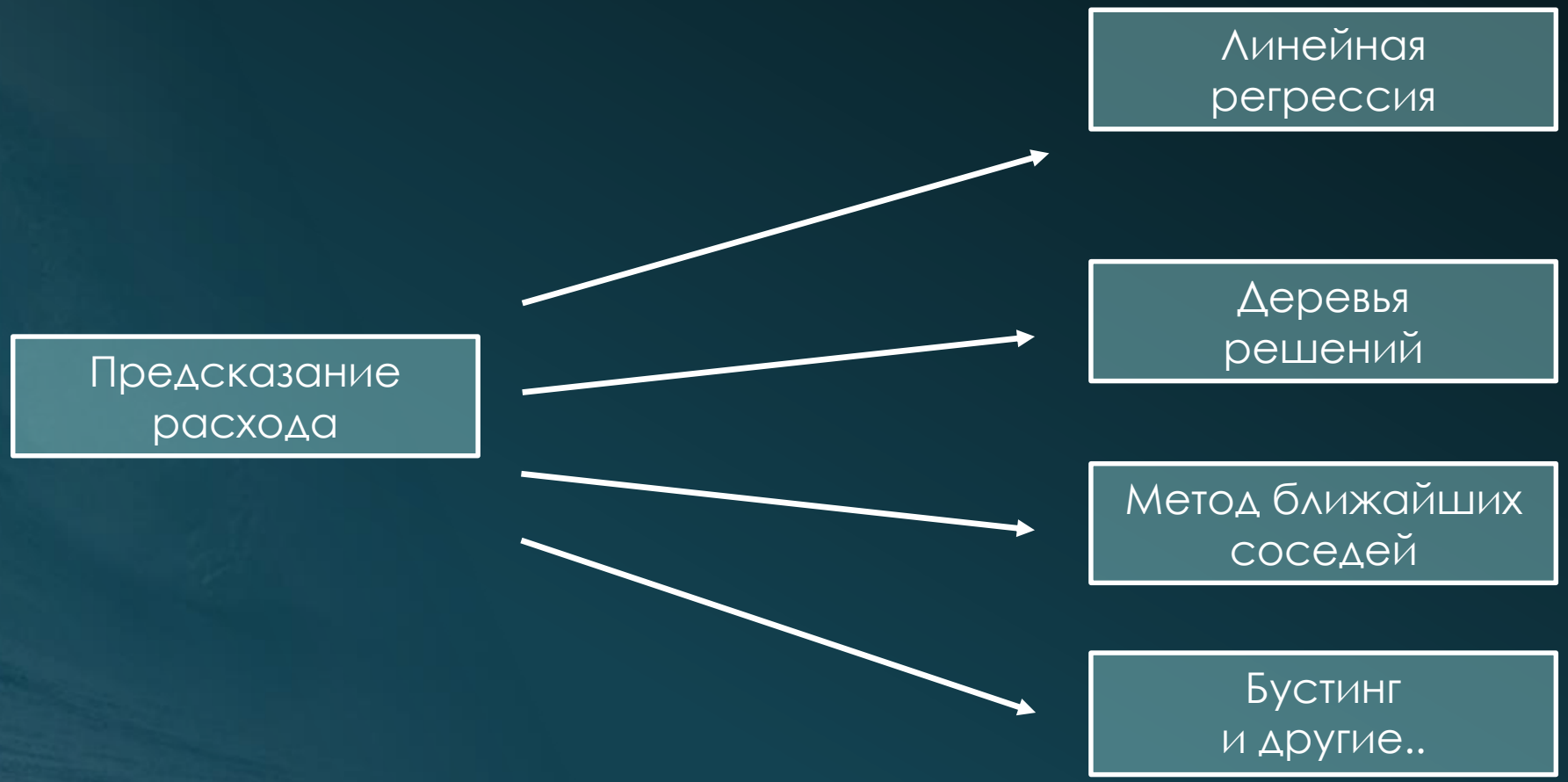
- $X_1, \dots X_n \in \mathbb{R}^m$ – наблюдения (вектора признаков)
- $y_1, \dots y_n \in \mathbb{R}$ – непрерывная целевая переменная (расход)
- Делаем допущение, что (X_i, y_i), $i = 1, \dots n$ – независимы и одинаково распределены

В чем особенности задачи?

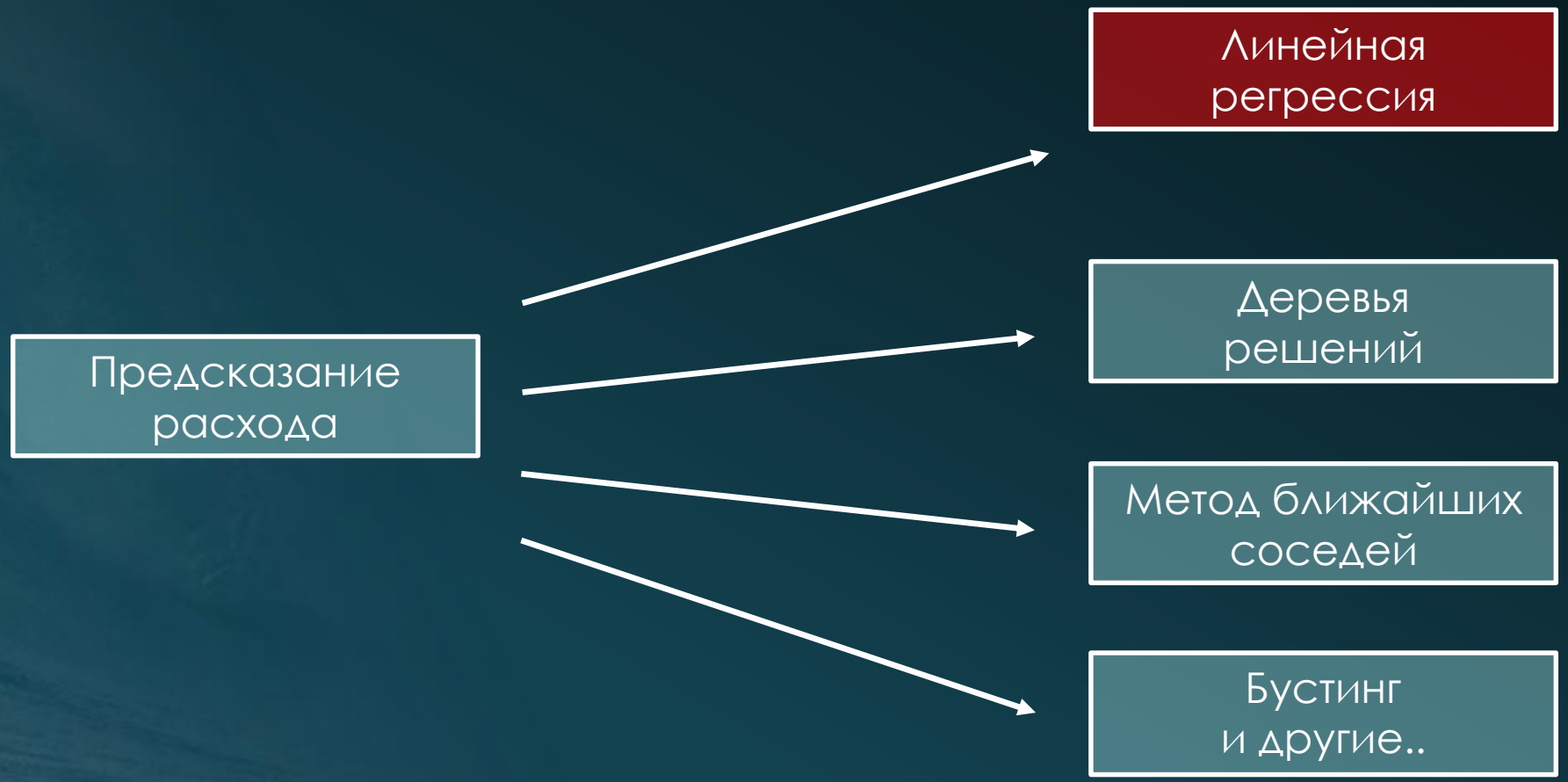
- Может быть мало данных или данные с пропусками
- Расход может сильно варьироваться
- «Хвост» распределения может быть очень длинным



Какой метод выбрать?



Какой метод выбрать?



Линейная регрессия

Модель линейной регрессии

Основное предположение модели: y – линейная функция от признаков $(X_1, \dots X_m)$

$$y = a_0 + a_1 X_1 + \dots + a_m X_m + \varepsilon$$

- $a_0, a_1, \dots, a_m$ - набор констант (веса модели или параметры)
- ε – случайная величина (ошибка)

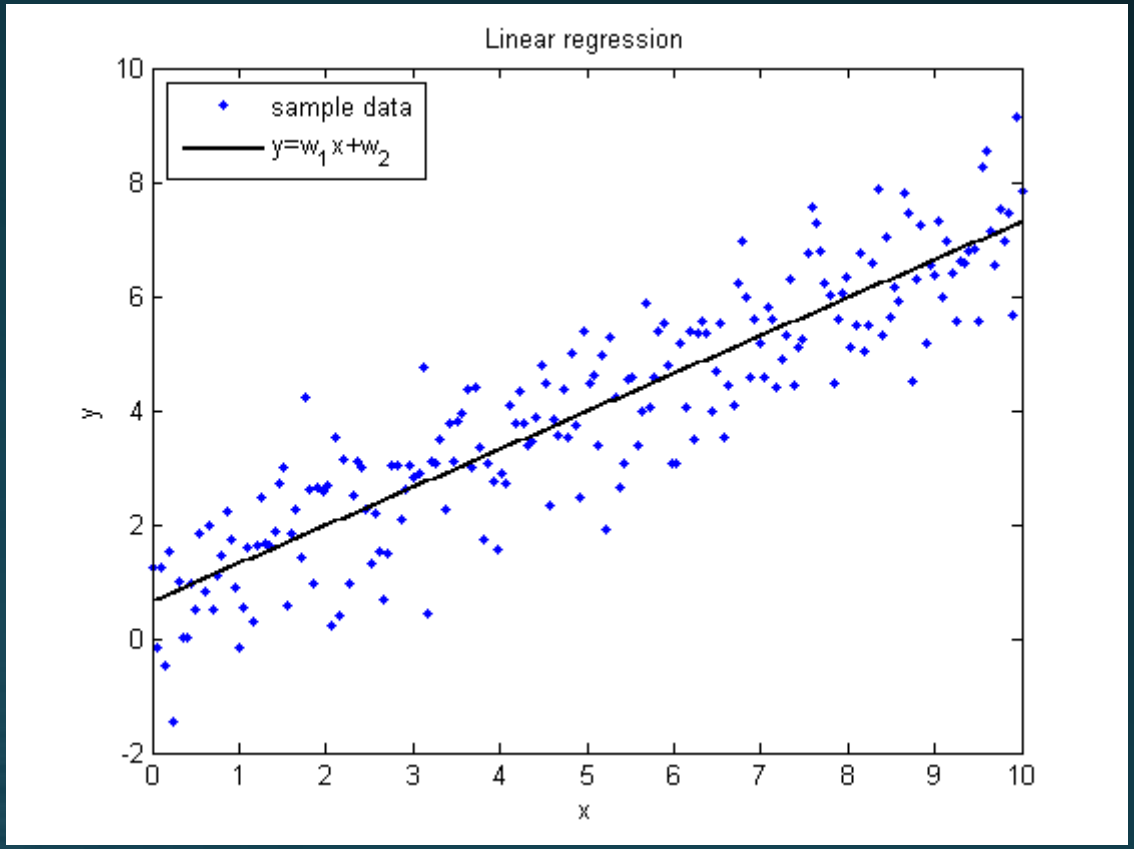
Для i-ого наблюдения: $y_i = a_0 + a_1 * x_{i,1} + \dots + a_m * x_{i,m} + \varepsilon_i$

Матричная запись: $y = Xa + \varepsilon$, $y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, a = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$

$$X = \begin{bmatrix} 1 & \cdots & x_{1,m} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_{n,m} \end{bmatrix} = \begin{bmatrix} x_{1,0} & \cdots & x_{1,m} \\ \vdots & \ddots & \vdots \\ x_{n,0} & \cdots & x_{n,m} \end{bmatrix}, \text{ и } x_{i0} = 1$$

Постановка задачи

- По наблюдениям построить оценки весов $a_0, a_1, \dots a_m$
- Оценить ошибки весов и ошибку предсказаний модели



Чего-то не хватает….

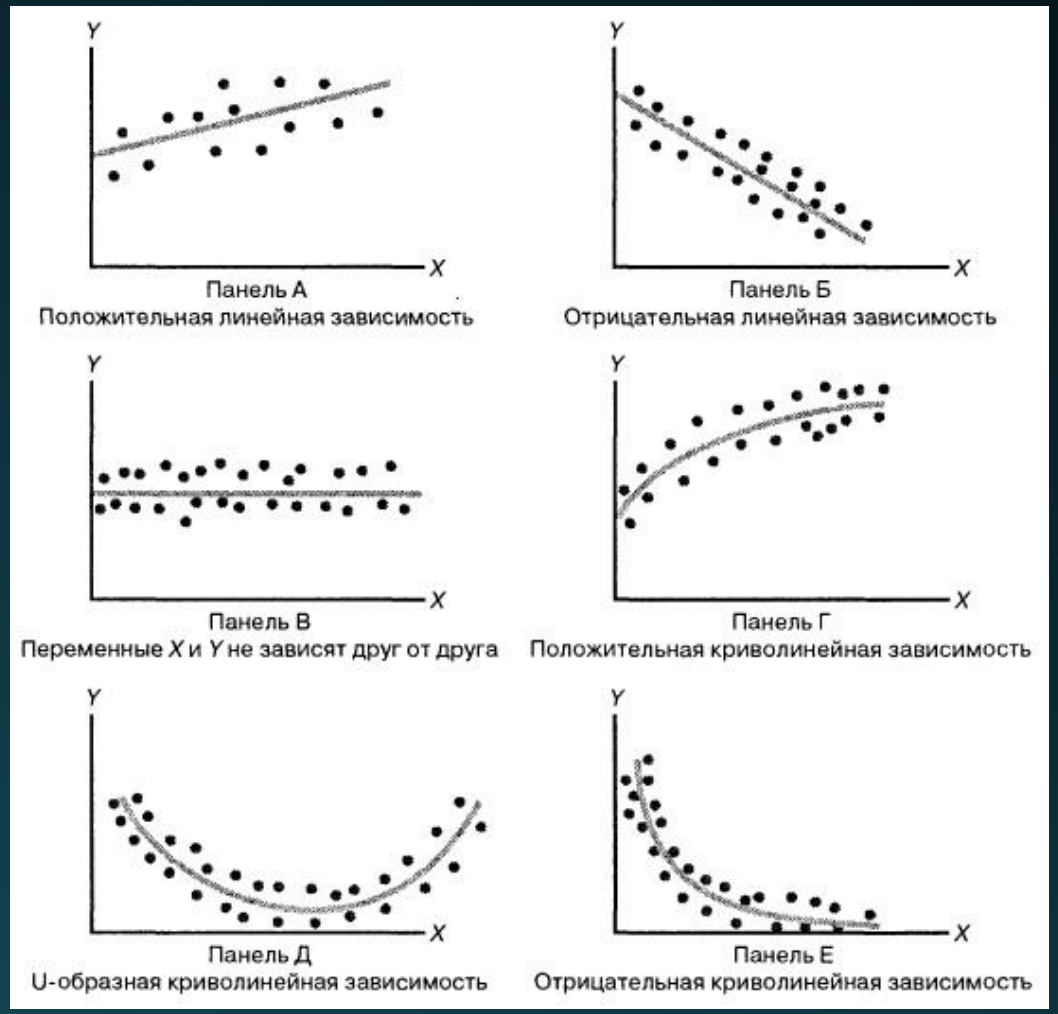
Не может быть все так просто в постановке задачи

Когда можно применять модель линейной регрессии?

