Задача ??. Для регрессионной модели $Y_i = \beta x_i + \varepsilon_i$, $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{D}(\varepsilon_i) = \sigma^2$, $\operatorname{cov}(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$, докажите частный случай теоремы Гаусса–Маркова. А именно, покажите, что

(a) МНК-оценка $\hat{\beta} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}$ является линейной по вектору $Y = (Y_1, \dots, Y_n)^T$,

(b) МНК-оценка $\hat{\beta} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}$ является несмещенной оценкой параметра β,

(c) МНК-оценка $\hat{\beta} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}$ имеет наименьшую дисперсию в классе всех линейных по $Y = (Y_1, \dots, Y_n)^T$, несмещенных оценок параметра β.

Решение.

(a) Пусть $Y = (Y_1, \dots, Y_n)^T$, $Z = (Z_1, \dots, Z_n)^T$ и $\lambda \in \mathbb{R}$. Функция $\hat{\beta}(X, Y) = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}$ является линейной по переменной Y, поскольку

$$\hat{\beta}(X, Y + Z) = \frac{\sum_{i=1}^n x_i (Y_i + Z_i)}{\sum_{i=1}^n x_i^2} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2} + \frac{\sum_{i=1}^n x_i Z_i}{\sum_{i=1}^n x_i^2} = \hat{\beta}(X, Y) + \hat{\beta}(X, Z)$$

и

$$\hat{\beta}(X, \lambda Y) = \frac{\sum_{i=1}^n x_i (\lambda Y_i)}{\sum_{i=1}^n x_i^2} = \lambda \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2} = \lambda \hat{\beta}(X, Y).$$

(b) Оценка $\hat{\beta} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}$ является несмещенной, поскольку

$$\mathbb{E}[\hat{\beta}] = \mathbb{E}\left[\frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}\right] = \frac{1}{\sum_{i=1}^n x_i^2} \mathbb{E}\left[\sum_{i=1}^n x_i Y_i\right] = \frac{1}{\sum_{i=1}^n x_i^2} \sum_{i=1}^n x_i \mathbb{E}[Y_i] = \frac{1}{\sum_{i=1}^n x_i^2} \sum_{i=1}^n x_i \beta x_i = \beta.$$

(c) Прежде всего, заметим, что всякая линейная по вектору $Y = (Y_1, \dots, Y_n)^T$ оценка $\tilde{\beta}$ имеет вид:

$$\tilde{\beta} = \sum_{i=1}^n c_i Y_i.$$

Представим МНК-оценку $\hat{\beta} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}$ в виде

$$\hat{\beta} = \sum_{i=1}^n a_i Y_i,$$

где $a_i := \frac{x_i}{\sum_{j=1}^n x_j^2}$, $i = 1, \dots, n$. Тогда оценку $\tilde{\beta}$ можно переписать в виде

$$\tilde{\beta} = \sum_{i=1}^n (a_i + b_i) Y_i,$$

где $b_i := c_i - a_i$, $i = 1, \dots, n$.

Далее, учитывая, что МНК-оценка $\hat{\beta}$ является несмещенной, т.е. $\mathbb{E}[\hat{\beta}]=\beta$ для любого $\beta\in\mathbb{R}$, получаем соотношение

$$\mathbb{E}[\hat{\beta}]=\mathbb{E}\left[\sum_{i=1}^{n}a_iY_i\right]=\sum_{i=1}^{n}a_i\mathbb{E}[Y_i]=\sum_{i=1}^{n}a_i\beta x_i=\beta\sum_{i=1}^{n}a_ix_i=\beta,$$

верное при любом $\beta\in\mathbb{R}$. Следовательно, верное и при $\beta=1$. Значит,

$$\sum_{i=1}^{n}a_ix_i=1. \qquad (*)$$

В свою очередь, оценка $\tilde{\beta}$ также является несмещенной, т.е. $\mathbb{E}[\tilde{\beta}]=\beta$ для любого $\beta\in\mathbb{R}$. Следовательно, имеет место соотношение

$$\mathbb{E}[\tilde{\beta}]=\mathbb{E}\left[\sum_{i=1}^{n}(a_i+b_i)Y_i\right]=\sum_{i=1}^{n}(a_i+b_i)\mathbb{E}[Y_i]=\sum_{i=1}^{n}(a_i+b_i)\beta x_i=\beta\sum_{i=1}^{n}(a_i+b_i)x_i=\beta,$$

верное при любом $\beta\in\mathbb{R}$. Стало быть, верное и при $\beta=1$. Поэтому

$$\sum_{i=1}^{n}(a_i+b_i)x_i=1. \qquad (**)$$

Из соотношений $(*)$ и $(**)$ получаем, что

$$\sum_{i=1}^{n}b_ix_i=0. \qquad (***)$$

В свою очередь, отсюда следует, что

$$\sum_{i=1}^{n}a_ib_i=\sum_{i=1}^{n}\frac{x_i}{\sum_{j=1}^{n}x_j^2}b_i=\frac{1}{\sum_{j=1}^{n}x_j^2}\underbrace{\sum_{i=1}^{n}x_ib_i}_{=0}=0. \qquad (****)$$

Теперь сравним дисперсии $\mathrm{D}(\hat{\beta})$ и $\mathrm{D}(\tilde{\beta})$. Имеем:

$$\mathrm{D}(\hat{\beta})=\mathrm{D}(\sum_{i=1}^{n}a_iY_i)=\sum_{i=1}^{n}\mathrm{D}(a_iY_i)=\sum_{i=1}^{n}a_i^2\,\mathrm{D}(Y_i)=$$

$$=\sum_{i=1}^{n}a_i^2\,\mathrm{D}(\beta x_i+\varepsilon_i)=\sum_{i=1}^{n}a_i^2\,\mathrm{D}(\varepsilon_i)=\sum_{i=1}^{n}a_i^2\sigma^2=\sigma^2\sum_{i=1}^{n}a_i^2;$$

$$\mathrm{D}(\tilde{\beta})=\mathrm{D}(\sum_{i=1}^{n}(a_i+b_i)Y_i)=\sum_{i=1}^{n}\mathrm{D}((a_i+b_i)Y_i)=\sum_{i=1}^{n}(a_i+b_i)^2\,\mathrm{D}(Y_i)=$$

$$=\sum_{i=1}^{n}(a_i+b_i)^2\,\mathrm{D}(\beta x_i+\varepsilon_i)=\sum_{i=1}^{n}(a_i+b_i)^2\sigma^2=\sigma^2\sum_{i=1}^{n}(a_i+b_i)^2=$$

$$=\sigma^2\sum_{i=1}^{n}a_i^2+2\sigma^2\underbrace{\sum_{i=1}^{n}a_ib_i}_{=0\ (\text{по}\ (****))}+\sigma^2\underbrace{\sum_{i=1}^{n}b_i^2}_{\geq 0}\geq\sigma^2\sum_{i=1}^{n}a_i^2=\mathrm{D}(\hat{\beta}),$$

т.е. дисперсия МНК-оценки $\hat{\beta}$ нестрого меньше, чем дисперсия произвольной линейной по вектору $Y=(Y_1,\ldots,Y_n)^T$ и несмещенной оценки $\tilde{\beta}$. $\square$