Математическая статистика
Лекция 10

CSC, 12 апреля 2022

Линейная модель

Пусть данные представлены в виде набора векторов $\left(y_i, x_{i,1}, \dots, x_{i,k}\right)$, $i = 1, \dots, n$.

Линейная модель предполагает, что

$$Y_i | (X_i = x_i) = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_k x_{i,k} + \varepsilon_i,$$

где $\beta_0, \dots, \beta_k$ — неизвестные параметры, а ε_i — случайная ошибка.

Линейная модель

Немного обозначений:

— $x \in \mathbb{R}^{n \times k}$ — матрица $\left(x_{i,j}\right)_{i,j=1}^{n,k}$, x_i — i-я строка, $x_{:,j}$ — j-й столбец. Столбцы x называются независимыми переменными, факторами, предикторами, фичами, …

— $y = (y_1, \dots, y_n)$ — зависимая переменная, отклик, таргет, …

— $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ — ошибки

Пока не оговорено иное, первым фактором является вектор единиц, что позволяет записать модель в виде:

$$Y_i | (X_i = x_i) = x_i \beta + \varepsilon_i.$$

Неслучайные факторы

Рассмотрим упрощение описанной модели, которое состоит в том, что матрица x является фиксированной, а случайным является только вектор ε.

В этой модели анализ становится существенно проще, так как перестает зависеть от распределения X, но страдает применимость к реальным задачам. В частности, мы не будем рассматривать никакие асимптотические свойства, так как x фиксирована и n никуда не стремится.

Как модель выглядит формально:

— $x \in \mathbb{R}^{n \times k}$ — фиксирована и известна,

— y — известная реализация случайного вектора Y, имеющего распределение $Y = x\beta + \varepsilon$, где β — неизвестный вектор параметров, ε — случайный вектор ошибок.

Метод наименьших квадратов

Пусть у нас есть оценка β^* вектора β. С помощью нее мы можем построить предсказание $\hat{y} = x\beta^*$. Вектор $e(\beta^*) = y - \hat{y}$ называется вектором остатков. Заметим, что в случае $\beta^* = \beta$ вектор остатков равен вектору ошибок: $e(\beta) = \varepsilon$.

Будем искать β^* в виде $\arg\min\limits_{\phi} \sum_{i=1}^{n} (y_i - x_i\phi)^2 = \arg\min\limits_{\phi} \sum_{i=1}^{n} e_i(\phi)^2$. Такая оценка называется оценкой метода наименьших квадратов или МНК оценкой.

Метод наименьших квадратов

Пусть у нас есть оценка β^* вектора β. С помощью нее мы можем построить предсказание $\hat{y} = x\beta^*$. Вектор $e(\beta^*) = y - \hat{y}$ называется вектором остатков. Заметим, что в случае $\beta^* = \beta$ вектор остатков равен вектору ошибок: $e(\beta) = \varepsilon$.

Будем искать β^* в виде $\arg\min\limits_{\phi} \sum_{i=1}^{n} (y_i - x_i\phi)^2 = \arg\min\limits_{\phi} \sum_{i=1}^{n} e_i(\phi)^2$. Такая оценка называется оценкой метода наименьших квадратов или МНК оценкой.

Чтобы найти β^*, заметим, что $\sum_i (y_i - x_i\beta^*)^2 = |y - x\beta^*|^2$, то есть β^* минимизирует расстояние от y до линейной оболочки факторов $\langle x_{:,1}, \dots, x_{:,k} \rangle$. Значит $x\beta^*$ это проекция, поэтому $x^{\mathrm{T}}(y - x\beta^*) = 0 \Longrightarrow x^{\mathrm{T}}x\beta^* = x^{\mathrm{T}}y \Longrightarrow \beta^* = \left(x^{\mathrm{T}}x\right)^{-1}x^{\mathrm{T}}y$, если x имеет ранг k. Матрица $\hat{h} \coloneqq x\left(x^{\mathrm{T}}x\right)^{-1}x^{\mathrm{T}}$ называется hat matrix: $\hat{y} = x\beta^* = \hat{h}y$.

Классическая модель линейной регрессии

В зависимости от дополнительных предположений, оценка β^* будет обладать теми или иными свойствами. Мы рассмотрим классическую модель:

1. $\mathbb{E}Y = x\beta \Longrightarrow \mathbb{E}\varepsilon = 0$ — линейность,
2. $\mathbb{D}\varepsilon_i = \sigma^2$ — гомоскедастичность,
3. $\operatorname{cov}(\varepsilon_i, \varepsilon_j) = \mathbb{E}\varepsilon_i\varepsilon_j$ — некоррелированность остатков,
4. $rank(x) = k$ — неколлинеарность факторов.