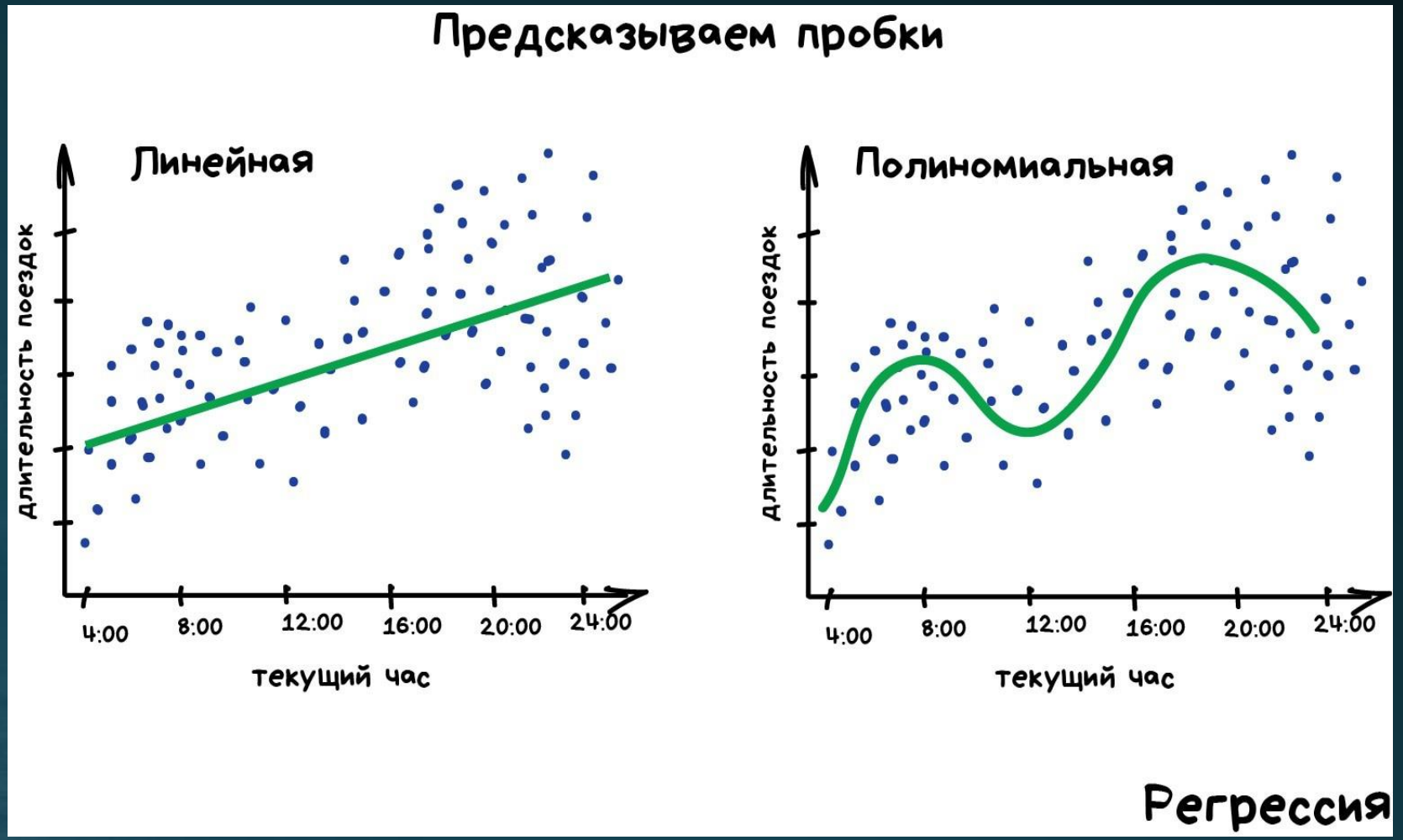
Признаки должны хоть отдаленно линейно зависеть от целевой переменной и не зависеть друг от друга)

Что делать если это не так?

1. Исключить такой признак из модели
2. Строить другую модель
3. Преобразовать «неподходящие» признаки
   - $x \to x^{\alpha}$ (полиномиальная модель)
   - $x \to \log(x)$
   - $x \to e^{\alpha x}$
4. Сделать несколько признаков из одного



Когда можно применять модель линейной регрессии?



Дополнительные условия

Дополнительно на модель надо наложить следующие ограничения (проверяется после подбора параметров модели):

- Математическое ожидание случайных ошибок равно 0

$$\forall i: \quad E[\varepsilon_i] = 0$$

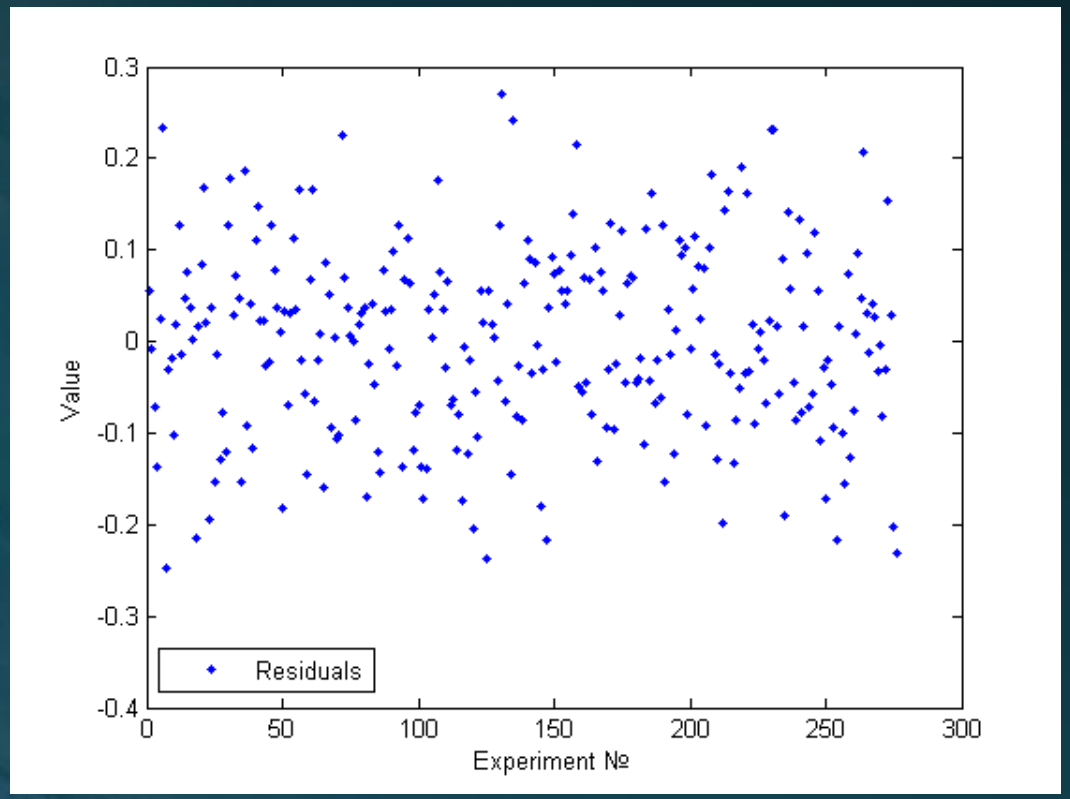
- Дисперсия случайных ошибок одинакова и конечна (гомоскедастичность)

$$\forall i: Var(\varepsilon_i) = \sigma^2 < \infty$$

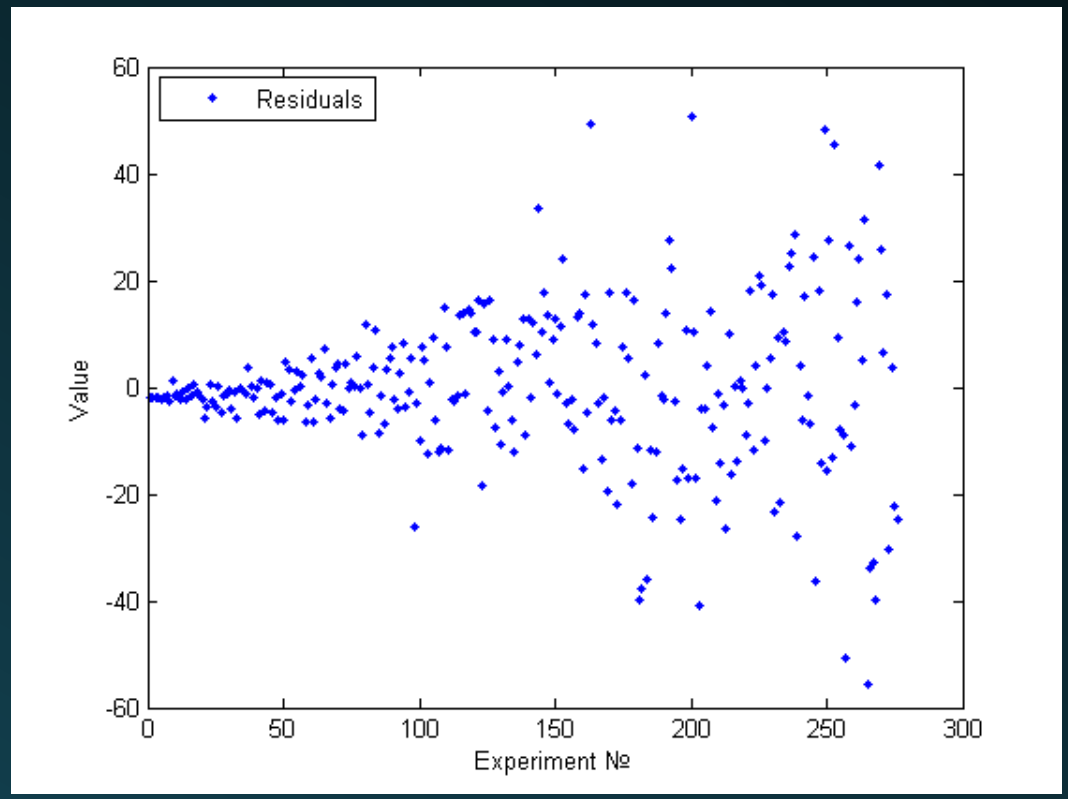
- Случайные ошибки не скоррелированы (независимы)

