ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ
УПРАВЛЕНИЯ

Основан в 1919 году

Методы и технологии машинного обучения

Лекция 3: Линейные регрессионные модели

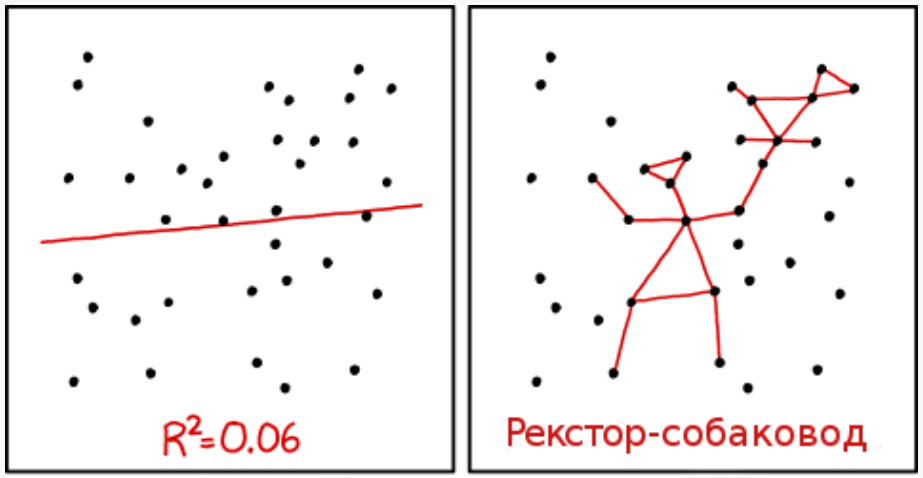
Светлана Андреевна Суязова (Аксюк)
sa_aksyuk@guu.ru

осенний семестр 2021 / 2022 учебного года

План лекции

- Непрерывный Y: линейная регрессия

- Качественные регрессоры, взаимодействие регрессоров
- Выбор оптимальной модели



Я не доверяю линейной регрессии, если угадать направление связи по графику разброса труднее, чем нарисовать новое созвездие

xkcd.com/1725/

Линейная регрессия линейна по параметрам

$$Y = f(X) + \epsilon$$

$$f(X) = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j,$$

где $\hat{\beta}_0$, $\hat{\beta}_j$ – оценки параметров; X_j – регрессоры:

- непрерывные (количественные) переменные;
- базисные функции ($\log X$, $\sqrt{X}$, X^2);
- полиномиальные представления: $X_2 = X_1^2, X_3 = X_1^3$;
- фиктивные переменные (*dummy*);
- взаимодействия между переменными: $X_3 = X_1 \cdot X_2$.

$$RSS(\beta) = \sum_{i=1}^{n} \Big(y_i - f(x_i) \Big)^2 = \sum_{i=1}^{n} \Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \Big)^2 \to \min$$

МНК-оценки: $\hat{\beta} = (X^T X)^{-1} X^T \mathrm{y}$

Допущения:

- остатки случайны и соответствуют условиям Гаусса-Маркова;
- переменные $\mathrm{x}_1, \ldots, \mathrm{x}_j, \ldots, \mathrm{x}_p$ некоррелированы.