Пункты 2–3: $\Big\}\ \operatorname{cov}(\varepsilon) = \sigma^2 I_n$

Теорема Гаусса—Маркова

Если выполнены предположения 1—4, то

— $\mathbb{E}\beta^* = \beta$,

— $\operatorname{cov}(\beta^*) = \sigma^2 \left(x^{\mathrm{T}} x\right)^{-1}$,

— β^* является эффективной оценкой в классе линейных несмещенных оценок: $\{\phi \,|\, \phi = ay, a \in \mathbb{R}^{k\times n}, \mathbb{E}\phi = \beta\}$.

Классическая модель линейной регрессии

Теорема Гаусса—Маркова

Если выполнены предположения 1—4, то

- $\mathbb{E}\beta^* = \beta$,
- $\operatorname{cov}(\beta^*) = \sigma^2 \left(x^{\mathrm{T}} x\right)^{-1}$,
- β^* является эффективной оценкой в классе линейных несмещенных оценок: $\{\phi \big| \phi = ay, a \in \mathbb{R}^{k \times n}, \mathbb{E}\phi = \beta\}$.

Эффективность означает, что $\operatorname{cov}(\phi) - \operatorname{cov}(\beta^*)$ неотрицательно определена $\forall\phi$, что равносильно $\mathbb{D}\left(c^{\mathrm{T}}\beta^*\right) \leq \mathbb{D}\left(c^{\mathrm{T}}\phi\right)$ для любых $c \in \mathbb{R}^k$ и ϕ.

Частные случаи:

- $c = i$-й орт: $c^{\mathrm{T}}\phi = \phi_i$, то есть $\mathbb{D}(\beta_i^*) = \sigma^2 \left(x^{\mathrm{T}} x\right)^{-1}_{i,i} \leq \mathbb{D}(\phi_i)$,
- $c =$ новое наблюдение x_{n+1}: $c^{\mathrm{T}}\phi = \hat{y}_{n+1}$, то есть $\mathbb{D}\left(\hat{y}(\beta^*)\right) \leq \mathbb{D}\left(\hat{y}(\phi)\right)$.

Классическая модель линейной регрессии

Теорема Гаусса—Маркова

Если выполнены предположения 1—4, то

- $\mathbb{E}\beta^* = \beta$,
- $\operatorname{cov}(\beta^*) = \sigma^2 \left(x^{\mathrm{T}} x\right)^{-1}$,
- β^* является эффективной оценкой в классе линейных несмещенных оценок: $\{\phi \mid \phi = ay, a \in \mathbb{R}^{k \times n}, \mathbb{E}\phi = \beta\}$.

Док-во: $\beta^* = \left(x^{\mathrm{T}} x\right)^{-1} x^{\mathrm{T}} Y = \left(x^{\mathrm{T}} x\right)^{-1} x^{\mathrm{T}} (x\beta + \varepsilon) = \beta + \left(x^{\mathrm{T}} x\right)^{-1} x^{\mathrm{T}} \varepsilon$.

Тогда, во-первых, $\mathbb{E}\beta^* = \beta + \left(x^{\mathrm{T}} x\right)^{-1} x^{\mathrm{T}} \mathbb{E}\varepsilon = \beta$,

во-вторых, $\operatorname{cov}(\beta^*) = \left(x^{\mathrm{T}} x\right)^{-1} x^{\mathrm{T}} \underbrace{\operatorname{cov}(\varepsilon)}_{\sigma^2 I_n} x \left(x^{\mathrm{T}} x\right)^{-1} = \sigma^2 \left(x^{\mathrm{T}} x\right)^{-1}$.

Классическая модель линейной регрессии

Теорема Гаусса—Маркова

Если выполнены предположения 1—4, то

- $\mathbb{E}\beta^* = \beta$,
- $\operatorname{cov}(\beta^*) = \sigma^2\left(x^{\mathrm{T}}x\right)^{-1}$,
- β^* является эффективной оценкой в классе линейных несмещенных оценок: $\{\phi \mid \phi = ay, a \in \mathbb{R}^{k\times n}, \mathbb{E}\phi = \beta\}$.

Док-во: $\beta^* = \left(x^{\mathrm{T}}x\right)^{-1}x^{\mathrm{T}}Y = \left(x^{\mathrm{T}}x\right)^{-1}x^{\mathrm{T}}(x\beta + \varepsilon) = \beta + \left(x^{\mathrm{T}}x\right)^{-1}x^{\mathrm{T}}\varepsilon$.