$$\forall i \neq j: \quad cov(\varepsilon_i, \varepsilon_j) = 0$$

Дополнительные условия

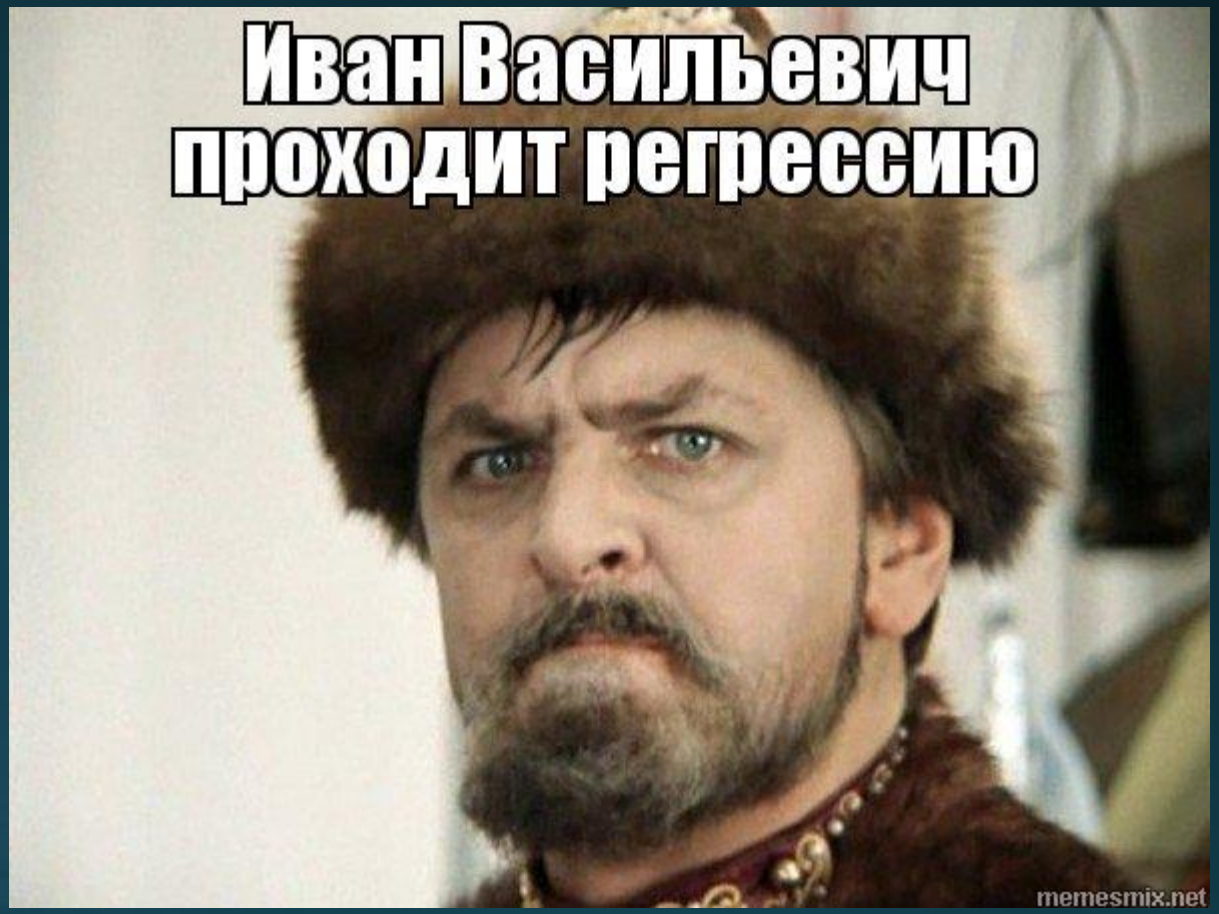


Гомоскедастичность



Гетероскедастичность

Оценка коэффициентов



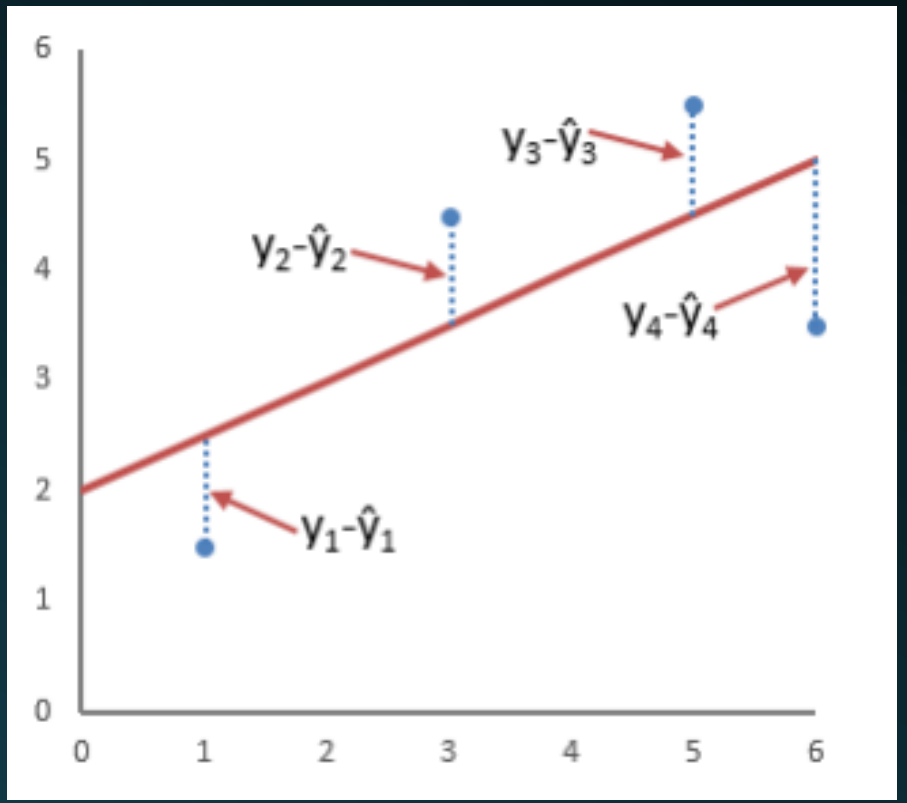
Метод наименьших квадратов

Пусть $\tilde{a}$ - некая оценка коэффициентов.

Квадратичный функционал ошибки:

$$Q(\tilde{a}) = \left|\left|\hat{\varepsilon}\right|\right|^2 = \left|\left|\hat{y}(\tilde{a}) - y\right|\right|^2 = \left|\left|X\tilde{a} - y\right|\right|^2 = \\ = \frac{1}{2n}(y - Xa)^T(y - Xa)$$

МНК: минимизации $Q(\tilde{a})$



Оценка коэффициентов с помощью

МНК – значения аргументов, на которых квадратичный функционал ошибки принимает наименьшее значение

$$\hat{a} = arg\min_a Q(a)$$

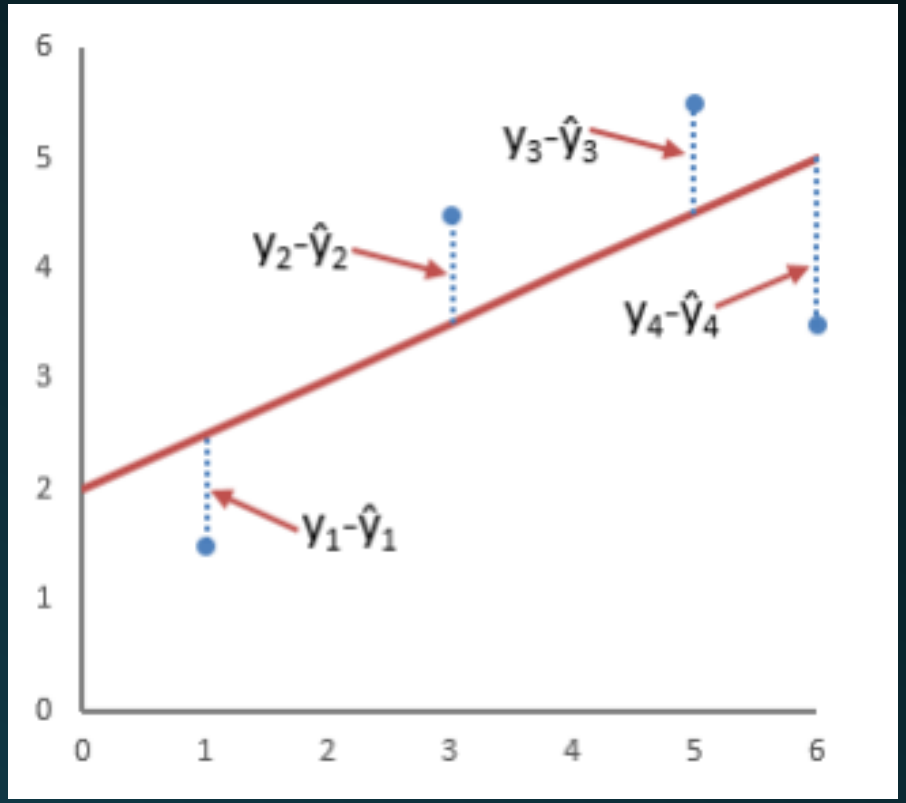
Метод наименьших квадратов

И на самом деле, в случае задачи линейной регрессии существует решение задачи минимизации в явном виде!

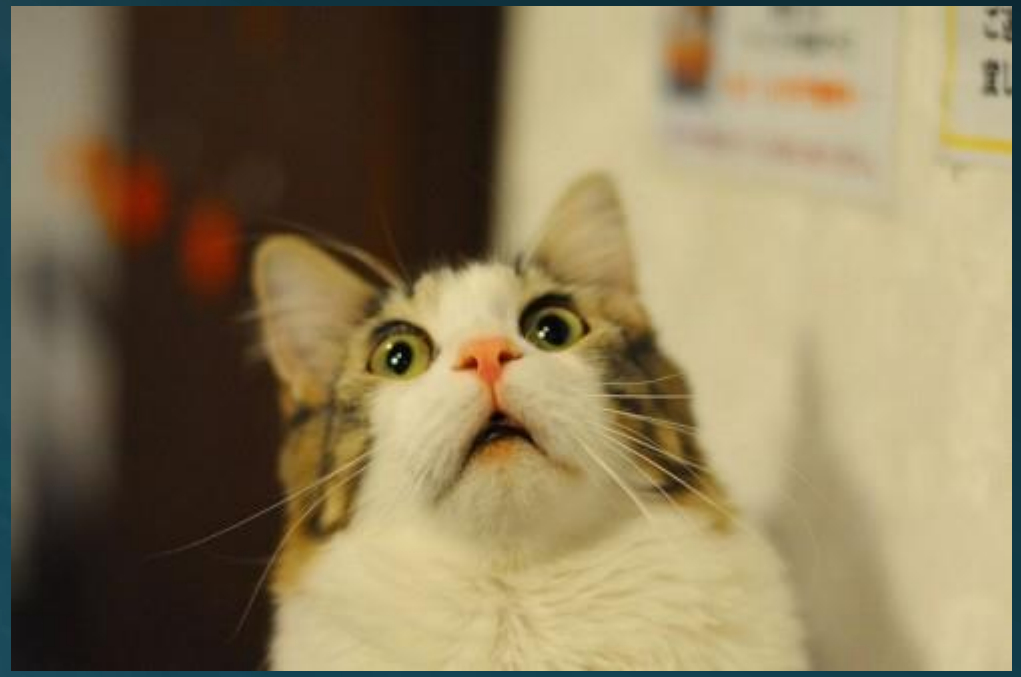
Верно следующее:

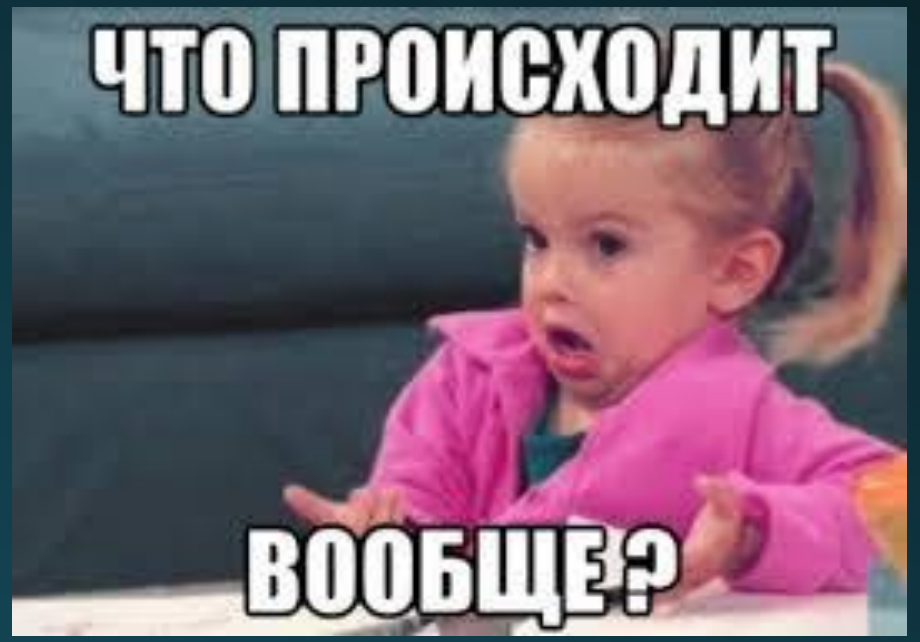
Утверждение: Если для матрицы $X^T X$ существует обратная, то существует единственное решение задачи $Q(a) \to min$:

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



Оценка коэффициентов





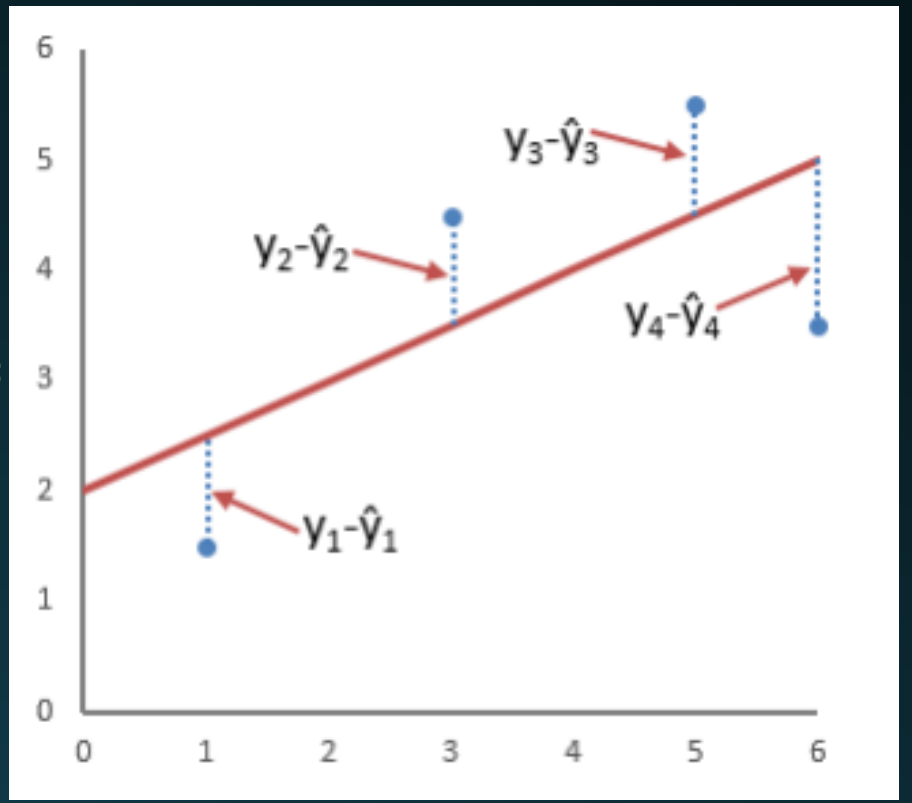
Покажу на простом примере =)

Будем искать зависимость y = ax+b

Квадратичный функционал ошибки:

$$Q(\tilde{a}) = \left|\left|\hat{\varepsilon}\right|\right|^2 = \left|\left|\hat{y}(\tilde{a}) - y\right|\right|^2 = \left|\left|ax + b - y\right|\right|^2 = \sum (ax_i + b - y_i)^2 \to min$$

Чтобы вычислить минимум что нужно сделать?

