Семинар 7.
Безусловное прогнозирование.

1. Проверьте формулу (из лекции) для среднеквадратической ошибки прогноза:

$$\mathbb{E}\left(\hat{y}_{n+1} - y_{n+1}\right)^2 = \sigma^2\left(1 + x_{n+1}\left(X^TX\right)^{-1}x_{n+1}^T\right).$$

Здесь $\hat{y}_{n+1}$ — прогнозное значение зависимой переменной для нового наблюдения, y_{n+1} — истинное значение зависимой переменной для нового наблюдения, x_{n+1} — $1 \times k$ вектор-строка объясняющих переменных для нового наблюдения.

Решение:
Ошибка прогноза равна

$$\hat{y}_{n+1} - y_{n+1} = x_{n+1}\hat{\beta} - (x_{n+1}\beta + \varepsilon_{n+1}).$$

Поскольку оценка $\hat{\beta}$ несмещенная и $\mathbb{E}(\varepsilon_{n+1}) = 0$, то математическое ожидание ошибки прогноза равно 0:

$$\mathbb{E}(\hat{y}_{n+1} - y_{n+1}) = \mathbb{E}(x_{n+1}\hat{\beta}) - x_{n+1}\beta - \mathbb{E}(\varepsilon_{n+1}) = x_{n+1}\beta - x_{n+1}\beta = 0.$$

Отсюда получаем:

$$\begin{aligned}\mathbb{E}(\hat{y}_{n+1} - y_{n+1})^2 &= V(\hat{y}_{n+1} - y_{n+1}) = V(x_{n+1}\hat{\beta} - x_{n+1}\beta - \varepsilon_{n+1}) = V(x_{n+1}\hat{\beta} - \varepsilon_{n+1}) \\ &= V(x_{n+1}\hat{\beta}) + V(\varepsilon_{n+1}) - 2\text{Cov}(x_{n+1}\hat{\beta}, \varepsilon_{n+1}).\end{aligned}$$

Последнее слагаемое равно 0, так как ε_{n+1} и $\hat{\beta}$ некоррелированы. Таким образом,

$$\begin{aligned}\mathbb{E}(\hat{y}_{n+1} - y_{n+1})^2 &= V(x_{n+1}\hat{\beta}) + V(\varepsilon_{n+1}) = x_{n+1}V(\hat{\beta})x_{n+1}^T + \sigma^2 \\ &= \sigma^2 x_{n+1}(X^TX)^{-1}x_{n+1} + \sigma^2 = \sigma^2(1 + x_{n+1}(X^TX)^{-1}x_{n+1}^T),\end{aligned}$$

что и требовалось показать.

2. (самостоятельно, используя формулу из задания 1) Докажите равенство для среднеквадратической ошибки прогноза в случае парной регрессии с константой вида $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ следующему выражению:

$$\mathbb{E}\left(\hat{y}_{n+1} - y_{n+1}\right)^2 = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2)}\right).$$

Решение:
Из задачи 1 очевидно, что достаточно показать, что

$$x_{n+1}(X^TX)^{-1}x_{n+1}^T = \frac{1}{n} + \frac{(x_{n+1}-\overline{x})^2}{\sum(x_i-\overline{x})^2}.$$

В случае парной регрессии с константой $k = 2$ и

$$X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad x_{n+1} = \begin{bmatrix} 1 & x_{n+1} \end{bmatrix}.$$

