- Линейная регрессия. Основные предположения

Линейная регрессия – <u>линейная</u> зависимость одной переменно y от другой или нескольких других переменных x.

$y = f(x,b) + \varepsilon$, где:

$$f(x,b) = b_0 + b_1x_1 + b_2x_2 \dots = \sum_{i=1}^{n} b_i x_i = x^T b$$

ε — случайная ошибка модели

Предположения:
1) $E(\varepsilon_i) = 0$

2) $\sigma^2 = const$

3) $\forall\ i,j\ (i \neq j)\colon cov(\varepsilon_i, \varepsilon_j) = 0$

4) x_i — неслучайная величина

5*) ε имеет нормальное распределение (вроде как необязательное)

- Метод наименьших квадратов и его свойства

МНК – метод основанный на минимизации суммы квадратов отклонений некоторых функций от экспериментальных данных.

В методе мы находим такой b из $f(x,b)$, чтобы $f(x,b)$ была максимально «близко» к y.

$Xb = f(x,b)$

$$SS(b) = \sum_{i=1}^{n} (y_i - f(x_i, b_i))^2 = \varepsilon^T \varepsilon = \sum_{i=1}^{n} \varepsilon_i^2$$

$\hat{b} = \arg\min_{\mathrm{b}} SS(b)$

Ищем b такое, чтобы сумма квадратов отклонений была минимальной

$\varepsilon = y - Xb \Rightarrow SS = \varepsilon^T \varepsilon = (y - Xb)^T (y - Xb)$

Дифференцируем функцию по вектору параметров b и приравниваем производные к нулю:

$(X^T X) b = X^T y \Rightarrow b = (X^T X)^{-1} X^T y$

Свойства МНК:

1) Несмещенность оценки. Из предположения №1

2) Состоятельность. Из предположения №1 и независимости x и ε

3) Эффективная. Из предположения №2 и №3

- Основная теорема о линейной регрессии. Следствия из нее

$\hat{b}$ и $SS(\hat{b})$ – независимы; $SS(\hat{b})$ и $SS(b) - SS(\hat{b})$ – независимы

$\hat{b} \sim N(b, \sigma^2 X^T X)$

$m = rank(X^T X)$

$$\frac{SS(\hat{b})}{\sigma^2} \sim \chi^2(n-m)$$

$$\frac{SS(b) - SS(\hat{b})}{\sigma^2} \sim \chi^2(m)$$

Следствия:

1) $(\hat{b}_i - b_i)\sqrt{\frac{n-m}{A_{ii}^{-1} SS(\hat{b})}} \sim T(n-m)$, доверительный интервал для b_i,

$t -$ критерий $(H_0: b_i = 0)$, где $A = X^T X$

2) $\frac{n-m}{m} \cdot \frac{SS(b) - SS(\hat{b})}{SS(\hat{b})} \sim F(m,\ n-m)$

- Остаточная дисперсия. Коэффициент детерминации

Остаточная дисперсия - это мера разброса остатков (ошибок) модели, которые не объясняются независимыми переменными. Она является одним из показателей точности модели и может быть использована для оценки качества подгонки модели к данным. Чем меньше остаточная дисперсия, тем лучше модель соответствует данным.

$$\sigma^2_{\text{ост}} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-m}$$

Коэффициент детерминации (R^2) — это доля дисперсии зависимой переменной, объясняемая рассматриваемой моделью зависимости, то есть объясняющими переменными. Более точно — это единица минус доля необъяснённой дисперсии в дисперсии зависимой переменной.

$$R^2 = 1 - \frac{\sigma^2}{\sigma_y^2}$$

Где σ_y^2 – дисперсия случайной величины y, а σ^2 – условная дисперсия ($D(y|x)$) зависимой переменной

Проблема применения R^2 заключается в том, что его значение увеличивается от добавления в модель новых переменных, даже если эти переменные никакого отношения к объясняемой переменной не имеют.