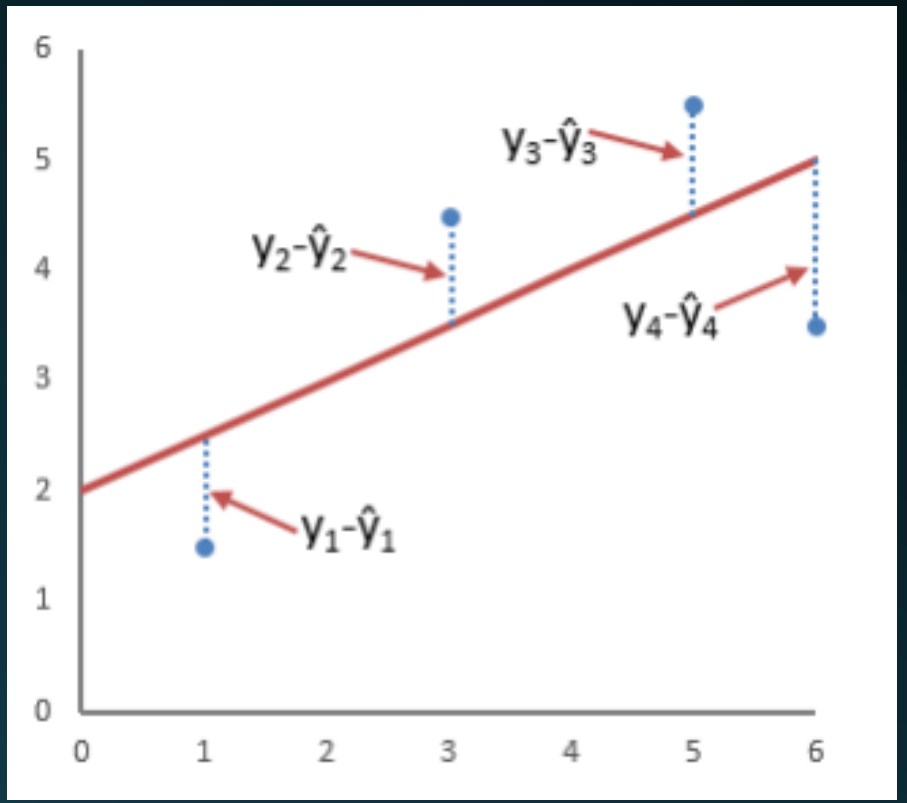


Покажу на простом примере =)

Будем искать зависимость y = ax+b

Квадратичный функционал ошибки:

$$Q(\tilde{a}) = \left|\left|\hat{\varepsilon}\right|\right| = \left|\left|\hat{y}(\tilde{a}) - y\right|\right| = \left|\left|ax + b - y\right|\right| =$$
$$= \sum (ax_i + b - y_i)^2 \to min$$



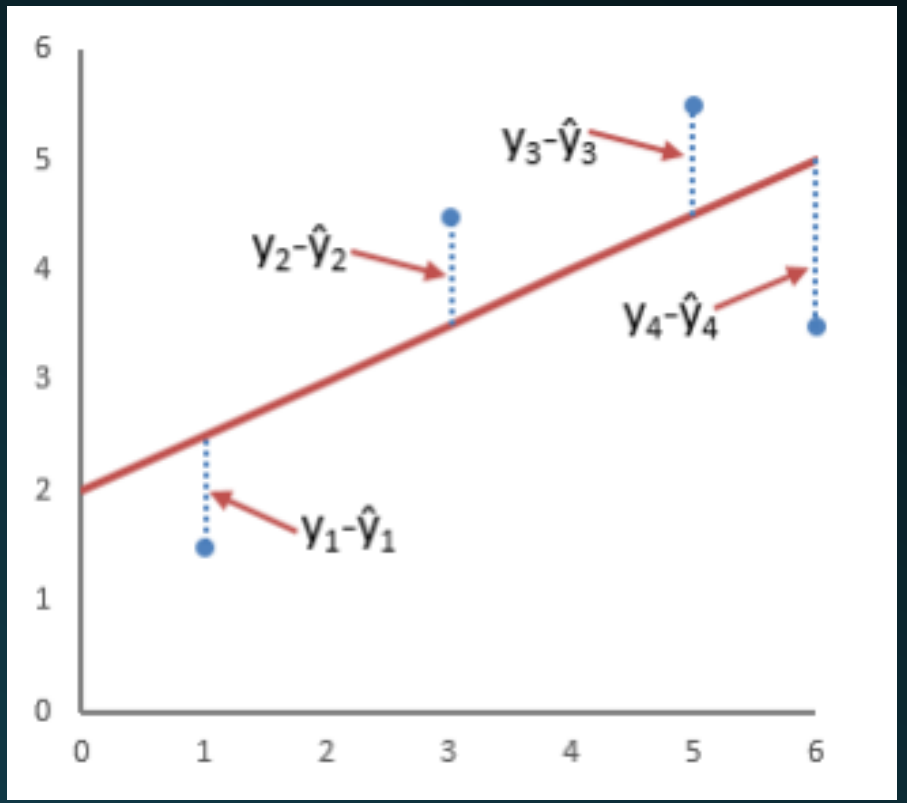
Чтобы вычислить минимум что нужно сделать?

Посчитать производные и приравнять к 0 =)

Покажу на простом примере =)

$$Q(\tilde{a}) = \left|\left|\hat{\varepsilon}\right|\right| = \left|\left|\hat{y}(\tilde{a}) - y\right|\right| = \left|\left|ax + b - y\right|\right| =$$
$$= \sum (ax_i + b - y_i)^2 \to min$$

$$\begin{cases} \frac{\partial Q}{\partial a} = 0 \\ \frac{\partial Q}{\partial b} = 0 \end{cases} <=> \begin{cases} \sum 2x_i(ax_i + b - y_i) = 0 \\ \sum 2(ax_i + b - y_i) = 0 \end{cases}$$



Отсюда находим значения коэффициентов регрессии:

$$\hat{b} = \bar{y} - \hat{k}\bar{x}$$

$$\hat{k} = \frac{\sum x_i y_i - \frac{1}{n}\sum y_i \sum x_i}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}$$

Чем хороша полученная оценка?

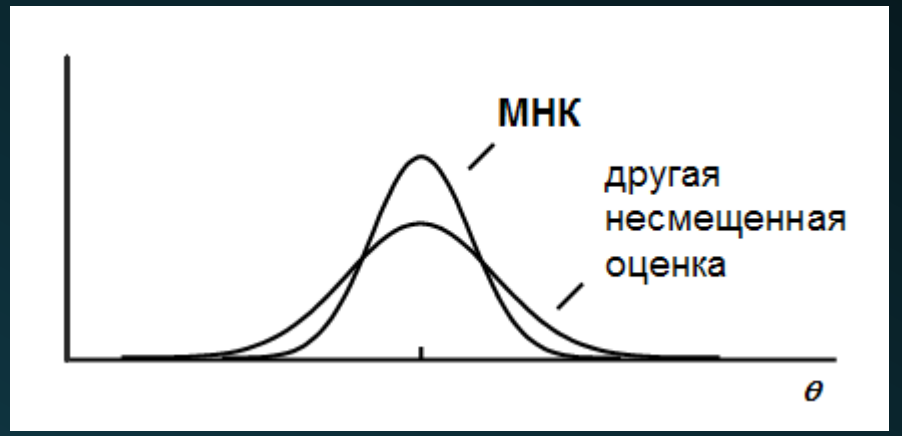
Пусть $\hat{a}$ - оценка полученная с помощью МНК.

Тогда:

1. Оценки МНК несмещенные
$$E\hat{a} = a$$
2. Оценки МНК эффективны в классе линейных оценок

Для любой линейной несмещенной оценки $\hat{b}$
$$D_a(\hat{a}) \leq D_a(\hat{b})$$

3. Оценки МНК состоятельны
$$\hat{a}_n \xrightarrow{P} a \quad \text{или} \quad P(|\widehat{a_n} - a| > \varepsilon) \to 0, n \to \infty$$



Вероятностная интерпретация

Метод максимального правдоподобия

Пусть $X_1, X_2, .., X_n$ - независимые одинаково распределенные случайные величины (наблюдения), распределение которых задается параметром θ

Функция распределения $P_\theta(x) = P(x|\theta)$.

Правдоподобием $\mathcal{L}$ называется вероятность появления фиксированной наблюдаемой выборки, как функции от параметра θ.

$$\mathcal{L}(\theta) = P(X_1 = x_1, \dots X_n = x_n \mid \theta) = \prod_{i=1}^{n} P_\theta(x_i)$$

ММП заключается в максимизации функции правдоподобия по θ, т.е. поиска такого параметра θ, при котором появление наблюдаемой выборки будет наиболее вероятным.

Оценка $\hat{\theta}$,на которой достигается максимум, называется оценкой максимального правдоподобия

Метод максимального правдоподобия

Если предположить, что ошибки в модели $\boldsymbol{y = Xa + \varepsilon}$ имеют нормальное распределение: $\varepsilon \sim \boldsymbol{N_n(0, \sigma^2 I)}$, то неизвестные параметры - $\boldsymbol{a}$ и $\boldsymbol{\sigma^2}$.

И для случайных величин - ошибок можно выписать правдоподобие:

$$\boldsymbol{\mathcal{L}}(a, \sigma^2) = \prod_{i=1}^{n} P_\theta(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{e_i^2}{2\sigma^2}\right) = \frac{1}{(2\pi)^{0.5n}\sigma^n} \exp\left(-(y - Xa)^T (y - Xa)/2\sigma^2\right)$$

Максимизация правдоподобия эквивалентна минимизации логарифма правдоподобия. А следовательно верно следующее утверждение:

Утверждение: Если существует обратная к матрице $X^T X$, то оценка ММП существует и совпадает с оценкой МНК.

Оценка качества модели

Вспомним, что было в задаче классификации…

Критерии качества. Классификация

- Хотим предсказывать класс нашего наблюдения.

$$Y = \{-1, 1\}$$

- Составляется матрица ошибок
- Метрика, оценивающая качество модели, выбирается исходя из потребностей задачи

	Y = 1	Y = -1
$\hat{Y} = 1$	True Positive (TP)	False Positive (FP)
$\hat{Y} = -1$	False Negative (FN)	True Negative (TN)

Ошибка 2 рода
«пропуск цели»

Ошибка 1 рода
«ложная тревога»

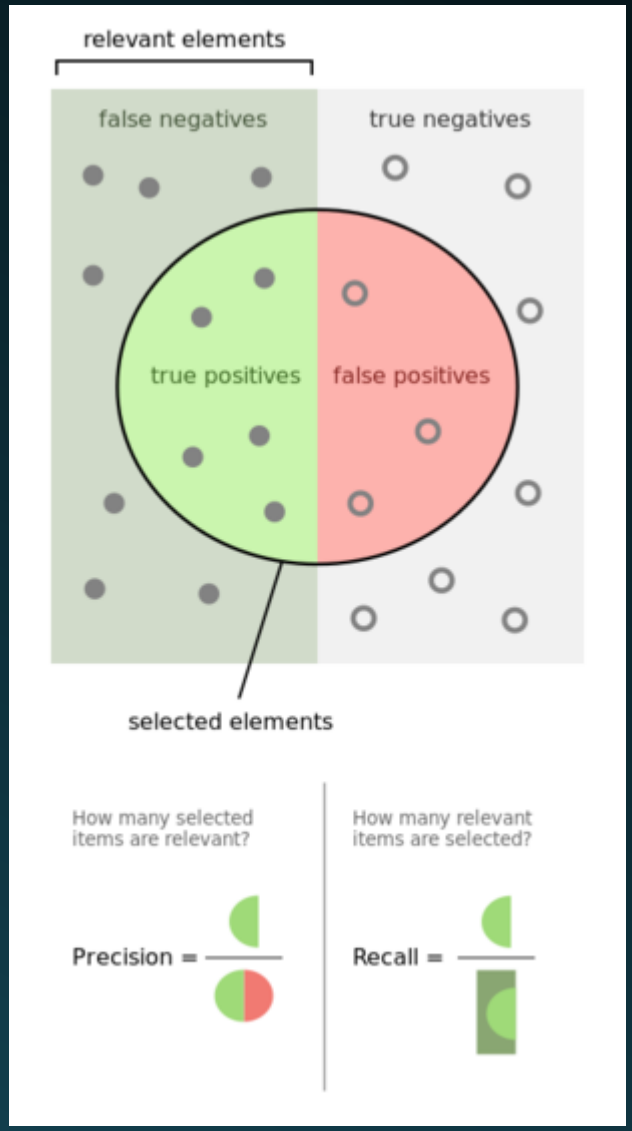
Метрики задач классификации

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$\text{Precision} = \frac{TP}{TP + FP} = PPV$$

$$Recall = \frac{TP}{TP + FN} = TPR$$

$$F_1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 * recall * precision}{precision + recall}$$



А как на счет регрессии?

Коэффициент детерминации

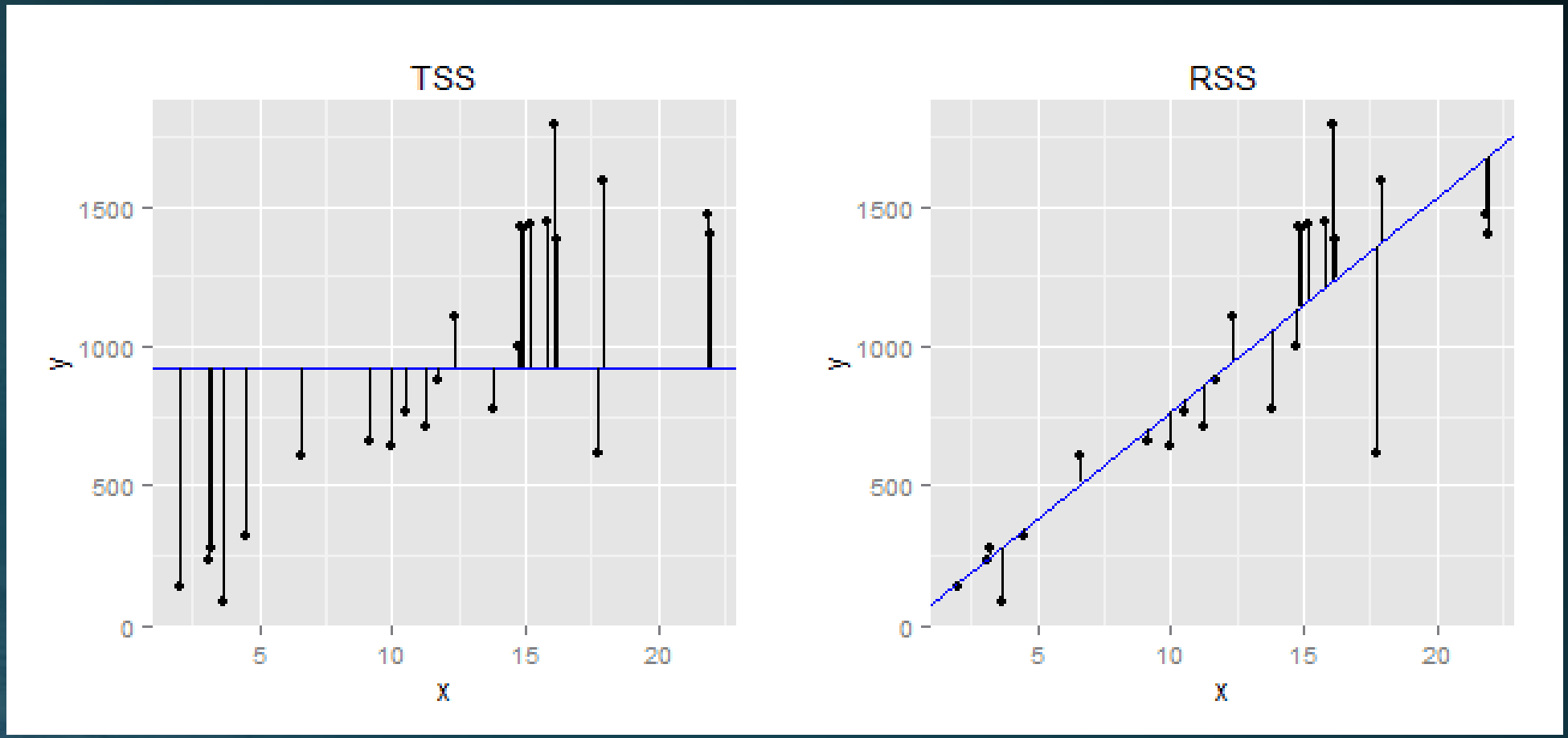
- Общая дисперсия зависимой переменной y имеет следующий вид (total sum of squares), где $\bar{y}$ - выборочное среднее

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- Сумма квадратов ошибок в оценке регрессии (residual sum of squares)

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

Коэффициент детерминации



Коэффициент детерминации

- Общая дисперсия зависимой переменной y имеет следующий вид (total sum of squares), где $\bar{y}$ - выборочное среднее

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- Сумма квадратов ошибок в оценке регрессии (residual sum of squares)