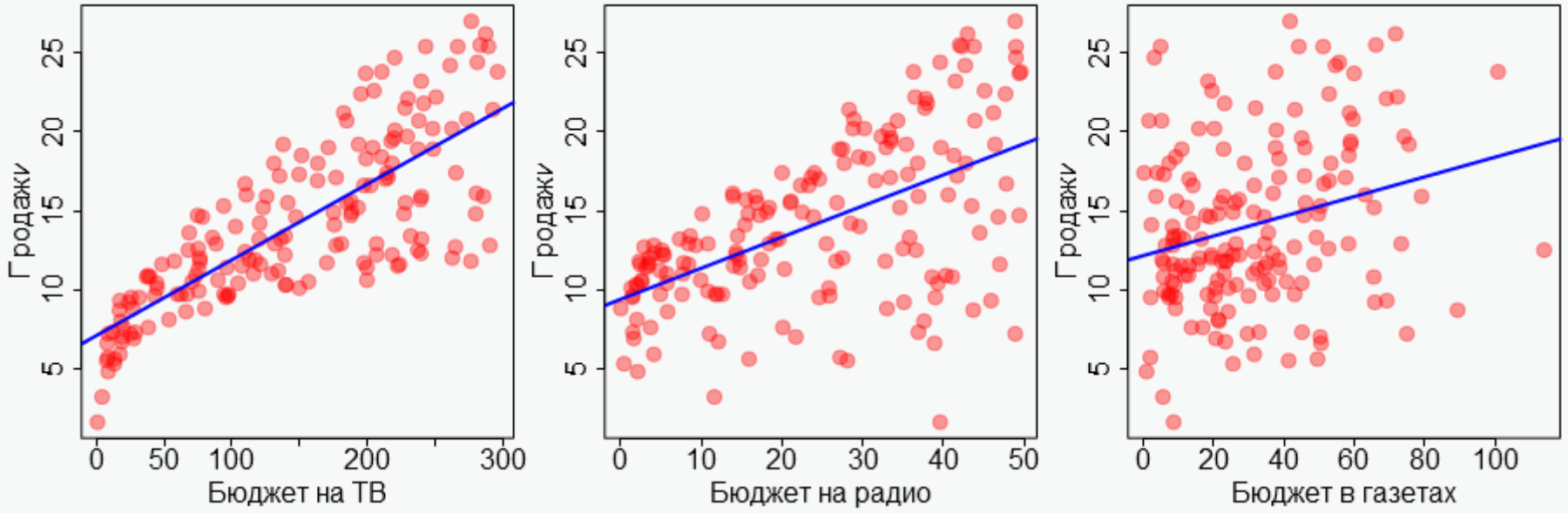
Теорема Гаусса-Маркова: МНК-оценки обладают наименьшей дисперсией в классе линейных несмещённых оценок:

$$Var(\hat{\beta}) = (X^T X)^{-1} \hat{\sigma}^2, \text{ где } \hat{\sigma}^2 = \hat{Var}(\epsilon)$$

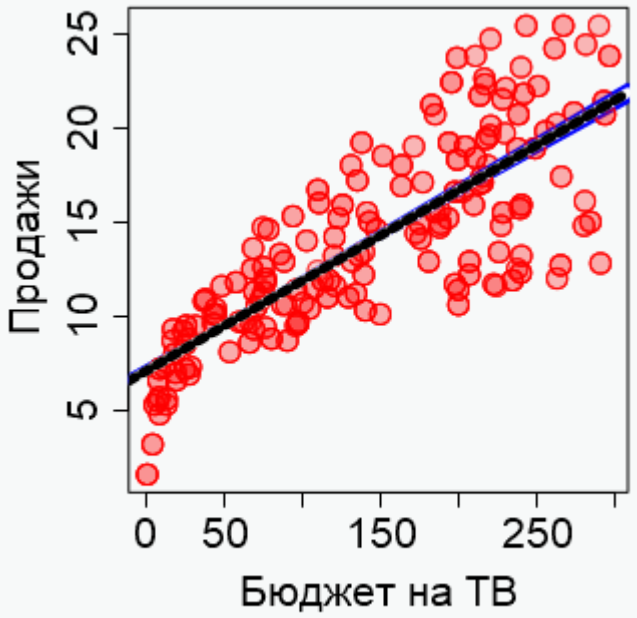
Однако, если пожертвовать несмещённостью, можно уменьшить дисперсию оценок параметров (LASSO, ридж-регрессия)