Тогда, во-первых, $\mathbb{E}\beta^* = \beta + \left(x^{\mathrm{T}}x\right)^{-1}x^{\mathrm{T}}\mathbb{E}\varepsilon = \beta$,

во-вторых, $\operatorname{cov}(\beta^*) = \left(x^{\mathrm{T}}x\right)^{-1}x^{\mathrm{T}}\operatorname{cov}(\varepsilon)x\left(x^{\mathrm{T}}x\right)^{-1} = \sigma^2\left(x^{\mathrm{T}}x\right)^{-1}$.

$\mathbb{E}\phi = a\mathbb{E}Y = ax\beta = \beta, \forall\beta \Longrightarrow ax = I_k$. $\operatorname{cov}(aY) = a\operatorname{cov}(\varepsilon)a^{\mathrm{T}} = \sigma^2 aa^{\mathrm{T}}$.

$$aa^{\mathrm{T}} - \left(x^{\mathrm{T}}x\right)^{-1} = aa^{\mathrm{T}} - ax\left(x^{\mathrm{T}}x\right)^{-1}x^{\mathrm{T}}a^{\mathrm{T}} = a\left(I_k - x\left(x^{\mathrm{T}}x\right)^{-1}x^{\mathrm{T}}\right)a^{\mathrm{T}} \geq 0.$$

Случай нормальных ошибок

Рассмотрим $RSS = \sum_i (y_i - x_i\beta^*)^2$ — residual sum of squares.

— $\hat{\sigma}^2 = \frac{1}{n-k} RSS$ является несмещенной оценкой σ^2,

Случай нормальных ошибок

Если $Y \sim \mathcal{N}(x\beta, \sigma^2 I_n)$, или, что то же самое, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, то верен более сильный результат:

— β^* является оценкой ММП:

$$\mathcal{L}(y|\beta) = \frac{1}{(2\pi)^{n/2}\sigma^n} \prod_i e^{-\frac{(y_i - x_i\beta)^2}{\sigma^2}} \Longrightarrow \log \mathcal{L}(y|\beta) = \mathrm{const} - \frac{1}{\sigma^2} \sum_i (y_i - x_i\beta)^2$$

— $RSS \sim \sigma^2 \cdot \chi^2(n-k)$ и не зависит от β^*,

— $\forall c \in \mathbb{R}^k$ выполняется $\frac{c^{\mathrm{T}}(\beta^* - \beta)}{\hat{\sigma}\sqrt{c^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}c}} \sim T(n-k)$,

— β^* является эффективной в классе всех несмещенных оценок,

Разные интервалы для нормальных ошибок

Ниже $t \sim T(n-k)$.

- Координата β_i: $\beta_i = \beta_i^* + t \cdot \hat{\sigma}\sqrt{(x^{\mathrm{T}}x)_{i,i}^{-1}}$
- Матожидание таргета: $\bar{y}_{n+1} = \hat{y}_{n+1} + t \cdot \hat{\sigma}\sqrt{x_{n+1}(x^{\mathrm{T}}x)^{-1}x_{n+1}^{\mathrm{T}}}$
- Таргет: $y_{n+1} = \hat{y}_{n+1} + t \cdot \hat{\sigma}\sqrt{1 + x_{n+1}(x^{\mathrm{T}}x)^{-1}x_{n+1}^{\mathrm{T}}}$
- Дисперсия ошибок: $\sigma^2 = RSS/\chi^2(n-k)$

Навороченные гипотезы

С оценкой y и β и проверкой гипотез относительно β_i мы справились. Но что с остальными гипотезами?

- Предсказание y
- Оценка β_i
- Гипотезы о параметрах: $H_0\colon \beta_i = c$
- Значимость модели: $H_0\colon \beta_2 = \dots = \beta_k = 0$
- Понижение размерности: $H_0\colon \beta_{i_1} = \beta_{i_2} = \dots = \beta_{i_m} = 0, \{i_1, \dots, i_m\} \subset \{1, \dots, k\}$

Кроме того, как проверить выполнение предположений?

Навороченные гипотезы

Обе гипотезы (и еще масса других) проверяются с помощью одной и той же техники: сравнения моделей.

Постановка такая: имеется две модели

- длинная — $Y = x\beta + z\gamma + \varepsilon$, $\gamma \in \mathbb{R}^m$,
- короткая — $Y = x\beta + \varepsilon$.

Требуется выяснить, правда ли, что истинная модель короткая?

Иначе говоря, $H_0\colon \gamma = 0$.

