Лекция 7

Линейная регрессия: продолжение

Курс: Введение в DS на УБ и МиРА (весна, 2022)

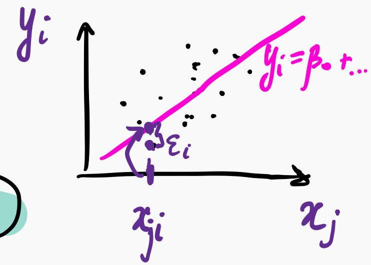
Преподаватель: Владимир Омелюсик

16 мая 2022 г.

В предыдущих сериях

- Параметрическое тестирование гипотез (статистики, p-value, доверительные интервалы).
- Непараметрическое тестирование гипотез (χ^2-критерий согласия Пирсона).
- Линейная регрессия: начало.

Линейная регрессия: напоминание



- Верим, что данные пришли из модели

предп. $\longrightarrow$

$$y_i = \underbrace{\beta_0 + \beta_1 x_{1,i} + \ldots + \beta_k x_{k,i}} + \varepsilon_i$$

- y_i – целевая (зависимая переменная), $x_{j,i}$ – регрессоры (независимые переменные), β_j – коэффициенты, ε_i – случайная ошибка.

(параметры)

- Предполагаем, что $x_{j,i}$ и $\boxed{\beta_j \text{ – константы.}}$ – обяз-о
- Линейная регрессия линейна по β_j.

ЛР

$\ln y_i = \beta_0 + \beta_1 x_i^2 + \varepsilon_i$

- Хотим по данным оценить $\underbrace{\hat{\beta}_0, \ldots, \hat{\beta}_k}$, чтобы делать предсказания как

$y_i = \beta_0 + \beta_1^2 x_i + \varepsilon_i$ – не ЛР

Оценив. $\longrightarrow$

$$\boxed{\hat{y}_i = \hat{\beta}_0 + \ldots + \hat{\beta}_k x_k}$$

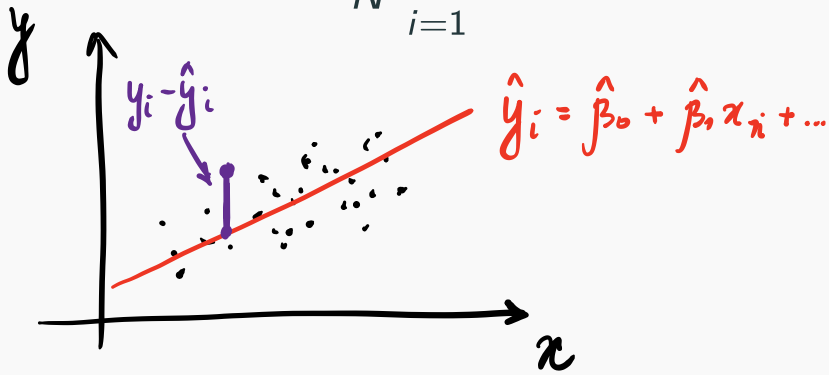
Метод наименьших квадратов

- Будем минимизировать усреднённую сумму квадратов отклонений истинного y_i от предсказанного:

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

– среднеквадр. ошибки (MSE, mean squared error)

$$= \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\beta}_0 - \ldots - \hat{\beta}_k x_{k,i})^2 \to \min_{\hat{\beta}_0,\ldots,\hat{\beta}_k}$$



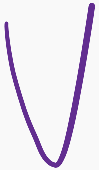
Пример: регрессия на константу

- Модель:

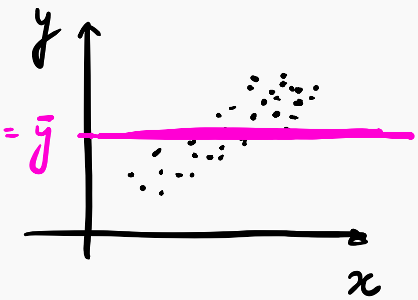
$$\boxed{y_i = \beta_0 + \varepsilon_i}$$

$$\hat{y} = \hat{\beta}_0$$

- Ищем оценку:



$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\beta}_0)^2 \to \min_{\hat{\beta}_0}$$



$$MSE'_{\hat{\beta}_0} = \cancel{\frac{1}{N}(-2)} \sum_{i=1}^{N}(y_i - \hat{\beta}_0) = 0$$

$$\sum_{i=1}^{N} y_i = \sum_{i=1}^{N} \hat{\beta}_0 = N \cdot \hat{\beta}_0$$

$$\Rightarrow \hat{\beta}_0 = \frac{\sum_{i=1}^{N} y_i}{N} = \bar{y}$$

$$\hat{y} = \hat{\beta}_0 = \bar{y}$$

Упр: $y_i = \beta_1 x_i + \varepsilon_i$

$$\hat{y} = \hat{\beta}_1 x_i$$

$$\hat{\beta}_1 = \frac{\sum_i x_i y_i}{\sum_i x_i^2}$$

Пример: парная регрессия

- Модель:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \varepsilon_i$$

- Ищем оценку:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat\beta_0 - \hat\beta_1 x_{1,i})^2 \to \min_{\hat\beta_0,\hat\beta_1}$$

- Есть готовые формулы:

$$\hat\beta_1 = \frac{\sum_{i=1}^{N}(y_i - \bar y)(x_i - \bar x)}{\sum_{i=1}^{N}(x_i - \bar x)^2},$$

$$\hat\beta_0 = \bar y - \hat\beta_1 \bar x.$$

$$\bar y = \frac{\sum_{i=1}^{N} y_i}{N}$$

$$\bar x = \frac{\sum_{i=1}^{N} x_i}{N}$$

Формула оценок для множественной регрессии

$$X = \begin{bmatrix} 1 & | & | & & | \\ 1 & x_1 & x_2 & \ldots & x_k \\ \vdots & & & & | \\ 1 & | & | & & \end{bmatrix}$$

$N \times (k+1)$

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{bmatrix}$$

$$y_i = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_k x_{k,i} + \varepsilon_i$$

$$\underset{N\times 1}{y} = \underset{N\times(k+1)}{X}\ \underset{(k+1)\times 1}{\beta} + \underset{N\times 1}{\varepsilon}$$

- Запишем модель в матричном виде

$$\beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix} \qquad y_i - \hat{y}_i \approx \hat{\varepsilon}_i$$

$$\boxed{y = X\beta + \varepsilon,}$$

$$\underset{N\times 1}{\hat{y}} = X\underset{(k+1)\times 1}{\hat{\beta}}$$

$$\|y - \hat{y}\|_2^2 = \sum_i (y_i - \hat{y}_i)^2$$

- Запишем задачу в матричном виде

$$MSE = \frac{1}{N}\|y - X\hat{\beta}\|_2^2 = (y - X\hat{\beta})^T(y - X\hat{\beta}) \to \min_{\hat{\beta}}$$

$\underset{p\times 1}{v}$ – вектор $\quad \|v\|_2^2 = \sqrt{\sum_{i=1}^{p} v_i^2}$ – евклид. норма v