Пример 1 (маркетинговый план): Advertising

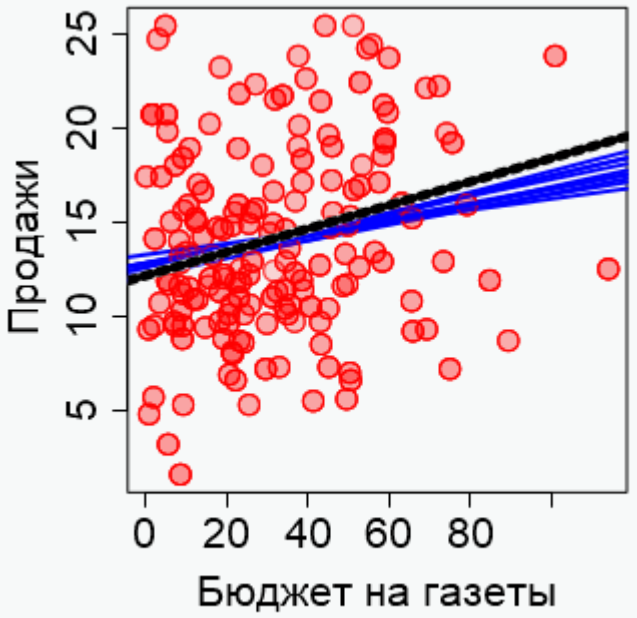
- $n = 200, p = 3$;
- обучающая выборка: 85%;
- `Sales` – объём продаж продукта, тыс. единиц;
- `TV` – размер рекламного бюджета на ТВ, тыс.долл.;
- `Radio` – рекламный бюджет на радио;
- `Newspaper` – рекламный бюджет в газетах.



Дисперсия оценок и устойчивость модели



10 выборок из обучающей по 80% наблюдений



10 выборок из обучающей по 80% наблюдений

Модель	Оценка.коэфф.b_1	Ошибка.коэфф.b_1
для ТВ (n=170)	0.048	0.0030
для газет (n=170)	0.062	0.0178

Два подхода к отбору объясняющих переменных

1. Эконометрический: на основе проверки гипотез. Ключевые метрики – P-значения для параметров, скорректированный R-квадрат, информационные критерии качества (Акаике, Байесовский и т.д.).

2. Машинного обучения: на основе точности модели (MSE на тестовой выборке), методов сжатия и снижения размерности.

План лекции

- Непрерывный Y: линейная регрессия
- Качественные регрессоры, взаимодействие регрессоров
- Выбор оптимальной модели

Пример интерпретации модели с качественными регрессорами

Пример 2 (зарплаты, Москва, 2012): wages.ru

Цель: построить модель, чтобы обосновать влияние различных факторов на размер среднемесячной заработной платы.

Данные: Подвыборка данных по 150 жителям Москвы из репрезентативной выборки по индивидуумам 21-ой волны обследования (2012г.) «Российского мониторинга экономического положения и здоровья населения НИУ-ВШЭ (RLMS-HSE)» (http://www.hse.ru/rlms'http://www.hse.ru/rlms>).

Пример 2 (зарплаты, Москва, 2012): wages.ru

- `salary` – среднемесячная зарплата после вычета налогов за последние 12 месяцев (рублей);
- `male` – пол: **1** – мужчина, **0** – женщина;
- `educ` – уровень образования:

  - **1** – 0-6 классов,
  - **2** – незаконченное среднее (7-8 классов),
  - **3** - незаконченное среднее плюс что-то еще,
  - **4** – законченное среднее,
  - **5** – законченное среднее специальное,
  - **6** – законченное высшее образование и выше;

- `forlang` - иност. язык: **1** – владеет, **0** – нет;
- `exper` – официальный стаж с 1.01.2002 (лет).

Модель 1: $\widehat{\text{salary}} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{male}$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	30213	4050.45	7.46	0.0000
male	18076	5705.78	3.17	0.0019

$$\text{male} = 0\text{: } \widehat{\text{salary}} = 30213 + 18076 \cdot 0 = 30213$$

$$\text{male} = 1\text{: } \widehat{\text{salary}} = 30213 + 18076 \cdot 1 = 48289$$

Модель 2: $\hat{\text{salary}} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{exper} + \hat{\beta}_2 \cdot \text{male}$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	15123	10064.43	1.50	0.1355
exper	1728	1056.70	1.64	0.1044
male	17637	5674.28	3.11	0.0023

$\text{male} = 0$: $\hat{\text{salary}} = 15123 + 1728 \cdot \text{exper}$

$\text{male} = 1$:
$\hat{\text{salary}} = 15123 + 1728 \cdot \text{exper} + 17637 = 32760 + 1728 \cdot \text{exper}$

Модель 3:

$$\hat{\text{salary}} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{exper} + \hat{\beta}_2 \cdot \text{male} + \hat{\beta}_3 \cdot \text{male} \cdot \text{exper}$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	16857	13368.19	1.26	0.2097
exper	1530	1459.69	1.05	0.2967
male	13905	19679.44	0.71	0.4812
exper:male	421	2125.01	0.20	0.8433

Коэффициенты модели **незначимы**; эффект взаимодействия `exper:male` наименее значим.