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

- Объясненная дисперсия (explained sum of squares)

$$ESS = \sum_{i=1}^{n} (\bar{y} - \hat{y}_i)^2$$

- **R^2 - коэффициент детерминации**

$$R^2 = 1 - \frac{RSS}{TSS}$$

Коэффициент детерминации

- Общая дисперсия зависимой переменной y имеет следующий вид (total sum of squares), где $\bar{y}$ - выборочное среднее

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- Сумма квадратов ошибок в оценке регрессии (residual sum of squares)

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

- Объясненная дисперсия (explained sum of squares)

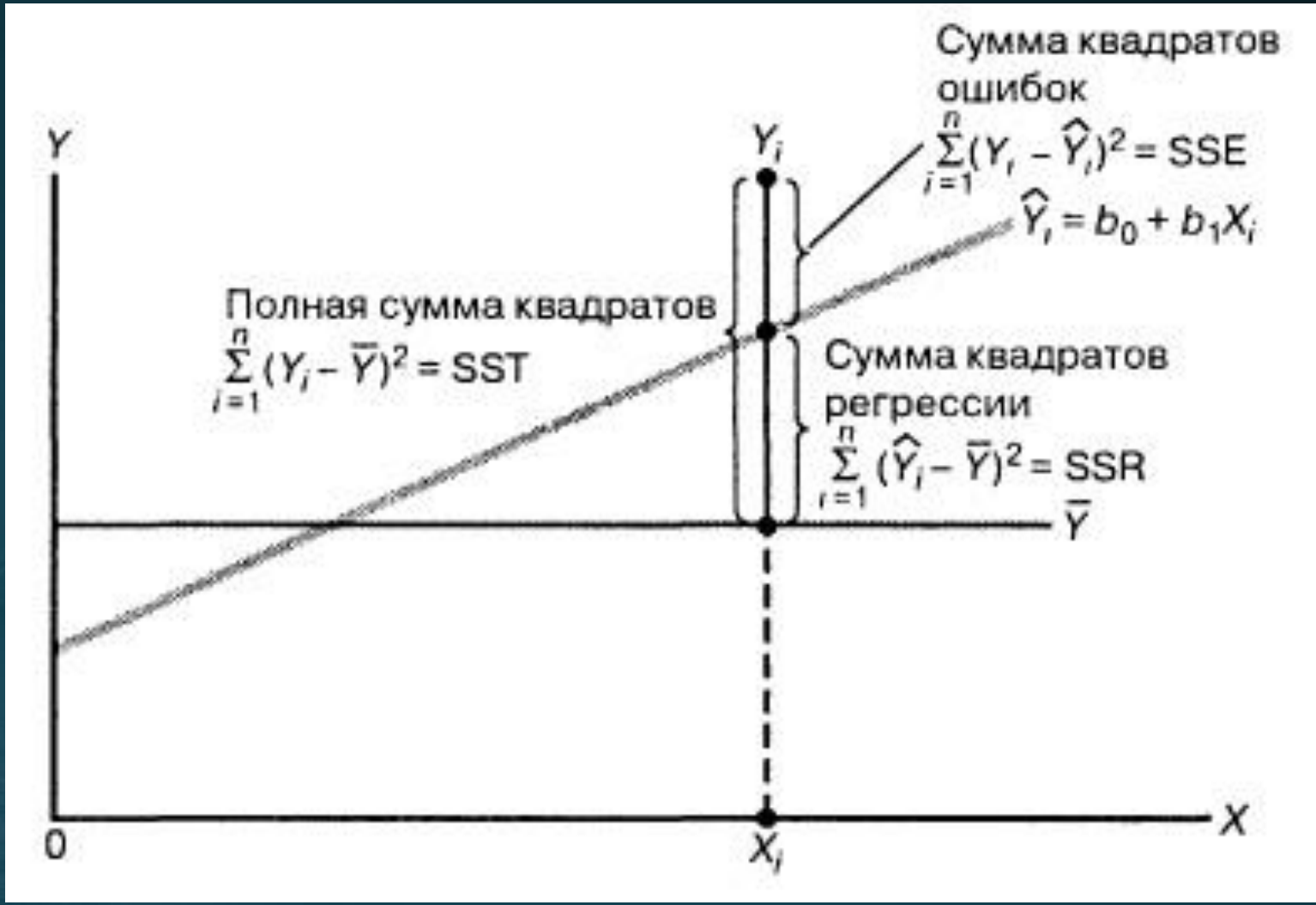
$$ESS = \sum_{i=1}^{n} (\bar{y} - \hat{y}_i)^2$$

- **R^2 - коэффициент детерминации**

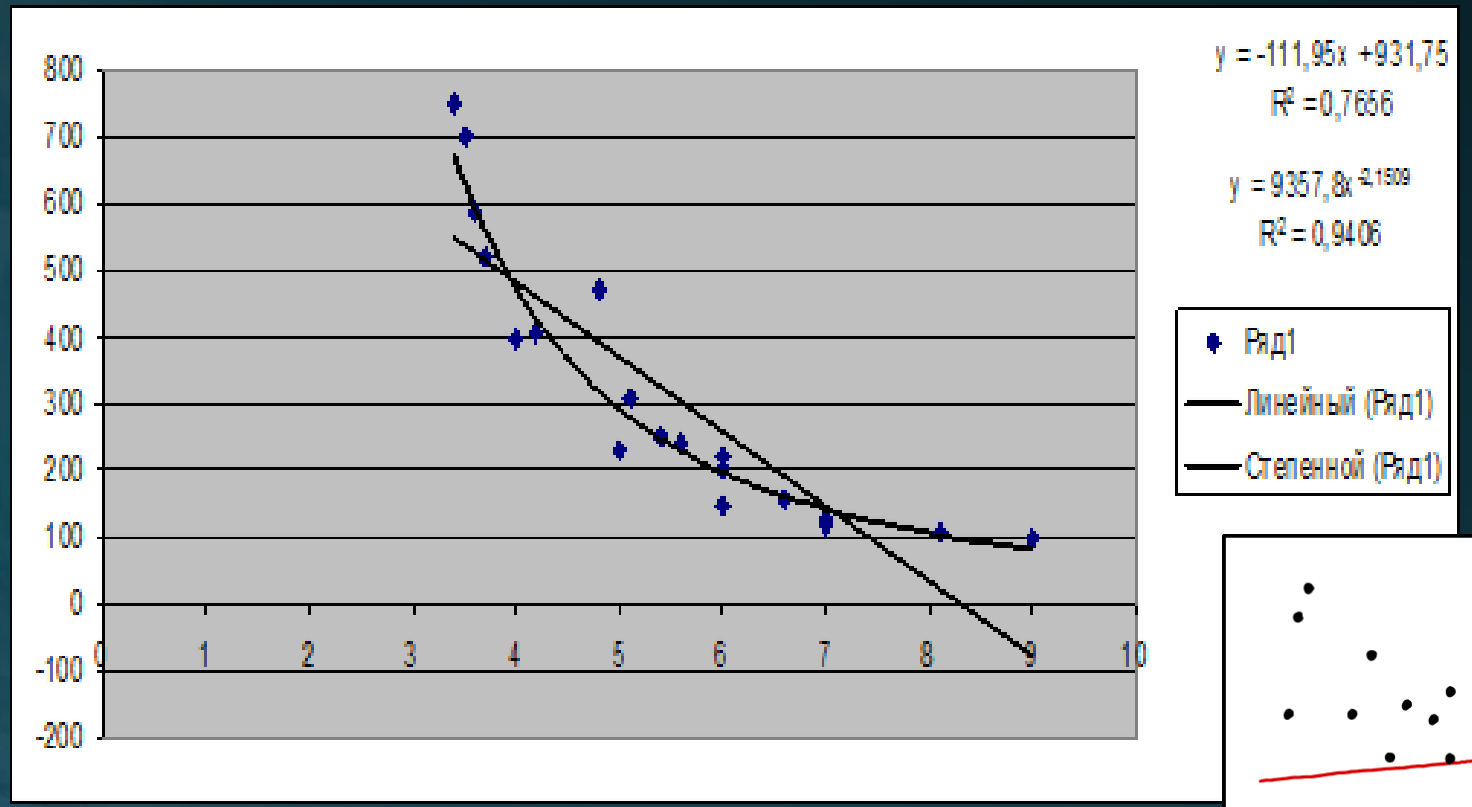
$$R^2 = 1 - \frac{RSS}{TSS}$$

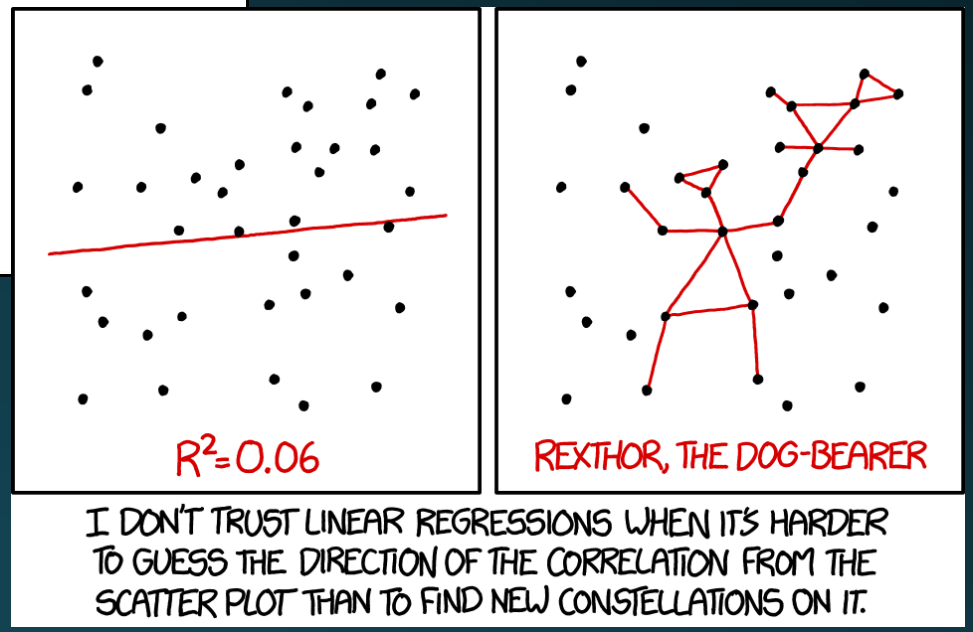
В случае прогнозирования расходов это и есть знаменитый коэффициент Нэша Сатклиффа

Коэффициент детерминации



Коэффициент детерминации





Метрика качества модели

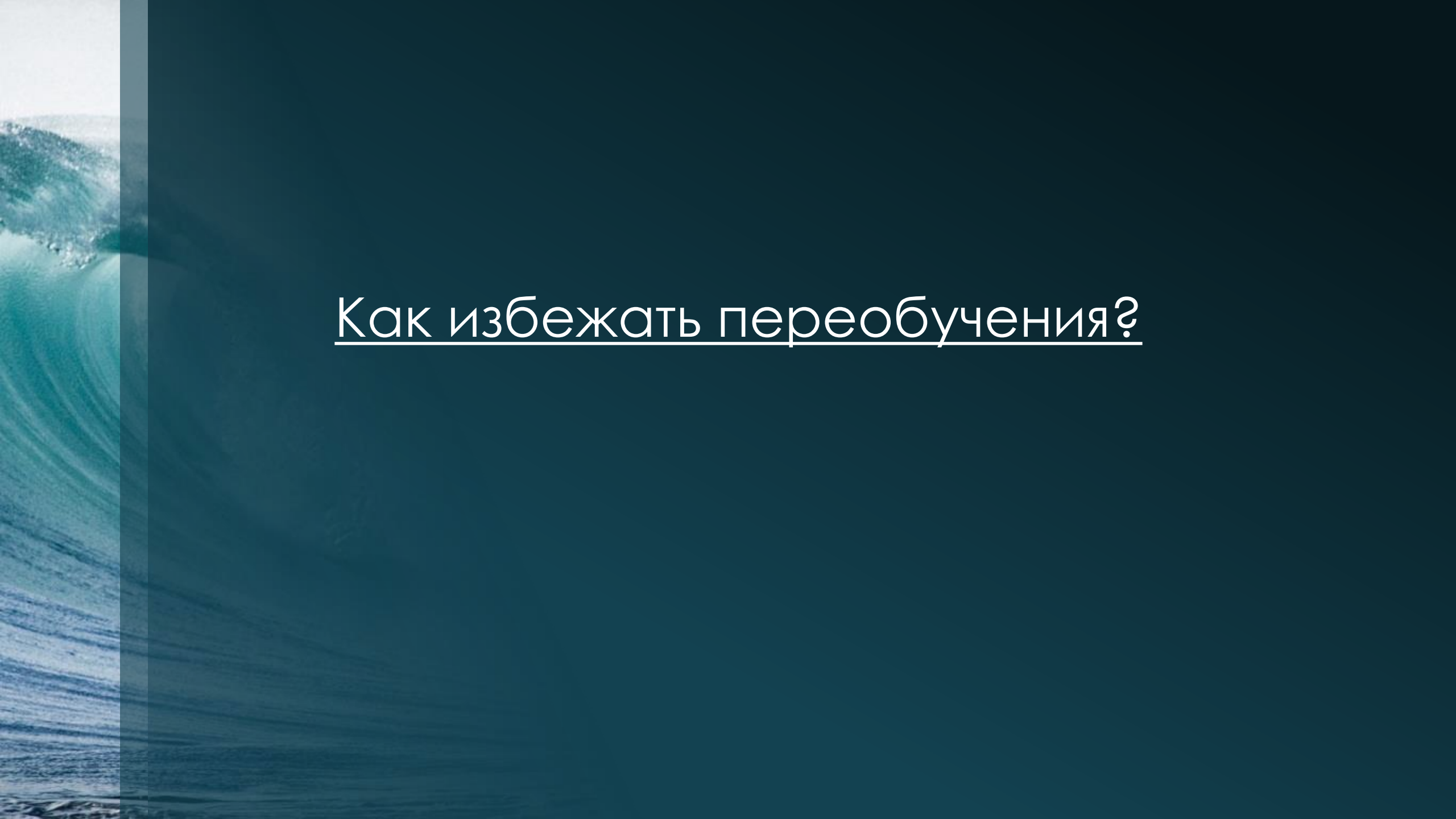
$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2,\ RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2,\ ESS = \sum_{i=1}^{n}(\bar{y} - \hat{y}_i)^2\ ,\ R^2 = 1 - \frac{RSS}{TSS}$$

Утверждения:

Если $\hat{y}_i$ - оценка МНК, то верны следующие свойства:

1. $TSS = ESS + RSS$
2. $R^2 = 1 - \frac{RSS}{TSS} = \frac{ESS}{ESS+RSS}$
3. $R^2 = 0 \Leftrightarrow \text{ESS} = 0 \Leftrightarrow \sum_{i=1}^{n}(\bar{y} - \hat{y}_i)^2 = 0 \Leftrightarrow \hat{y} = \bar{y}$ т.е. оценка – константа
4. $R^2 = 1 \Leftrightarrow \text{RSS} = 0 \Leftrightarrow \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 0 \Leftrightarrow \hat{y}_i = y_i$ для $\forall i$; т.е. оценка - идеальна

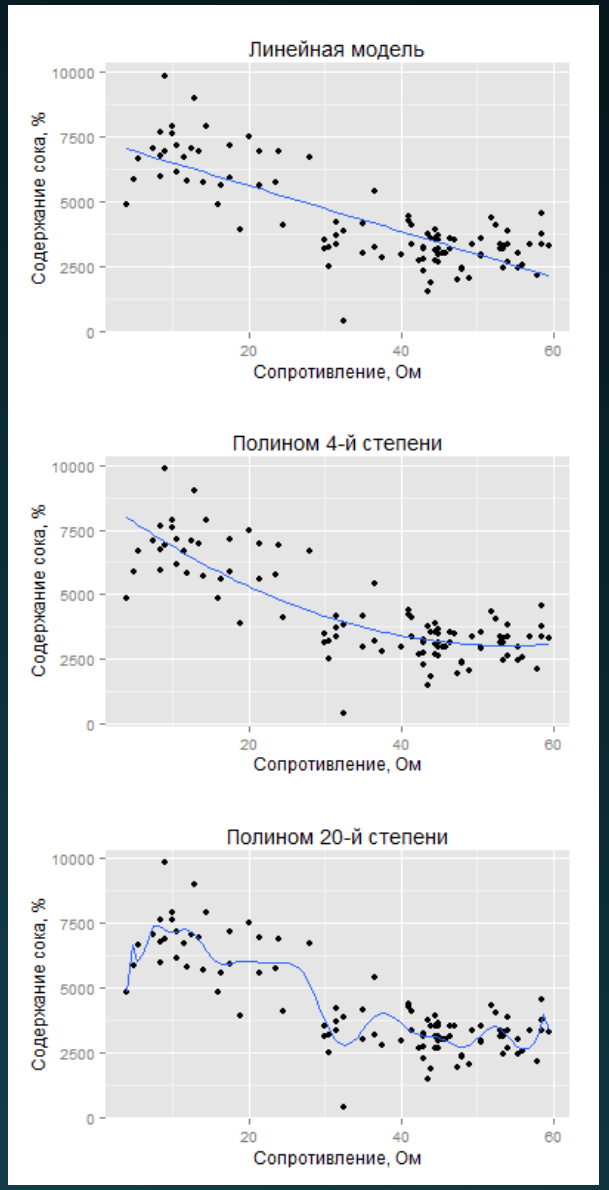
R^2 оценка качества модели. Чем выше R^2 – тем лучше модель



Как избежать переобучения?

Как избежать переобучения?

- Что делать если признаков слишком много?
- Или мы подбирали параметры модели и в какой-то момент случилось переобучение?



Пример переобучения

Строим полиномиальную модель. Оптимизируем степень полинома n.

$n = 2: x \to x, x^2$

Пример переобучения

Строим полиномиальную модель. Оптимизируем степень полинома n.

$n = 20: x \to x, x^2, \dots x^{20}$

Пример переобучения

Строим полиномиальную модель. Оптимизируем степень полинома n.

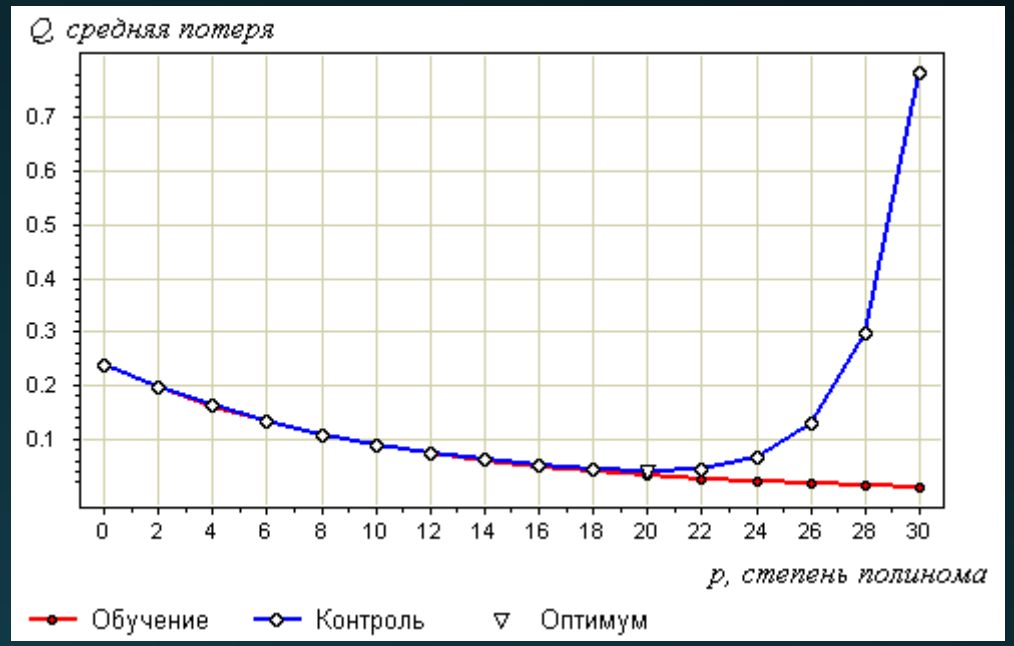
$n = 40: x \to x, x^2, \dots x^{40}$

Пример переобучения

Что же получается?

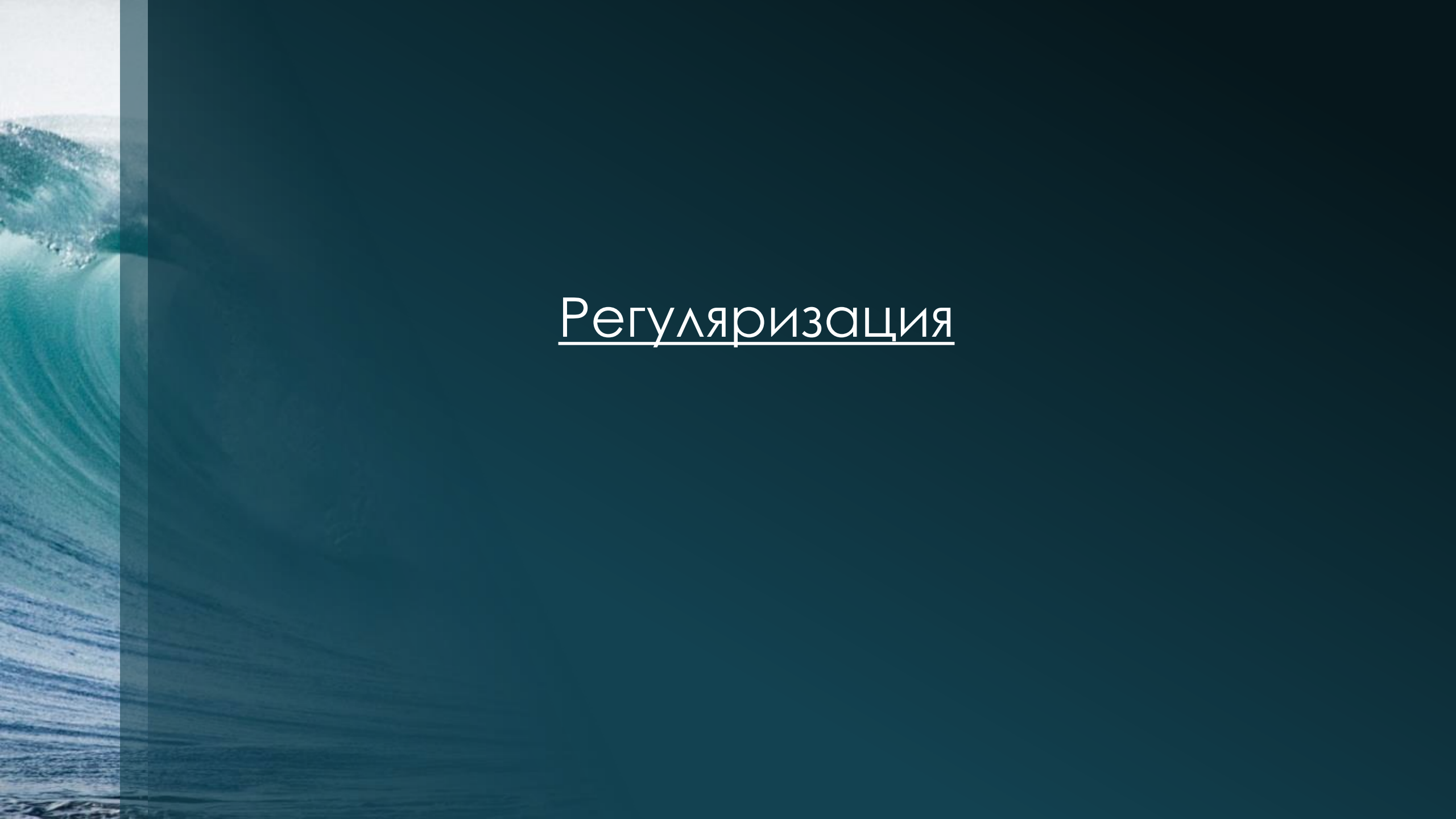
Изначально модель действительно становилась лучше при увеличении степени полинома, но потом в погоне за точностью на обучающей выборке мы просто подстроились под наши данные.

Тем самым серьезно ухудшилось качество на тестовой выборке.



Несоответствие между правильностью на обучающем наборе и правильностью на тестовом наборе является явным признаком переобучения и поэтому мы должны попытаться найти модель, которая позволит нам контролировать сложность

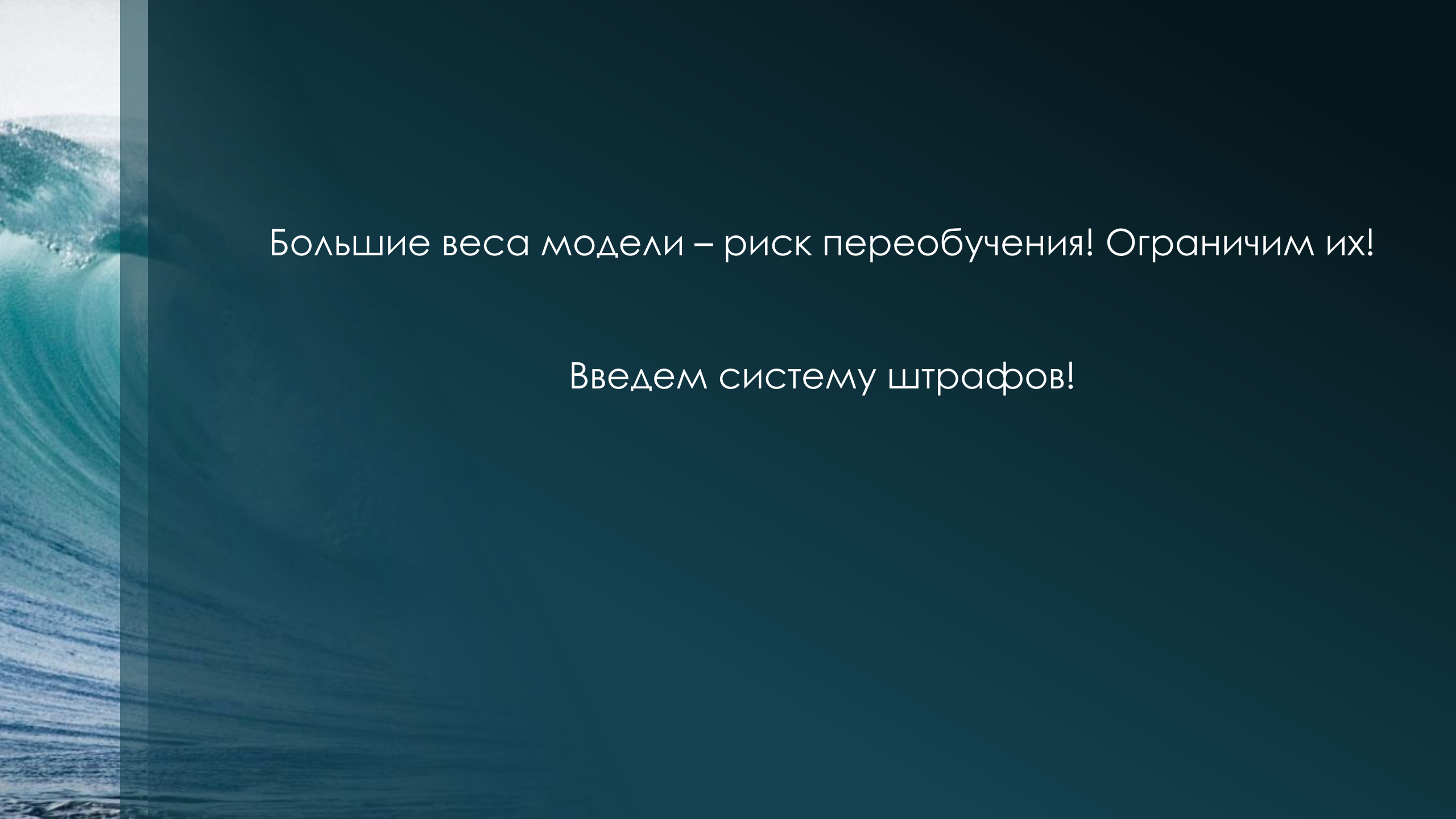
Что делать в таких ситуациях??



Регуляризация

В чем может выражаться переобучение?

- В погоне за точностью веса могут начать становится слишком большими
- Слишком большое количество признаков может сделать модель нестабильной
- Наличие скоррелированых признаков тоже может привести к переобучению



Большие веса модели – риск переобучения! Ограничим их!

Введем систему штрафов!

Гребневая регрессия

Модифицируем функцию потерь. Добавим к функционалу ошибки регуляризатор и будем минимизировать уже получившуюся величину.