- Тогда

$$\hat{\beta}_1 = \frac{\sum_i x_i y_i}{\sum_i x_i^2}$$

$$\hat{\beta} = \underbrace{(X^T X)^{-1} X^T y}_{(k+1)\times 1}$$

Интерпретация оценок коэффициентов

! При прочих равных

1. Lin-lin модель:

$$y_i = \beta_0 + \ldots + \beta_k x_{k,i} + \ldots + \varepsilon_i$$

$$\hat{y}_i = \hat{\beta}_0 + \ldots + \hat{\beta}_k x_{ki} + \ldots$$

   - При увеличении $x_{k,i}$ на единицу, y_i увеличится на β_k

2. Log-lin модель:

$$\log y_i = \beta_0 + \ldots + \beta_k x_{k,i} + \ldots + \varepsilon_i$$

$$y_i = e^{\beta_0 + \ldots + \beta_k (x_{ki}+1) + \ldots + \varepsilon_i}$$

   - При увеличении $x_{k,i}$ на единицу, y_i увеличится в e^{β_k} раз.
   - $e^{\beta_k} \approx 1 + \beta_k$ для маленьких $\beta_k \Rightarrow$при увеличении $x_{k,i}$ на единицу, y_i увеличится на β_k процентов.

Интерпретация оценок коэффициентов

3. Lin-log модель:

$\tilde{x}_{ki}$ $y_i = \beta_0 + \tilde{\beta}_1 \log x_i + \varepsilon_i$

$$y_i = \beta_0 + \ldots + \beta_k \log x_{k,i} + \ldots + \varepsilon_i$$

$y_i = \beta_0 x_i^{\beta_1} \varepsilon_i$

$\log y_i = \log \beta_0 + \beta_1 \log x_i + \varepsilon_i$

$y_i = \beta_0 + \beta_1 x_i^2 + \varepsilon_i \longrightarrow y_i = \beta_0 + 2\beta_1 \log x_i + \varepsilon_i$

- При увеличении $\log x_{k,i}$ на единицу, y_i увеличится на β_k.
- При увеличении $x_{k,i}$ на единицу, y_i увеличится на

$\tilde{y} = \tilde{\beta}_0 + \beta_1 \log x_i + \varepsilon_i$

$$\boxed{\beta_k[\log(x_{k,i} + 1) - \log(x_{k,i})]} = \beta_k \log \frac{x_{k,i} + 1}{x_{k,i}}$$

- При увеличении $x_{k,i}$ на один процент, y_i увеличится на

$$\beta_k \log \frac{1.01\cancel{x_{k,i}}}{\cancel{x_{k,i}}} = \underbrace{\beta_k \log 1.01}$$

- При увеличении $x_{k,i}$ на $p\%$, y увеличится на $\boxed{\beta_k \log([100 + p]/100).}$

Интерпретация оценок коэффициентов

4. Log-log модель:

$$\log y_i = \beta_0 + \ldots + \beta_k \log x_{k,i} + \ldots + \varepsilon_i$$

- При увеличении $x_{k,i}$ на $p\%$, y увеличится в $e^{\beta_k \log([100+p]/100)}$ раз $\Rightarrow$ на $\beta_k \log([100+p]/100)$ процентов.

Качество подгонки

M1 (4 регр.) M2 (8 регр)

- Среднеквадратичная ошибка:

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

800 300

- Средняя абсолютная ошибка:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

7.5 5.4

M2 лучше

- Коэффициент детерминации:

$R^2 \in [0,1]$

! только если есть β_0!

средние по $\hat{y}_i$

$$R^2 = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2 / N-1}{\sum_{i=1}^{N}(y_i - \bar{y})^2 / N-1} = \frac{sVar(\hat{y})}{sVar(y)}$$

среднее по y_i

$$sVar(x) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})$$

(доля выб. дисп. целевой перем, объясн. нашей моделью

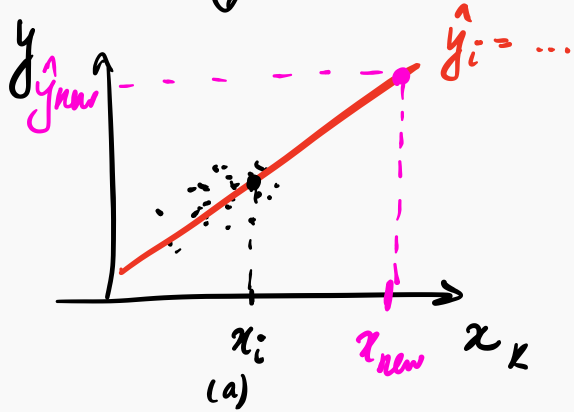
Miscellaneous

а) • Как делать предсказания на имеющейся выборке? $\hat{y}_i = \hat{\beta}_0 + \ldots + \hat{\beta}_k x_{ki}$