Тогда

$$\begin{aligned} x_{n+1}(X^TX)^{-1}x_{n+1}^T &= [1 \quad x_{n+1}] \begin{bmatrix} n & n\overline{x} \\ n\overline{x} & \sum x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ x_{n+1} \end{bmatrix} \\ &= \frac{1}{n\sum x_i^2 - n^2\overline{x}^2}[1 \quad x_{n+1}] \begin{bmatrix} \sum x_i^2 & -n\overline{x} \\ -n\overline{x} & n \end{bmatrix} \begin{bmatrix} 1 \\ x_{n+1} \end{bmatrix} \\ &= \frac{1}{n(\sum x_i^2 - n\overline{x}^2)} \left(\sum x_i^2 - 2n\overline{x}x_{n+1} + nx_{n+1}^2 \right) \\ &= \frac{1}{n(\sum x_i^2 - n\overline{x}^2)} \left(n(x_{n+1}^2 - 2\overline{x}x_{n+1} + \overline{x}^2) + \sum x_i^2 - n\overline{x}^2 \right) \\ &= \frac{n(x_{n+1}-\overline{x})^2}{n(\sum x_i^2 - n\overline{x}^2)} + \frac{\sum x_i^2 - n\overline{x}^2}{n(\sum x_i^2 - n\overline{x}^2)} \\ &= \frac{(x_{n+1}-\overline{x})^2}{\sum(x_i-\overline{x})^2} + \frac{1}{n}. \end{aligned}$$

Получаем, что среднеквадратическая ошибка прогноза $\hat{y}_{n+1}$ в случае парной регрессии с константой имеет вид:

$$\mathbb{E}(\hat{y}_{n+1} - y_{n+1})^2 = \sigma_\varepsilon^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1}-\overline{x})^2}{\sum(x_t-\overline{x})^2} \right),$$

чтд.

3. Для модели парной регрессии $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, $i = 1,...,10$, известно, что

$$\sum_{i=1}^{10} y_i = 8, \sum_{i=1}^{10} x_i = 40, \sum_{i=1}^{10} y_i^2 = 26, \sum_{i=1}^{10} x_i^2 = 200, \sum_{i=1}^{10} y_i x_i = 20.$$

Для некоторого наблюдения дано $x_{11} = 10$. Предполагая, что данное наблюдение удовлетворяет исходной модели,

(a) вычислите наилучший линейный несмещенный прогноз величины y_{11};

(b) оцените стандартную ошибку прогноза.

Решение:

(a) Имеем:

$$\hat{\beta}_2 = \frac{\sum_{i=1}^n x_i y_i - n\overline{xy}}{\sum_{i=1}^n x_i^2 - n\overline{x}^2} = \frac{20 - 10 \cdot 4 \cdot 0.8}{200 - 10 \cdot 4^2} = -0.3,$$

$$\hat{\beta}_1 = \overline{y} - \hat{\beta}_2 \overline{x} = 0.8 + 0.3 \cdot 4 = 2.$$

Тогда наилучший линейный несмещенный прогноз величины y_{11} равен:

$$\hat{y}_{11} = \hat{\beta}_1 + \hat{\beta}_2 x_{11} = 2 - 0.3 \cdot 10 = -1.$$

(b) Найдем сумму квадратов остатков RSS в исходной регрессии. Учитывая, что $\sum_{i=1}^n e_i = 0$ (верно только для модели с константой) и $\sum_{i=1}^n x_i e_i = 0$ (из условия первого порядка в оптимизационной задаче для МНК), получаем:

$$\begin{aligned} RSS &= \sum e_i^2 = \sum e_i(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = \sum e_i y_i \\ &= \sum (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) y_i \\ &= \sum y_i^2 - \hat{\beta}_1 \sum y_i - \hat{\beta}_2 \sum x_i y_i \\ &= 26 - 2 \cdot 8 + 0.3 \cdot 20 = 16. \end{aligned}$$

Оценку дисперсии ошибок получаем по формуле

$$\hat{\sigma}_\varepsilon^2 = \frac{RSS}{n-2} = \frac{16}{8} = 2.$$

Согласно формуле из задания 2 оценка среднеквадратичной ошибки прогноза есть величина

$$\delta = \hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x_s - \overline{x})^2}{\sum (x_t - \overline{x})^2}\right) = 2 \left(1 + \frac{1}{10} + \frac{(10-4)^2}{200 - 10 \cdot 4^2}\right) = 4.$$

Таким образом, оценка стандартной ошибки прогноза равна 2.