НИУ ВШЭ, ОП «Политология»
Курс «Введение в многомерный статистический анализ», 2023

Лекция 3 февраля
Парная линейная регрессия

Выведение оценок коэффициентов посредством метода наименьших квадратов

1. **Случай регрессии на контанту**

Запишем исходную спецификацию:

$$y_i = \beta_0 + \varepsilon_i$$

Перепишем в терминах модельных (предсказанных) значений, то есть, отклик (зависимая переменная) в среднем равна константе (некоторому постоянному значению):

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

Руководствуясь принципом МНК, минимизируем сумму квадратов остатков:

$$\frac{\partial \sum_{i=1}^{n}(y_i - \hat{\beta}_0)^2}{\partial \hat{\beta}_0} = 0$$

$$-2\sum_{i=1}^{n}(y_i - \hat{\beta}_0) = 0$$

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 = 0$$

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i}{n}$$

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

2. **Случай парной регрессии (один предиктор)**

Найдем оптимальную оценку константы ($\hat{\beta}_0$) в парной линейной регрессии, при которой сумма квадратов остатков будет минимальна.

Рассмотрим частную производную по $\hat{\beta}_0$:

$$\frac{\partial \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\partial \hat{\beta}_0} = 0$$

$$(-2)\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1\sum_{i=1}^{n} x_i = 0$$

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1\frac{\sum_{i=1}^{n} x_i}{n}$$

Мы получили оценку константы в парной регрессии:

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

Далее давайте получим по тому же принципу МНК-оценку коэффициента при предикторе в парной линейной регрессии.

Рассмотрим частную производную по $\hat{\beta}_1$:

$$\frac{\partial \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\partial \hat{\beta}_1} = 0$$

$$(-2)\sum_{i=1}^{n}(x_i)(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \hat{\beta}_0 x_i - \sum_{i=1}^{n} \hat{\beta}_1 x_i^2 = 0$$

Вспомним, что ранее мы уже получили оценку константы, подставим ее в уравнение:

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} (\bar{y} - \hat{\beta}_1 \bar{x}) x_i - \sum_{i=1}^{n} \hat{\beta}_1 x_i^2 = 0$$

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \bar{y} x_i + \sum_{i=1}^{n} \hat{\beta}_1 \bar{x} x_i - \sum_{i=1}^{n} \hat{\beta}_1 x_i^2 = 0$$

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \bar{y} + \hat{\beta}_1 \sum_{i=1}^{n} x_i \bar{x} - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i (x_i - \bar{x})} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\widehat{Cov}(x, y)}{\widehat{Var}(x)}}$$

Заметим, что справедливым по построению модели оказалось то, что:

- сумма остатков равна 0
  $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$
- корреляция между остатками и предиктором равна 0
  $\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$, или скалярное произведение векторов $\vec{x}$ и $\vec{\varepsilon}$ равно 0, то есть, данные векторы ортогональны. Таким образом, справедливость допущения экзогенности, предполагающего нулевую корреляцию между ошибками и объясняющими переменными, проверить, посчитав корреляцию между остатками и предикторами, не получится. Как мы увидели, нулевая корреляция между остатками и предикторами заведомо верна по построению регрессионной модели.