Statistical Data Analysis

Jana de Wiljes

wiljes@uni-potsdam.de

www.dewiljes-lab.com

30. Oktober 2022

Universität Potsdam

Best linear unbiased estimator (BLUE)

Linear estimator

Def: A linear estimator has the form

$$\hat{\beta}^L = \mathbf{b} + \mathbf{A}\mathbf{y} \tag{1}$$

where $\mathbf{b} \in \mathbb{R}^{(p+1)\times 1}$ and $\mathbf{A} \in \mathbb{R}^{(p+1)\times n}$.

Example: The LS-estimator:

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y} \tag{2}$$

is a linear estimator with $\mathbf{b} = \mathbf{0}$ and $\mathbf{A} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top$

Best linear unbiased estimator (BLUE)

Theorem: The LS-estimator is BLUE. This means that the LS-estimator has minimal variance among all linear and unbiased estimators $\hat{\beta}^L$

$$\mathrm{Var}(\hat{\beta}_j) \leq \mathrm{Var}(\hat{\beta}_j^L), \quad j = 0, \ldots, p. \tag{3}$$

Furthermore, for an arbitrary linear combination $\mathbf{c}^\top \hat{\beta}$ it holds that

$$\mathrm{Var}(\mathbf{c}^\top \hat{\beta}) \leq \mathrm{Var}(\mathbf{c}^\top \hat{\beta}^L) \tag{4}$$

Proof:

$$\mathbb{E}[\hat{\beta}^L] = \mathbb{E}][\underbrace{b + Ay}_{=\hat{\beta}^L}] = \mathbb{E}[b] + \mathbb{E}[A(X\beta + \epsilon)] \tag{5}$$

$$= b + \mathbb{E}[AX\beta] + A\underbrace{\mathbb{E}[\epsilon]}_{=0} = b + AX\beta = \beta \tag{6}$$

For the special: $\beta = 0 : b + AX\mathbf{0} = 0 \implies b = 0$
Short: for $\hat{\beta}^L$ unbiased $\implies b = 0$

Proof

Further: for $\hat{\beta}^L$ unbiased $\implies AX\beta = \beta$:

$$\begin{aligned} AX\beta = \beta &\iff (AX - I_{p+1})\beta = 0 && (7)\\ &\iff AX = I_{p+1} && (8) \end{aligned}$$

Note:

- I_{p+1} has full rank: $\text{rk}(I_{p+1}) = p + 1$
- X has full rank
- condition: $(rk)(A) = p + 1$ then
  $\text{rk}(AX) = \min(\text{rk}(X), \text{rk}(A))p + 1 = \text{rk}(I_{p+1}) = p + 1$

Let the matrix without los sof generality be of them form

$$A = (X^\top X)^{-1} X^\top + B \tag{9}$$

Inserting into unbiasedness condition $I_{p+1} = AX$ yields

$$I_{p+1} = AX = \underbrace{(X^\top X)^{-1} X^\top X}_{I_{p+1}} + BX = I_{p+1} + BX \tag{10}$$

$\implies BX = 0$

Proof

Note: $\text{Cov}(\hat{\beta}^L) = \sigma^2 AA^\top$

$$\begin{aligned}
\text{Cov}(\hat{\beta}^L) &= \sigma^2 AA^\top && (11)\\
&= \sigma^2((X^\top X)^{-1}X^\top + B)((X^\top X)^{-1}X^\top + B^\top) && (12)\\
&= \sigma^2((X^\top X)^{-1}X^\top + B)((X^\top X)^{-1}X^\top + B^\top) && (13)
\end{aligned}$$

We know $BB^\top$ is positive semi-definite and

$$\begin{aligned}
&\text{Cov}(\hat{\beta}^L) - \text{Cov}(\hat{\beta}) = \sigma^2 BB^\top \geq 0 && (14)\\
&c^\top \text{Cov}(\hat{\beta}^L)c - c^\top \text{Cov}(\hat{\beta})c = \sigma^2 BB^\top \geq 0 \quad \forall c \in \mathbb{R}^{p+1} && (15)\\
&\implies \text{Var}(c^\top \hat{\beta}^L) \geq \text{Var}(c^\top \hat{\beta}) && (16)\\
&\text{Var}(c^\top \hat{\beta}^L) = c^\top \text{Cov}(\hat{\beta}^L)c && (17)
\end{aligned}$$

and $\text{Var}(c^\top \hat{\beta}) = c^\top \text{Cov}(\hat{\beta})c$ As c is a abitrary we can choose for each $j = 0, \ldots, p$ $c = (0, \ldots, 1, 0 \ldots 0)$

$$\text{Var}(\hat{\beta}_j) = \text{Var}(c^\top \hat{\beta}^L) \geq \text{Var}(c^\top \hat{\beta}_j) \tag{18}$$

□

Coefficient of determination

Def: The coefficient of determination is defined by

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{19}$$

and measures the proportion of variability in y that is accounted for by the statistical model from the overall variation in y.

Coefficient of determination

Lemma: The method of least squares yields the following geometrical results:

(i) The fitted values $\hat{\mathbf{y}}$ are orthogonal to the residuals $\hat{\epsilon}$, i.e., $\hat{\mathbf{y}}^\top \hat{\epsilon} = 0$.