Модель	R.квадрат	R.квадрат.скорр	F.расч	MSE.тест
1	0.074	0.067	10.037	1.65e+08
2	0.094	0.079	6.423	4.96e+07
3	0.094	0.072	4.262	5.56e+07

План лекции

- Непрерывный Y: линейная регрессия
- Качественные регрессоры, взаимодействие регрессоров

- Выбор оптимальной модели

Первый подход: измерители точности с поправкой

C_p – оценка среднеквадратичной ошибки на контрольной выборке:

$$C_p = \frac{1}{n}\left(RSS + 2d\hat{\sigma}^2\right)$$

где $\hat{\sigma}^2$ – оценка дисперсии остатков ϵ для всех уникальных значений отклика регрессионной модели, d – количество предикторов, RSS – остаточная сумма квадратов регрессионной модели.

AIC – информационный критерий Акаике:

$$AIC = \frac{1}{n\hat{\sigma}^2}\left(RSS + 2d\hat{\sigma}^2\right)$$

В формуле опущена константа.

Первый подход: измерители точности с поправкой

BIC – байсовский информационный критерий:

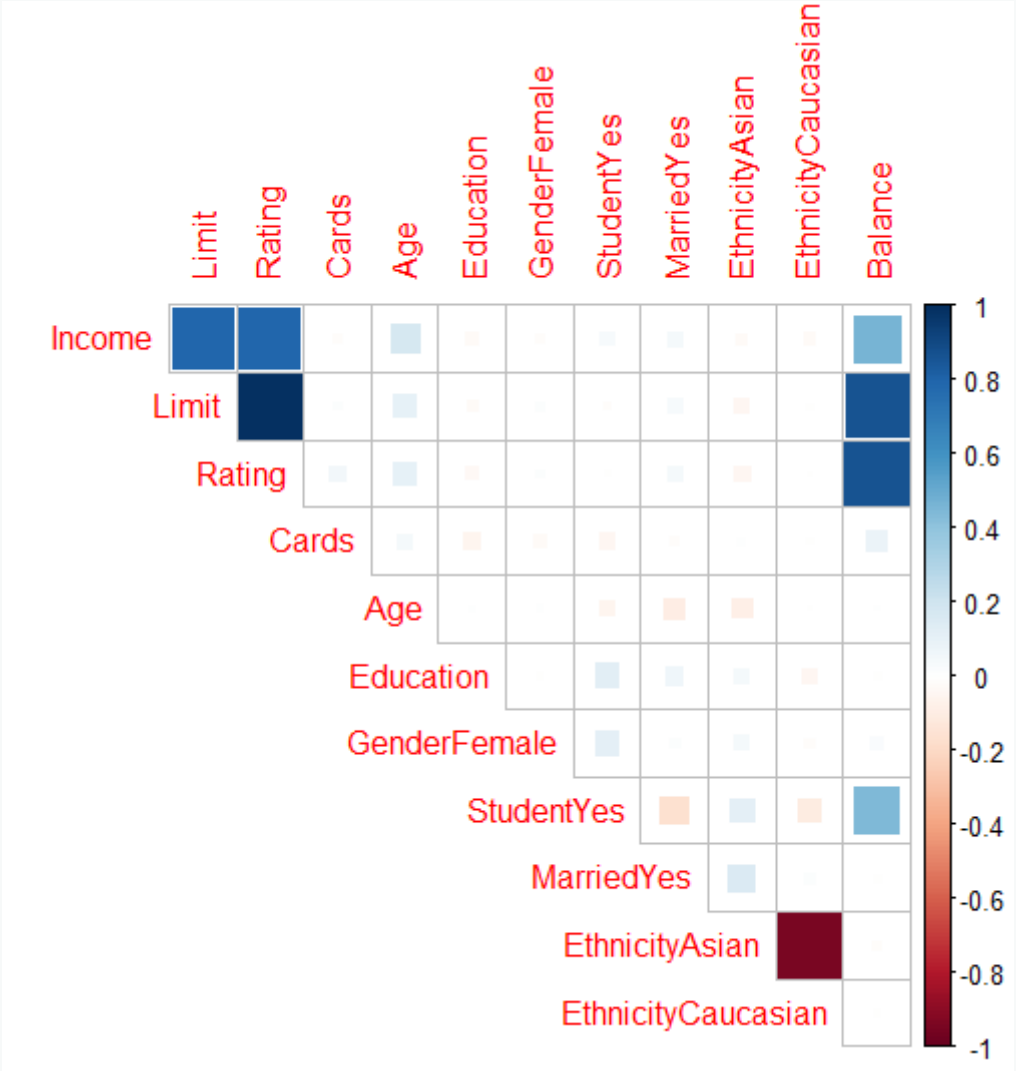
$$BIC = \frac{1}{n}\left(RSS + \log(n)d\hat{\sigma}^2\right)$$