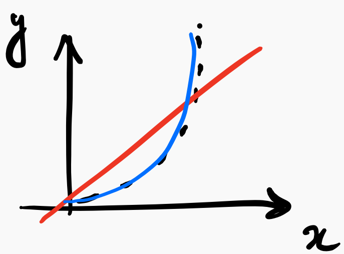
б) • Как делать предсказания на новой выборке?

в) • Как добавлять нелинейности?

$\hat{y}_{new} = \hat{\beta}_0 + \hat{\beta}_1 x_{new,1} + \ldots + \hat{\beta}_k x_{new,k}$



(в)



$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$

Проверка гипотезы о значимости отдельного коэффициента

$y_i = \beta_0 + \ldots + \beta_k x_{ki} + \ldots$

$\downarrow$

$H_0 : \beta_k = 0$ не отверг.

$$y = X\beta + \varepsilon,$$

Гипотеза о значим. β_j : $\begin{cases} H_0 : \beta_j = 0, \\ H_1 : \beta_j \neq 0 \end{cases}$

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

- Предполагаем, что $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ и все независимы.
- Можно показать, что

$$\hat\beta \sim \mathcal{MN}(\beta, \mathrm{Var}(\hat\beta))$$

$\begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix}$

- Также можно показать, что $\hat{\mathrm{Var}}(\hat\beta) = \underbrace{(X^T X)^{-1}}_{k \times k} \hat\sigma^2$. — число
- $\underbrace{\hat\sigma^2 = \dfrac{N \times MSE}{N - (k+1)}}$

$$\hat{\mathrm{Var}}(\hat\beta) = \begin{pmatrix} \hat{\mathrm{Var}}(\hat\beta_0) & \hat{\mathrm{cov}}(\hat\beta_0, \hat\beta_1) & \ldots & \hat{\mathrm{cov}}(\hat\beta_0, \hat\beta_k) \\ & \hat{\mathrm{Var}}(\hat\beta_1) & & \vdots \\ & & \ddots & \\ & & & \hat{\mathrm{Var}}(\hat\beta_k) \end{pmatrix}$$

Z-тест, t-тест и пример

- Используем Z-тест или t-тест:

$$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$$

$$\frac{\hat{\beta}_j - 0}{\sqrt{\hat{\mathrm{Var}}(\hat{\beta}_j)}} \sim t_{N-(k+1)} \to \mathcal{N}(0,1)$$

$$t = \frac{4-0}{0.8} = \frac{40}{8} = 5$$

H_0 отверг. $\Rightarrow$ коэф. значим

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$$

$$\hat{y}_i = 3 + 4x_i + 8z_i$$

	const	x	z
coef $(\hat{\beta}_j)$	3	4	8
se	0.5	0.8	1

standard error

$$se = \sqrt{\hat{\mathrm{Var}}(\hat{\beta}_j)}$$

$|||$

std

Можем пров.

1) $\begin{cases} H_0: \beta_j = \beta_0 \\ H_1: \beta_j \neq \beta_0 \end{cases}$

2) $\begin{cases} H_0: \beta_j = 0 \\ H_1: \beta_j \neq 0 \end{cases}$

H_0 не отверг. $\Rightarrow$ $\beta_j \, (x_j)$ не значим

Проверка гипотезы о значимости регрессии в целом

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

- Формулировка гипотезы о значимости в целом:

Гипотеза о знач. в целом

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0, \\ H_1 : \beta_1^2 + \ldots + \beta_k^2 > 0 \end{cases}$$

- Тестовая статистика сложная, имеет F-распределение.
- Легко проверять в софте.