Было:

$$Q(\tilde{a}) = \left|\left|\hat{\varepsilon}\right|\right|^2 = \left|\left|\hat{y}(\tilde{a}) - y\right|\right|^2 = \left|\left|ax + b - y\right|\right|^2 \to min$$

Стало:

$$Q(\tilde{a}) + \alpha\left|\left|w\right|\right|^2 \to min$$

где $\left|\left|w\right|\right|^2 = \sum w_i^2$

Это еще называется L2 регуляризацией

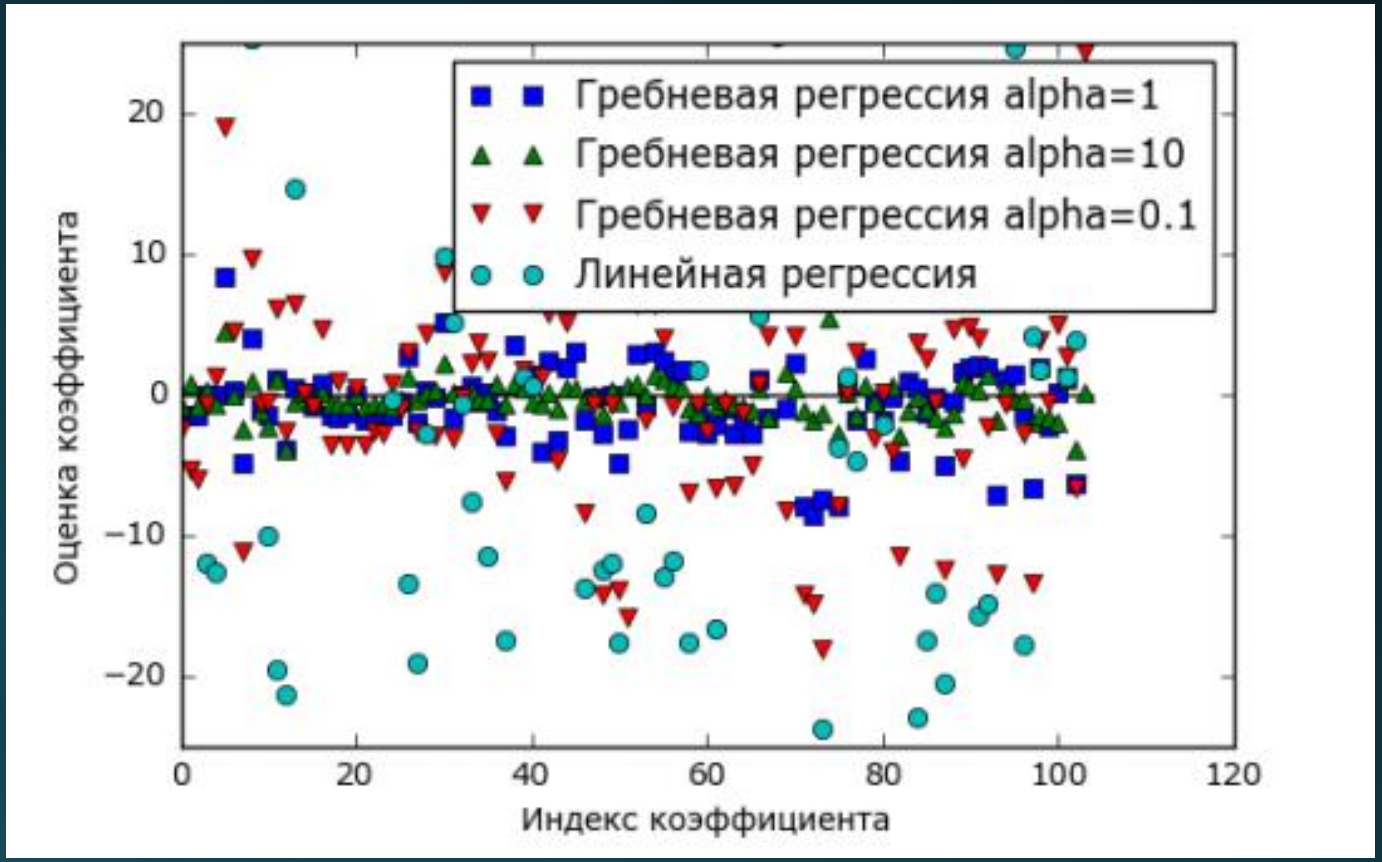
Гребневая регрессия

Посмотрим на α:

- $\alpha = 0$ – обычная регрессия, никаких ограничений на коэффициенты нет (сложная модель)
- $\alpha = 1$ – штраф за большие коэффициенты, следовательно модель имеет часть весов, близких к нулю (простая модель)
- $\alpha = 10$ – штраф еще больше, еще больше таких весов

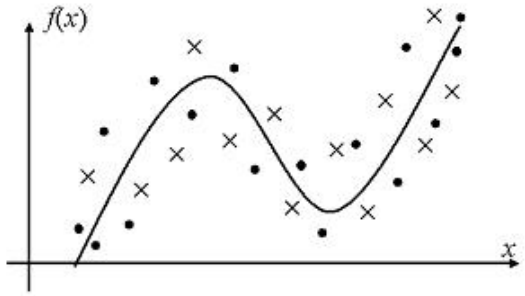
Более простая модель может давать меньшую правильность на обучающей выборке, но иметь лучшею обобщающую способность.

Гребневая регрессия



Чем больше альфа – тем проще модель и выше обобщающая способность. Но тут главное тоже не переборщить =)

Иллюстрация понятия переобучения: *а* - исходное множество экспериментальных измерений; *б* - максимально точный аппроксиматор сильно ошибается на новых измерениях



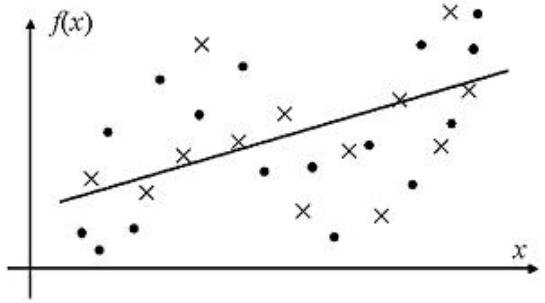


Иллюстрация метода регуляризации: *в* - регуляризованный аппроксиматор меньше ошибается на новых измерениях; *г* - слишком сильно регуляризованный аппроксиматор

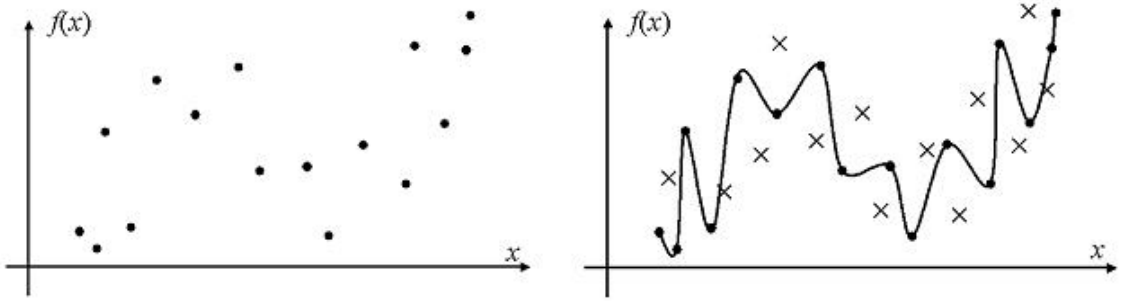


Иллюстрация понятия переобучения: *а* - исходное множество экспериментальных измерений; *б* - максимально точный аппроксиматор сильно ошибается на новых измерениях

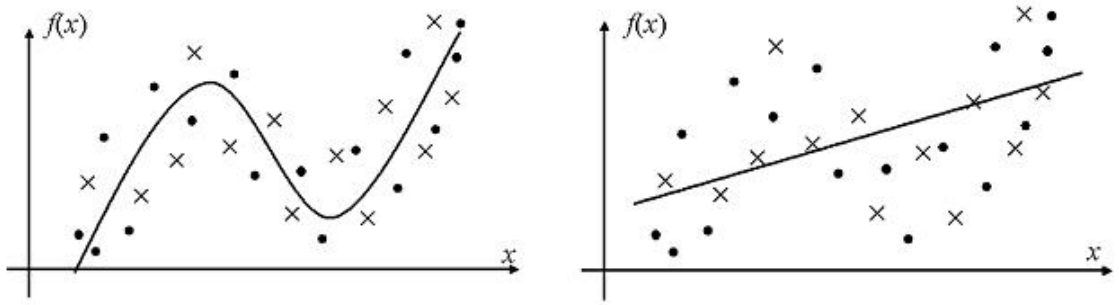


Иллюстрация метода регуляризации: *в* - регуляризованный аппроксиматор меньше ошибается на новых измерениях; *г* - слишком сильно регуляризованный аппроксиматор

Компромисс между слишком простой моделью и качеством на обучающей выборке можно определить с помощью параметра α. Нужен баланс.

Выбор α – например, по кросс валидации

Лассо регрессия регрессия

Альтернатива гребневой регрессии. Тоже сжимает веса модели, но несколько иным способом

$$Q(\tilde{a}) + \alpha \sum |w_i| \to min$$

Это еще называется L1 регуляризацией.

В чем отличие?

Лассо регрессия регрессия

Альтернатива гребневой регрессии. Тоже сжимает веса модели, но несколько иным способом

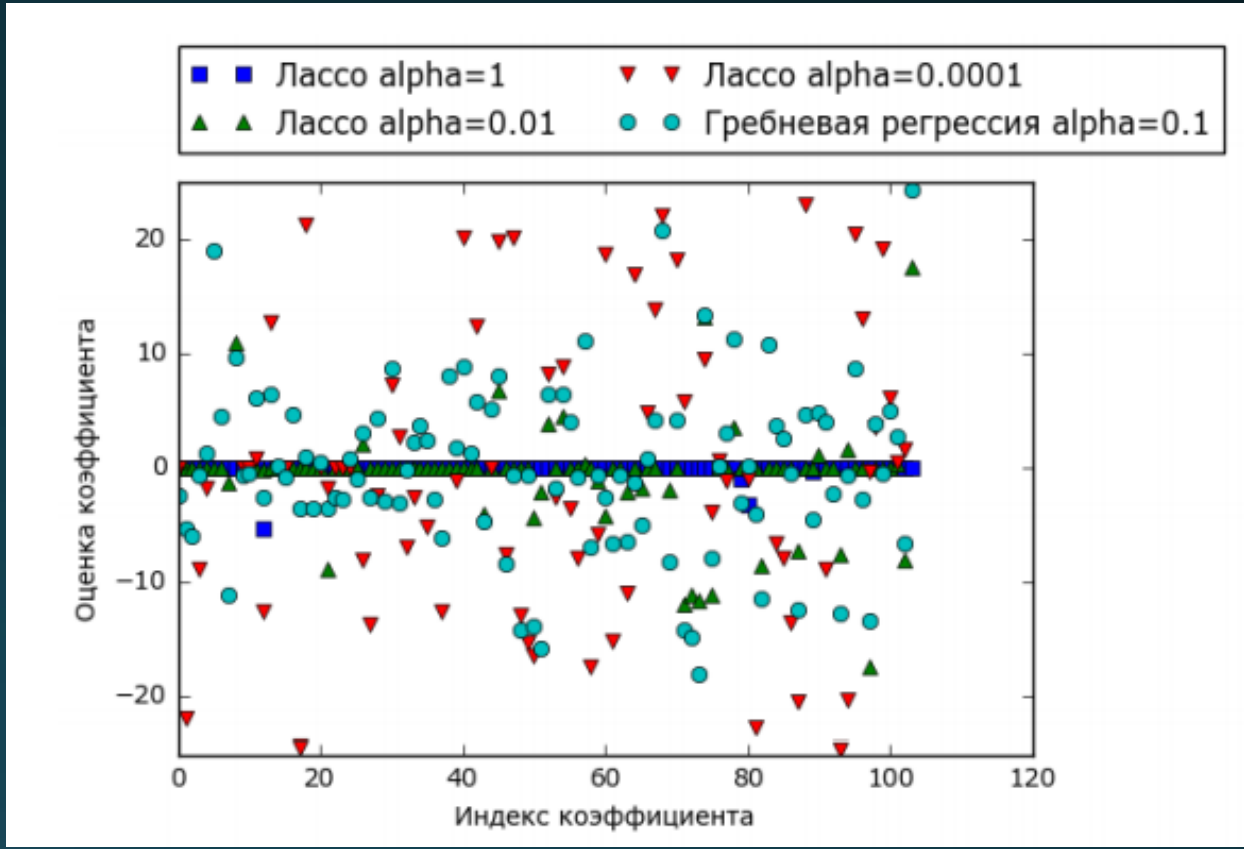
$$Q(\tilde{a}) + \alpha \sum |w_i| \to min$$

Это еще называется L1 регуляризацией.

В чем отличие? В результате такой регуляризации некоторые коэффициенты становятся равными точно нулю. Т.е. получается, что некоторые признаки полностью выкидываются из модели.

Параметр α определяет степень сжатия коэффициентов до нулевых значений

Лассо регрессия регрессия



- α = 1 – практически все веса = 0
- $\alpha = 0.0001$ - практически нерегуляризованная модель

Отбор признаков для модели

Критерии для проверки гипотез о значимости моделей и отдельных коэффициентов

Значимость всей модели в целом

Гипотеза H_0:
$a_i = 0$ для всех $i > 0$, т.е. модель в целом не значима

F – критерий:

F-статистика теста - $\frac{R^2/(m+1)}{(1-R^2)/(n-m-1)}$

Если $\frac{R^2/(m+1)}{(1-R^2)/(n-m-1)} > f_{m+1,n-m-1}(\alpha)$, то гипотеза отвергается и модель значима

Критерии для проверки гипотез о значимости моделей и отдельных коэффициентов

Значимость отдельного коэффициента (признака)

Гипотеза H_0:
$a_i = 0$ для некоторого $i > 0$, т.е. коэффициент для признака с номером i не значим (сам признак не значим)