Идея: посчитаем RSS у длинной модели (RSS_L) и короткой (RSS_S). Ясно, что

$$RSS_S \geq RSS_L,$$

но если H_0 верна, то разница должна быть относительно невелика.

Сравнение моделей

Длинная — $x\beta + z\gamma + \varepsilon$, короткая — $x\beta + \varepsilon$; $\beta \in \mathbb{R}^k$, $\gamma \in \mathbb{R}^m$.

F-критерий Фишера

Пусть ошибки нормальны. Тогда

$$\frac{(RSS_S - RSS_L)/m}{RSS_L/(n-k-m)} \sim F_{m,n-k-m}.$$

Критическая область правая.

Значимость модели

$$H_0: \beta_2 = \cdots = \beta_k = 0$$

Длинная модель — исходная модель $y = x\beta + \varepsilon$, $RSS_L = RSS$.

Короткая модель — $y = \beta_1 + \varepsilon$, $RSS_S = \sum_i (y_i - \bar{y})^2 = TSS$ — total sum of squares.

Статистика F-критерия: $\frac{(TSS-RSS)/(k-1)}{RSS/(n-k)} \sim F_{k-1,n-k}$.

Значимость модели

$$H_0\colon \beta_2 = \cdots = \beta_k = 0$$

Длинная модель — исходная модель $y = x\beta + \varepsilon$, $RSS_L = RSS$.

Короткая модель — $y = \beta_1 + \varepsilon$, $RSS_S = \sum_i (y_i - \bar{y})^2 = TSS$ — total sum of squares.

Статистика F-критерия: $\frac{(TSS-RSS)/(k-1)}{RSS/(n-k)} \sim F_{k-1,n-k}$.

Для величины $ESS \coloneqq TSS - RSS$ есть свое название — explained sum of squares.

Величина $R^2 = \frac{ESS}{TSS}$ называется коэффициентом детерминации. Она всегда находится в промежутке $[0,1]$ и характеризует качество модели: чем ближе к 1, тем лучше. Статистику критерия Фишера можно выразить через нее:

$$\frac{(TSS - RSS)/(k-1)}{RSS/(n-k)} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}.$$

Понижение размерности

$$H_0\colon \beta_{i_1} = \dots = \beta_{i_m} = 0$$

Длинная модель — исходная модель $y = x\beta + \varepsilon$, $RSS_L = RSS$.

Короткая модель — $y = x'\beta' + \varepsilon$, где x' это x без факторов $\{i_1, \dots, i_m\}$.

Статистика F-критерия: $\frac{(RSS_S - RSS)/m}{RSS/(n-k)} \sim F_{m,n-k}$.

В частном случае $H_0\colon \beta_i = 0$ получаем статистику $\frac{RSS_S - RSS}{RSS/(n-k)} \sim F_{1,n-k}$. Эта

статистика — квадрат статистики $\frac{\beta_i^* - \beta_i}{\hat{\sigma}\sqrt{(x^{\mathrm{T}}x)^{-1}_{i,i}}} \sim T(n-k)$.

Беды с регрессией

Какие бывают беды с регрессией?

Беды с предположениями:

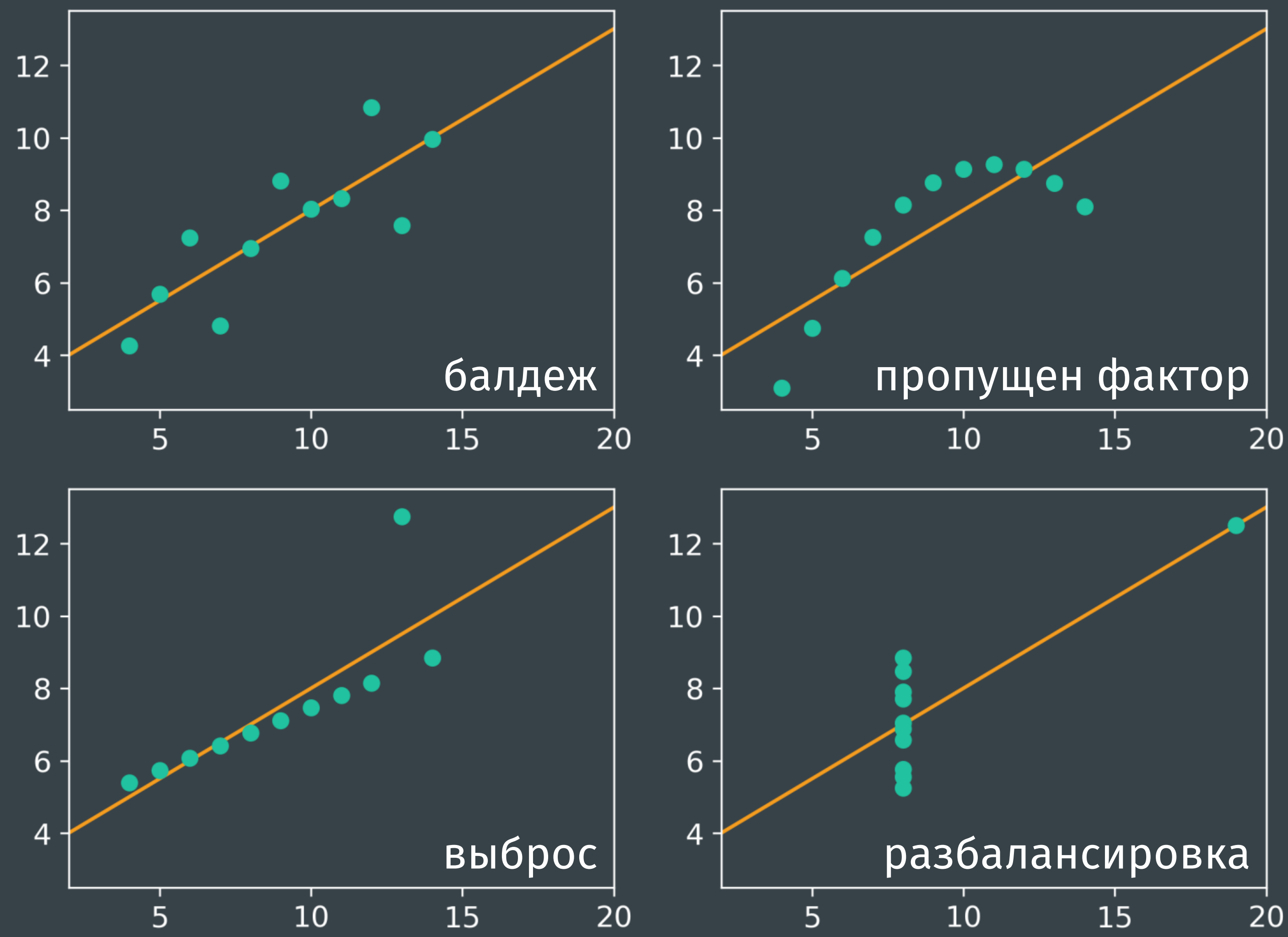
- Неверная спецификация модели (пропущен фактор) — $\mathbb{E}Y \neq x\beta$
- Гетероскедастичность — $\mathbb{D}\varepsilon_i \neq \sigma^2$
- Корреляция ошибок — $\operatorname{cov}(\varepsilon_i, \varepsilon_j) \neq 0$

Беды с данными:

- Выбросы
- Разбалансировка
- Мультиколлинеарность

Графики vs циферки: квартет Энскомба

$\bar{x}$	9
$\tilde{s}_x^2$	11
$\bar{y}$	7.5
$\tilde{s}_y^2$	4.125
ρ^*	0.816
β^*	(3, 0.5)
R^2	0.67



Не потеряли ли значимый фактор?

RESET-тест Рамсея

Короткая модель — исходная модель.

Длинная модель — $y = x\beta + \hat{y}^2\gamma_1 + \cdots + \hat{y}^{m+1}\gamma_m + \varepsilon$, где $\hat{y}^l$ это покоординатная степень $\hat{y} = x\beta^*$.

Статистика F-критерия: $\frac{(RSS - RSS_L)/m}{RSS_L/(n-k-m)} \sim F_{m,n-k-m}$.

Идея: $\langle \hat{y}^2, \ldots, \hat{y}^{m+1} \rangle$ зависит от степеней и всевозможных произведений столбцов x вплоть до степени $m+1$, поэтому если мы и правда что-то потеряли, зависящее от x, то оно будет скоррелировано с такими факторами. Нам не нужно сильно улучшить модель, нам нужно значимо ее улучшить.

Равная ли дисперсия у остатков?

Отсутствие гомоскедастичности называется гетероскедастичностью — это ситуация, когда $\mathbb{D}\varepsilon_i = \mathbb{E}\varepsilon_i^2$ зависит от x_i.

Идея: оценим $\mathbb{E}\varepsilon_i^2$ с помощью $e_i^2(\beta^*)$ — plug-in оценка по выборке объема 1.