В формуле опущена константа.

R^2_{adj} – скорректированный коэффициент детерминации:

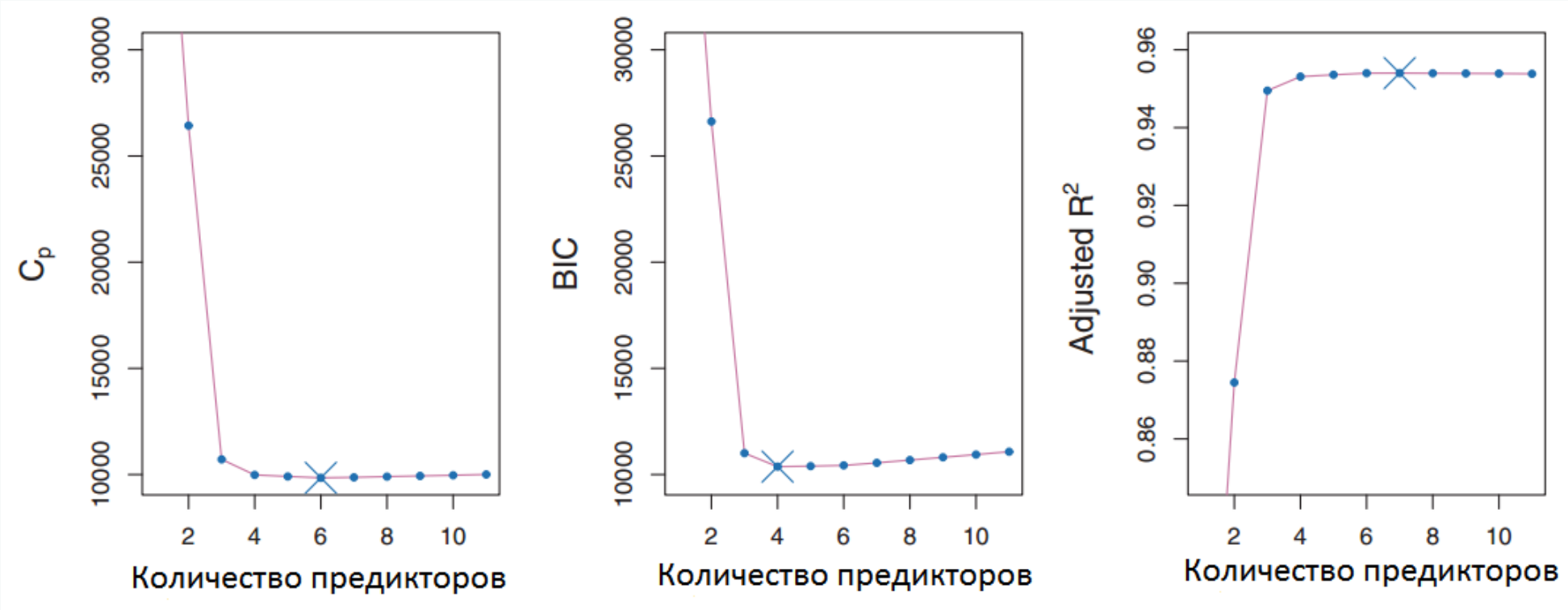
$$R^2_{adj} = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

При увеличении количества предикторов R^2 всегда растёт, а R^2_{adj} может как расти, так и снижаться.