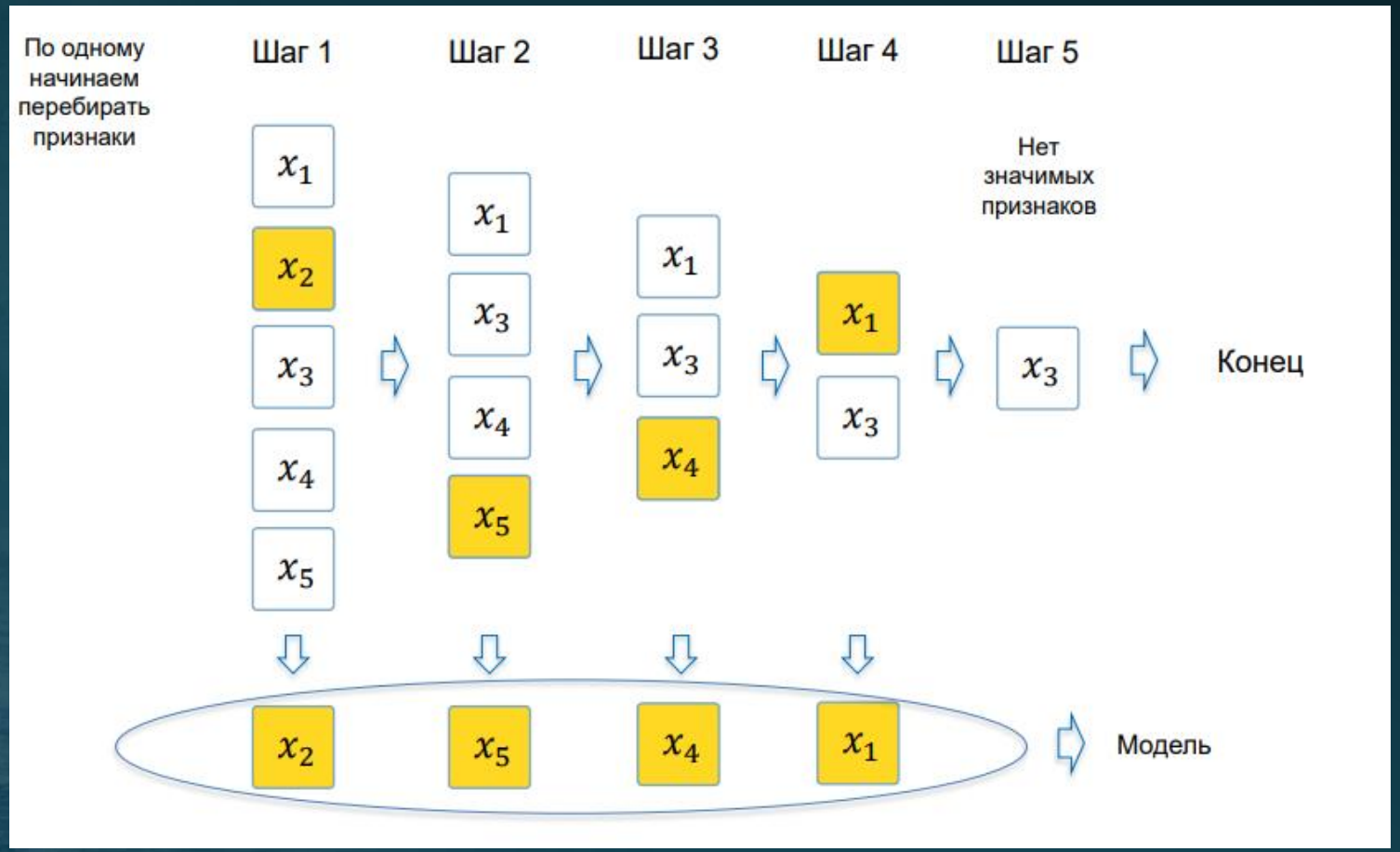
t – критерий:

t-статистика теста - $\frac{|\widehat{a_i}|}{\sqrt{s^2(X^TX)^{-1}}_{ii}}$, где $s^2 = \frac{y^T\left(I - X(X^TX)^{-1}X^T\right)y}{n-m-1}$

Если $\frac{R^2/(m+1)}{(1-R^2)/(n-m-1)} > t_{n-m-1}\left(\frac{\alpha}{2}\right)$, то гипотеза отвергается и коэффициент значим

Отбор переменных. Алгоритмы forward, backward, stepwise

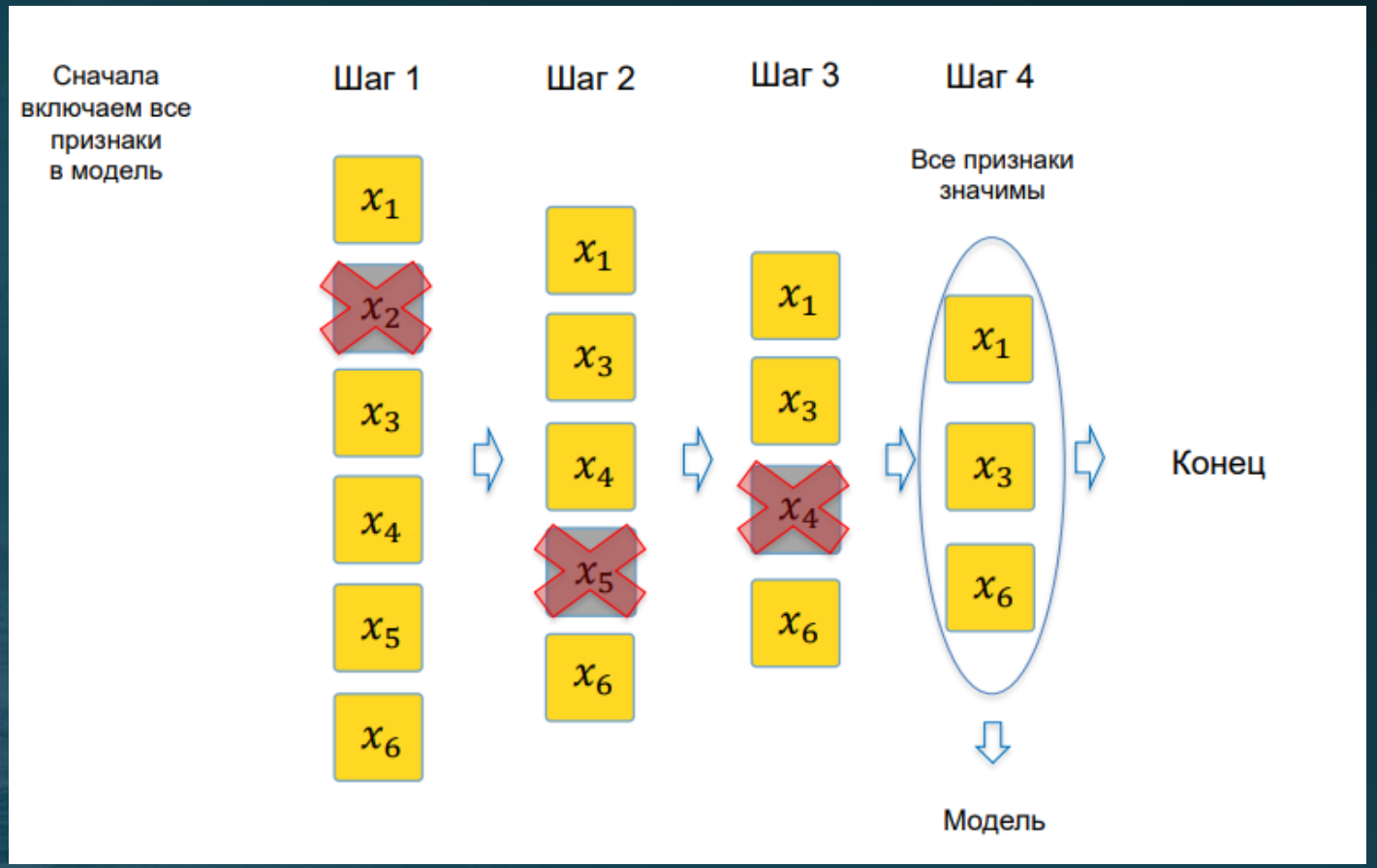
Forward Selection: Алгоритм основан на последовательном добавлении признаков в модель. На каждом шаге выбираем признак с минимальным p-value тестовой статистики значимости



Каждый раз добавляем признак, который дает наибольшее статистически значимое улучшение модели

Отбор переменных. Алгоритмы forward, backward, stepwise

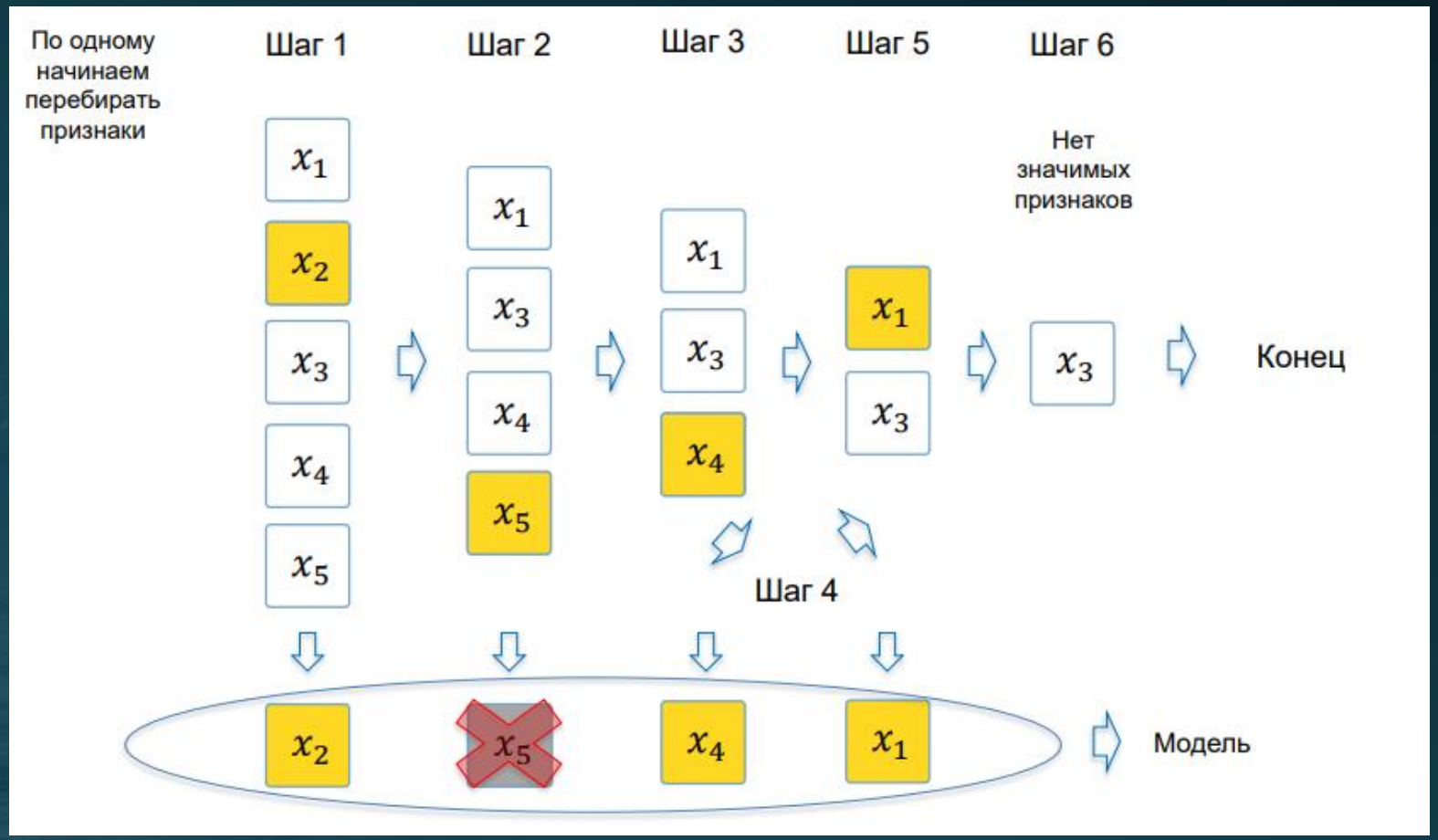
Backward Selection: Алгоритм основан на последовательном исключении признаков из модели. На каждом шаге выбираем признак с максимальным p-value тестовой статистики значимости

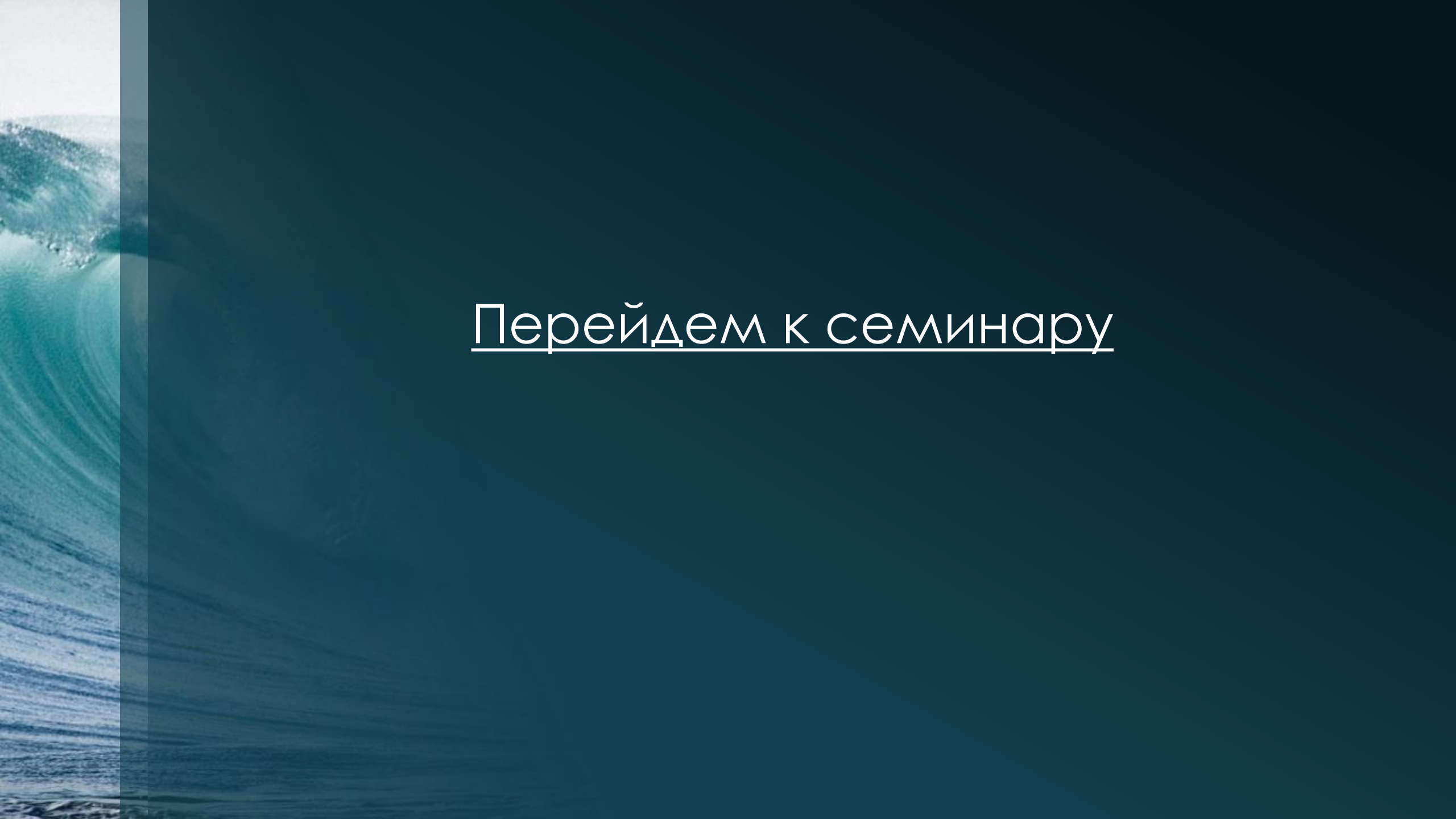


Каждый раз удаляем переменную (если это необходимо), потеря которой приводит к наиболее статистически незначимому ухудшению соответствия модели

Отбор переменных. Алгоритмы forward, backward, stepwise

Stepwise Selection: Комбинация Forward и Backward. После каждого включения проверяем, можно ли исключить какой-либо признак из уже включенных





Перейдем к семинару