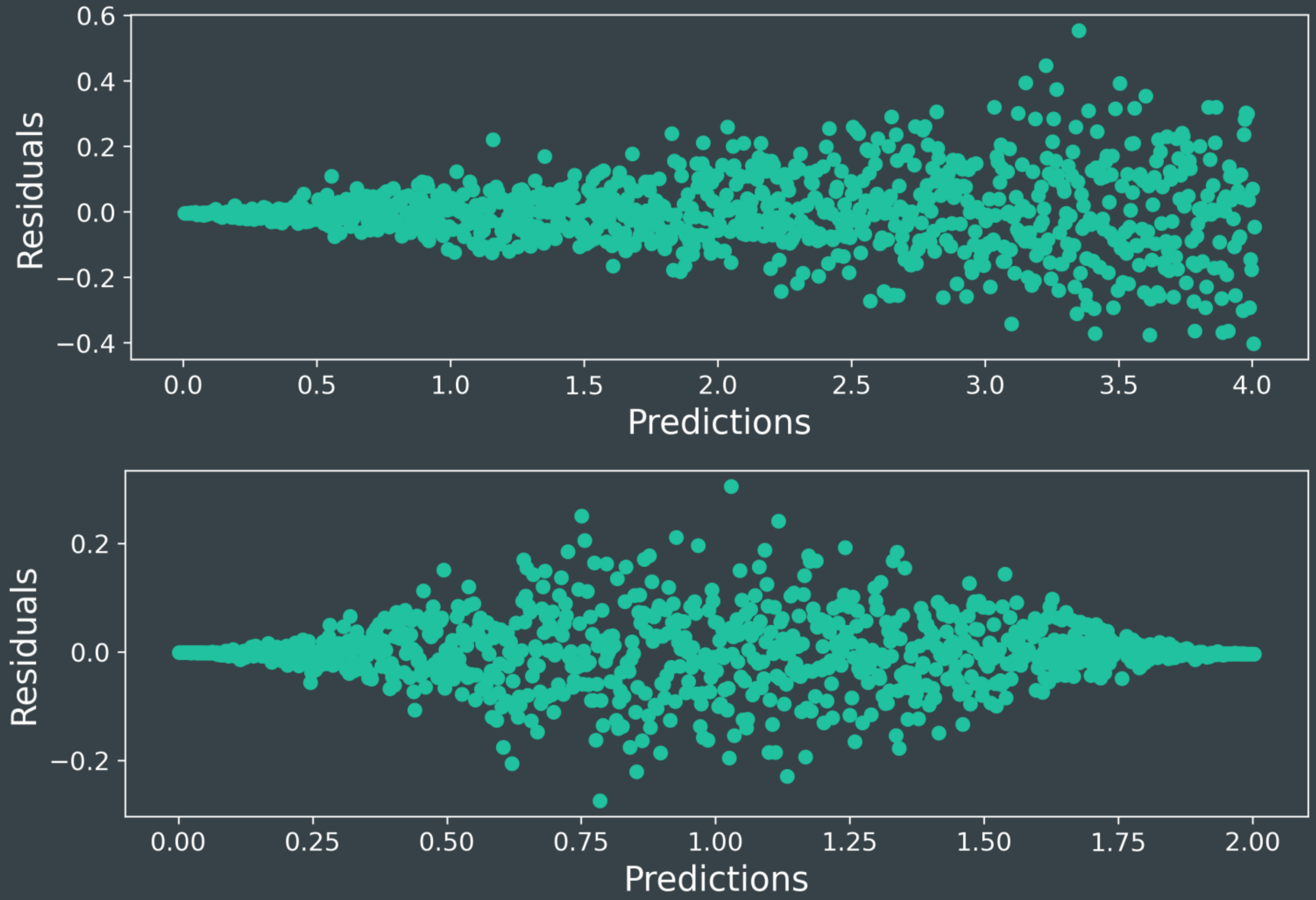
Тест Уайта

Короткая модель — $e^2(\beta^*) = \beta_1 + \varepsilon$.

Длинная модель — $e^2(\beta^*) = x\beta + x^2\gamma + \varepsilon$, где x^2 это матрица из квадратов и попарных произведений факторов (покоординатных).

Гетероскедастичность

Обычно гетероскедастичность выглядит как-то так. Вместо $\hat{y}$ по оси x можно брать любые функции от факторов.



Мультиколлинеарность

Бывает строгая: $rank(x) < k$, и нестрогая: $rank(x) = k$, $cond\left(x^{\mathrm{T}}x\right) \gg 1$.

Строгая является нарушением предположений и приводит к тому, что вектор β^* неединственен. Возникает из-за невнимательности и лечится удалением плохих факторов, пока ранг не нормализуется.

