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

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

- y_i – целевая (зависимая переменная), $x_{j,i}$ – регрессоры (независимые переменные), β_j – коэффициенты, ε_i – случайная ошибка.
- Предполагаем, что $x_{j,i}$ и β_j – константы.
- Линейная регрессия линейна по β_j.
- Хотим по данным оценить $\hat{\beta}_0$, ..., $\hat{\beta}_k$, чтобы делать предсказания как

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

Метод наименьших квадратов

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

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

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

Пример: регрессия на константу

- Модель:

$$y_i = \beta_0 + \varepsilon_i$$

- Ищем оценку:

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

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:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{N} (x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Формула оценок для множественной регрессии

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

- Запишем модель в матричном виде

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

- Запишем задачу в матричном виде

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

- Тогда

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

Интерпретация оценок коэффициентов

1. Lin-lin модель:

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

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

2. Log-lin модель:

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

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

Интерпретация оценок коэффициентов

3. Lin-log модель:

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

- 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

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

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

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

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

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

- Среднеквадратичная ошибка:

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

- Средняя абсолютная ошибка:

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

- Коэффициент детерминации:

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

Miscellaneous

- Как делать предсказания на имеющейся выборке?
- Как делать предсказания на новой выборке?
- Как добавлять нелинейности?

Проверка гипотезы о значимости отдельного коэффициента

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

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

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

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

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

- Используем Z-тест или t-тест:

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

Проверка гипотезы о значимости регрессии в целом

- Формулировка гипотезы о значимости 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.