

Занятие 1

Условное математическое ожидание и условная вариация

Задание 1. Известно, что совместный закон распределения случайных величин (далее – сл.в.) X и Y задан следующей таблицей.

$X \backslash Y$	0	1
1	0.23	0.17
2	0.32	0.09
3	0.12	...

Определите

1. условное математическое ожидание $E(Y|X = 1)$

Решение: Условное математическое ожидание $E(Y|X = x)$ показывает среднее значение сл.в. Y при условии заданного значения сл.в. X . Таким образом, для дискретной сл.в. будет справедливо, что

$$E(Y|X = x) = \sum_{i=1}^n y_i \times P(Y = y_i|X = x) = \sum_{i=1}^n y_i \times \frac{P(Y = y_i \cap X = x)}{P(X = x)}$$

$$\text{Для нашего примера } E(Y|X = 1) = 0 \times \frac{0.23}{0.23 + 0.17} + 1 \times \frac{0.17}{0.23 + 0.17} = 0.425$$

2. условную вариацию $Var(Y|X = 1)$

$$Var(Y|X = 1) = 0 \times \frac{0.23}{0.23 + 0.17} + 1 \times \frac{0.17}{0.23 + 0.17} - 0.425^2 = 0.244$$

Ряд распределения для Y^2 здесь такой же, как и исходный ряд распределения.

Выведение оценок коэффициентов парной линейной регрессии посредством метода наименьших квадратов

1. Для начала в качестве разминки рассмотрим случай регрессии на константу

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:

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

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

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

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

2. Случай парной регрессии (один предиктор)

Для начала запишем самую спецификацию модели:

Сначала в терминах истинных параметров:

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

, где y_i – зависимая переменная (отклик / объясняемая переменная),
 x_i – независимая переменная (предиктор / объясняющая переменная / регрессор),
 β_0 – коэффициент константы, β_1 – коэффициент наклона,
 ϵ – ошибка в регрессионной модели

Перепишем модель как результат оценивания по данным:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\epsilon}_i$$

,

где $\hat{\beta}_0, \hat{\beta}_1$ – оценки коэффициентов, $\hat{\epsilon}_i$ – остаток (то есть, оценка ошибки)

$$y_i = \hat{y}_i + \hat{\epsilon}_i$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = E(Y|X)$$

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:

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

Далее мы получим 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:

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

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

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

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