Данные `Credit`

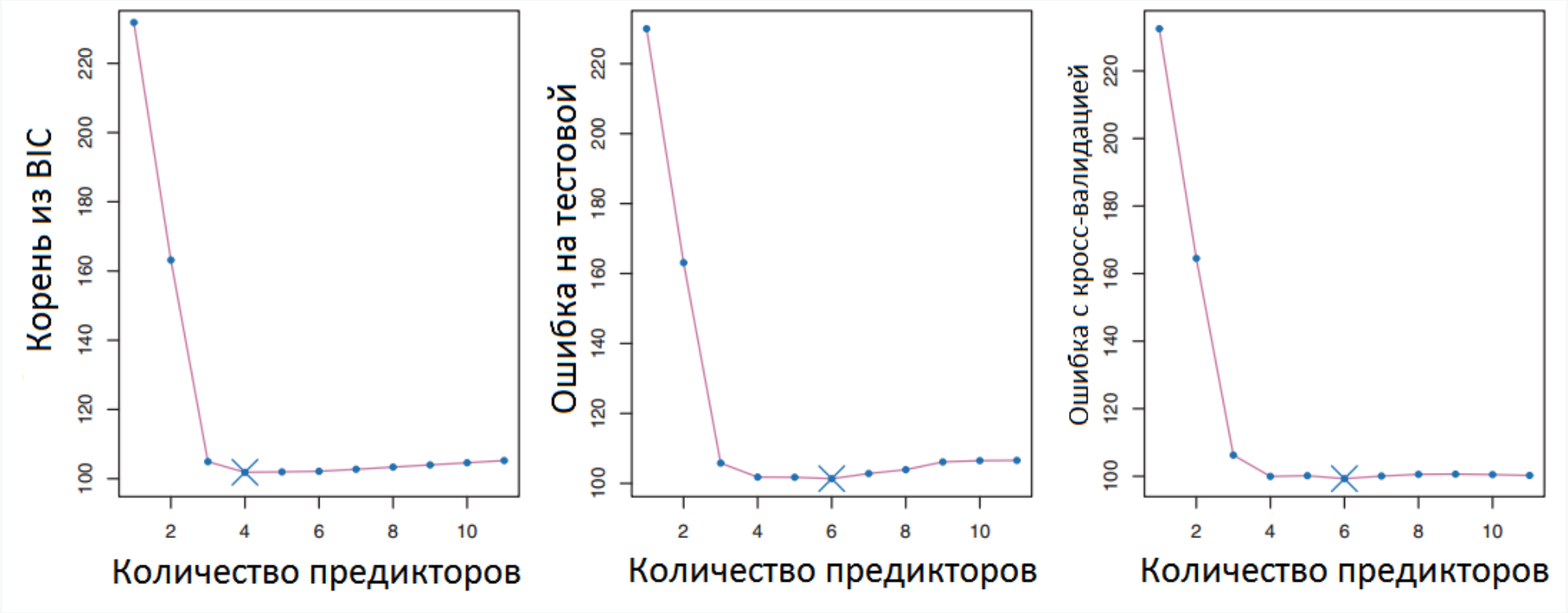
Измерители точности с поправкой



Данные `Credit`

Компромисс по минимумам C_p, AIC, максимуму R^2_{adj} и *простоте модели*: **4 объясняющих**

Второй подход: оценка ошибки непосредственно на проверочных данных



Данные `Credit`

Правило одной стандартной ошибки: (а) оценить стандартную ошибку оценок MSE ($\hat{\sigma}_{MSE}$); (б) выбрать модель в пределах $\pm\hat{\sigma}_{MSE}$ от $MSE_{\min}$

Источники

1. *Джеймс Г.*, *Уиттон Д.*, *Хасти Т.*, *Тибширани Р.* Введение в статистическое обучение с примерами на языке R. Пер. с англ. С.Э. Мастицкого – М.: ДМК Пресс, **2016** – 450 с.

2. Данные `Advertising` (http://www-bcf.usc.edu/~gareth/ISL/Advertising.csv).

3. Данные `wage.ru` (https://sites.google.com/a/kiber-guu.ru/msep/mag-econ/salary_data.csv?attredirects=0&d=1).

4. Данные `Credit` (https://rdrr.io/cran/ISLR/man/Credit.html).