Нестрогая возникает по разным причинам и приводит к тому, что дисперсия β_i^* повышается, а значимость уменьшается. Это следует из того, что $\mathbb{D}(\beta_i^*) = \sigma^2\left(x^{\mathrm{T}}x\right)_{i,i}^{-1} = \sigma^2/RSS_i$, где RSS_i это RSS для следующей модели:

$$x_{:,i} = x_{:,-i}\beta + \varepsilon.$$

Мультиколлинеарность можно заподозрить, если есть куча факторов, коэффициенты которых по отдельности не значимы, но гипотеза об одновременном равенстве нулю отвергается.

Bias-variance decomposition

Пусть мы построили нашу модель по некоторым данным $D_{[n]} = (X, Y)_{[n]}$. Как она будет работать на новом наблюдении $D = (X, Y)$?

Посчитаем MSE:

$$MSE = \mathbb{E}_{D, D_{[n]}} (Y - X\beta^*)^2 = \mathbb{E}_X \mathbb{D}_{D_{[n]}} (X\beta^*) + \mathbb{E}_X \left(X\beta - \mathbb{E}_{D_{[n]}} (X\beta^*) \right)^2 + \mathbb{D}\varepsilon.$$

Слагаемые справа имеют специальные названия:

- $\mathbb{E}_X \mathbb{D}_{D_{[n]}} (X\beta^*)$ — дисперсия, variance, переобучение, overfitting,
- $\mathbb{E}_X \left(X\beta - \mathbb{E}_{D_{[n]}} (X\beta^*) \right)^2$ — смещение, bias, недообучение, underfitting,
- $\mathbb{D}\varepsilon$ — irreducible error, неустранимая ошибка.

Bias-variance decomposition

$$MSE = \mathbb{E}_X \mathbb{D}_{D_{[n]}}(X\beta^*) + \mathbb{E}_X \left(X\beta - \mathbb{E}_{D_{[n]}}(X\beta^*)\right)^2 + \mathbb{D}\varepsilon.$$

Если выполняются предположения классической модели для всех реализаций $X_{[n]}$, то для МНК-оценки β^* выполняется:

— $\mathbb{E}_X \mathbb{D}_{D_{[n]}}(X\beta^*) = \mathbb{E}_X X \left[\mathbb{E}_{X_{[n]}} \left(X_{[n]}^{\mathrm{T}} X_{[n]}\right)^{-1}\right] X^{\mathrm{T}},$

— $\mathbb{E}_{X_{[n]}} \beta^* = \beta \Longrightarrow \mathbb{E}_X \left(X\beta - \mathbb{E}_{D_{[n]}}(X\beta^*)\right)^2 = 0,$

— $\mathbb{D}\varepsilon = \sigma^2.$

Таким образом смещение равно нулю, но дисперсия (которая зависит только от распределения $X_{[n]}$, которое мы всю дорогу игнорировали) при этом может быть очень большой.

Регуляризация

Что делать, если все же хочется уменьшить дисперсию?

Стандартный инструмент — регуляризация. С ее помощью можно уменьшить дисперсию ценой увеличения смещения.

На регуляризацию можно смотреть по разному, самое простое — как на введение штрафной функции.

Регуляризация

Раньше мы минимизировали функцию качества $\sum_i (y_i - x_i\beta)^2$.

Давайте минимизировать другую функцию:

- $L_2(\beta) = \sum_i (y_i - x_i\beta)^2 + \lambda \sum_{i>1} \beta_i^2$ — ridge regression,
- $L_1(\beta) = \sum_i (y_i - x_i\beta)^2 + \lambda \sum_{i>1} |\beta_i|$ — lasso regression,
- вариаций масса

Здесь $\lambda \in (0, \infty)$ — параметр, определяющий степень регуляризации и обеспечивающий компромисс между смещением и дисперсией. Важно, что свободный коэффициент не ругается!

Ridge & Lasso

Ридж-регрессия, она же L_2, она же гребневая, она же Тихонова, допускает аналитическое решение:

$$\beta^* = \left(x^{\mathrm{T}}x + \lambda I^*\right)^{-1} x^{\mathrm{T}} y,$$

где $I^* = \mathrm{diag}\{0, 1, \dots, 1\}$. Все работает даже если $x^{\mathrm{T}}x$ вырождена.

С Lasso такое не прокатит, придется оптимизировать численно.

Ridge & Lasso

Альтернативный взгляд:

– ридж регрессия эквивалентна оптимизационной задаче:

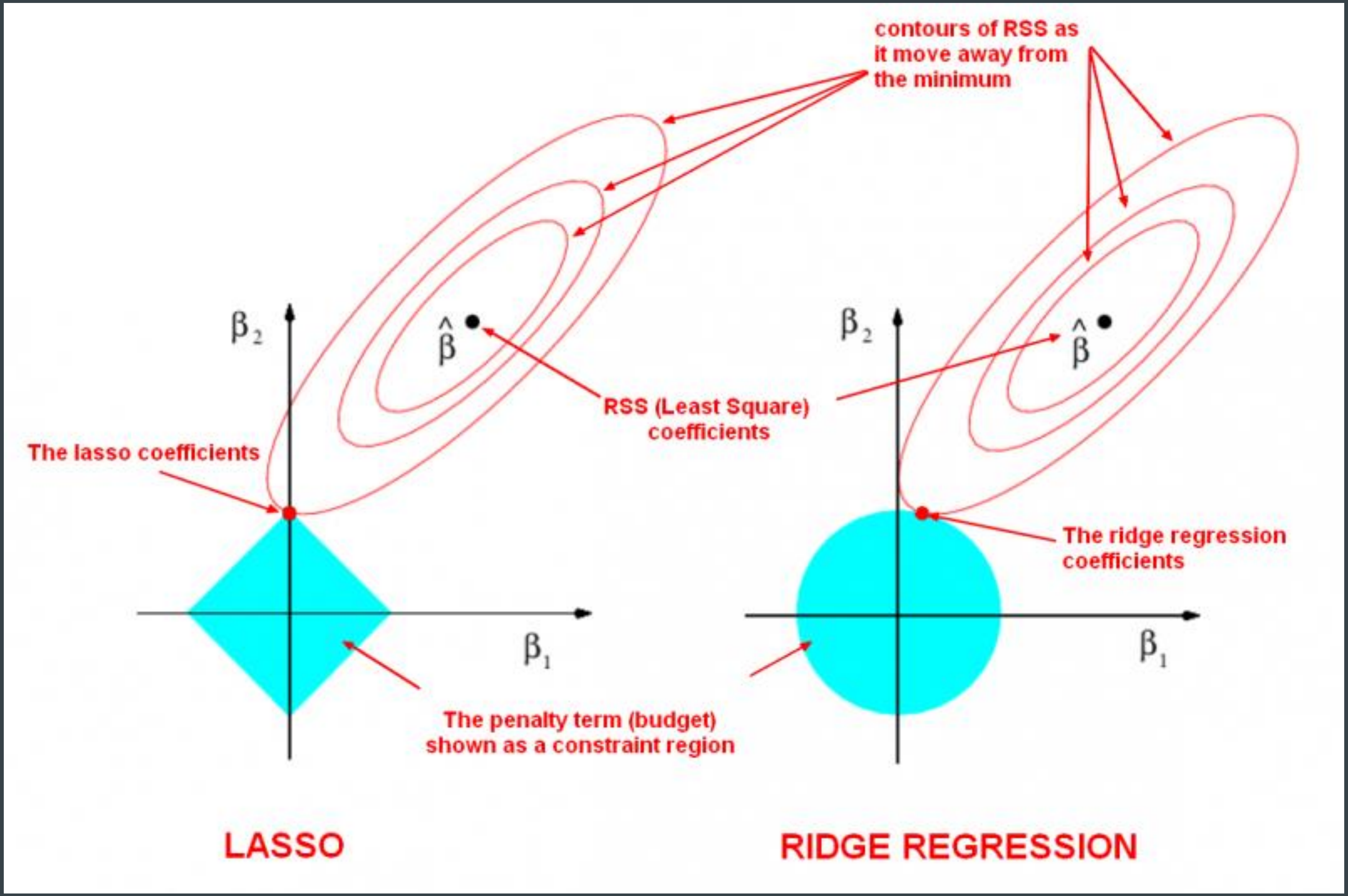
  $\sum_i (y_i - x_i\beta)^2 \to \min$, при условии $\sum_{i>1} \beta_i^2 \leq s$.

– лассо эквивалентна оптимизационной задаче:

  $\sum_i (y_i - x_i\beta)^2 \to \min$, при условии $\sum_{i>1} |\beta_i| \leq s$.

Увеличение λ соответствует уменьшению s.

Ridge & Lasso



contours of RSS as
it move away from
the minimum
β_2
$\hat{\beta}$
RSS (Least Square)
coefficients
The lasso coefficients
β_1
The ridge regression
coefficients
The penalty term (budget)
shown as a constraint region
LASSO
RIDGE REGRESSION

Вопросы?