(ii) The columns of $\mathbf{X}$ are orthogonal to the residuals $\hat{\epsilon}$, i.e., $\mathbf{X}^\top \hat{\epsilon} = 0$

(iii) The residuals are zero on average, i.e.,

$$\sum_{i=1}^{n} \hat{\epsilon}_i = 0 \quad \text{and} \quad \bar{\hat{\epsilon}} = \frac{1}{n} \sum_{i=1}^{n} \hat{\epsilon}_i = 0 \tag{20}$$

(iv) The mean of the estimated values

$$\bar{\hat{y}} = \frac{1}{n} \sum_{i=1}^{n} \hat{y}_i = \bar{y} \tag{21}$$

Proof

Proof of (i): We will be using $H = X(X^\top X)^{-1}X^\top$, $\hat{y} = X\hat{\beta}$ and $\hat{\epsilon} = y - X\hat{\beta} = y - \hat{y}$

$$\hat{y}^\top \hat{\epsilon} = (X(X^\top X)^{-1}X^\top y)^\top (y - X\hat{\beta}) \quad (22)$$

$$= y^\top X(X^\top X)^{-1}X^\top (y - Hy) \quad (23)$$

$$= y^\top X(X^{)-1}X^\top (y - Hy) \quad (24)$$

$$= y^\top H(Id - H)y = y^\top Hy - y^\top HHy = 0 \quad (25)$$

$\square$

Proof of (ii):

$$X^\top \hat{\epsilon} = X^\top (y - \hat{y}) \quad (26)$$

$$= X^\top y - X^\top Hy \quad (27)$$

$$= X^\top y - X^\top X(X^\top X)^{-1}X^\top y \quad (28)$$

$$= X^\top y - X^\top y = 0 \quad (29)$$

Proof

Proof of (iii): Reminder:

$$X = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & & \dots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix} \tag{30}$$

$\implies X_0 = (1, \dots, 1)^\top$

$$0 \underbrace{=}_{\text{using } (ii)} X_0^\top \hat{\epsilon} = 1^\top \hat{\epsilon} = \sum_{i=1}^{n} \hat{\epsilon}_i \tag{31}$$

$\square$

Proof of (iv): Using (iii) we have

$$\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} (y_i - \hat{\epsilon}_i) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \epsilon_i = \sum_{i=1}^{n} y_i \tag{32}$$

$\square$

Coefficient of determination

Lemma: The following decomposition holds:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}\hat{\epsilon}_i^2 \tag{33}$$

Proof: First we define the $n \times n$ matrix

$$C = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top \tag{34}$$

Note that C is symmetric and that

$$\begin{aligned} CC &= (I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top)(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top) && (35) \\ &= (I_n - 2\frac{1}{n}I_n\mathbf{1}\mathbf{1}^\top + \frac{1}{n^2}\mathbf{1}\underbrace{\mathbf{1}^\top\mathbf{1}}_{\frac{1}{n}}\mathbf{1}^\top) && (36) \\ &= (I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top) = C && (37) \\ & && (38) \end{aligned}$$

Proof

Let $a \in \mathbb{R}^n$ be an abitrary vector then $Ca = \begin{pmatrix} a_1 - \bar{a} \\ \vdots \\ a_n - \bar{a} \end{pmatrix}$ and

$$a^\top Ca = \sum_{i=1}^{n} (a_i - \bar{a})^2$$

Considering $y = \hat{y} + \hat{\epsilon}$ and multiply it with C yields:

$$Cy = C\hat{y} + \underbrace{C\hat{\epsilon}}_{\begin{pmatrix} \hat{\epsilon}_1 - \bar{\hat{\epsilon}} \\ \vdots \\ \hat{\epsilon}_n - \bar{\hat{\epsilon}} \end{pmatrix} \text{and note that} \frac{1}{n} \sum_{i=1}^{n} \hat{\epsilon}_i = 0} \tag{39}$$

$$= C\hat{y} + \hat{\epsilon} \tag{40}$$

Proof

Using that result we obtain

$$\begin{aligned} y^\top CCy &= (\hat{y}^\top C + \hat{\epsilon}^\top)(C\hat{y} + \hat{\epsilon}) \\ &= \hat{y}^\top CC\hat{y} + \hat{y}^\top C\hat{\epsilon} - \hat{\epsilon}^\top C\hat{y} + \hat{\epsilon}^\top\hat{\epsilon} \\ &= \hat{y}^\top CC\hat{y} + \hat{y}^\top C\hat{\epsilon} - \hat{\epsilon}^\top C\hat{y} + \hat{\epsilon}^\top\hat{\epsilon} \quad \text{using } CC = C \text{ and } C\hat{\epsilon} = \hat{\epsilon} \end{aligned}$$

Combining $CC = C$ and equation (40) leads to

$$y^\top CCy = y^\top Cy = \sum_{i=1}^{n}(y - \bar{y})^2 \tag{41}$$

Further we use that $\bar{\hat{y}} = \bar{y}$ and obtain

$$\hat{y}^\top C\hat{y} = \sum_{i=1}^{n}(\hat{y} - \bar{y})^2 \tag{42}$$

$$\implies \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}\hat{\epsilon}_i^2 \tag{43}$$

$\square$

Coefficient of determination

Lemma: The coefficient of determination R^2 can be transformed into

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\hat{\beta}^\top \mathbf{X}^\top \mathbf{y} - n\bar{y}^2}{\mathbf{y}^\top \mathbf{y} - n\bar{y}^2} \tag{44}$$

Def: The corrected coefficient of determination $\bar{R}^2$ is defined by

$$\bar{R}^2 = 1 - \left(\frac{n-1}{n-p-1}\right)(1 - R^2) \tag{45}$$

Asymptotic Properties of the LS-Estimator

Proposition: Consider the setting

$$\mathbf{y}_n = \mathbf{X}_n\beta + \epsilon_n \quad \text{with } \mathbb{E}[\epsilon_n] = \mathbf{0} \quad \text{and } Cov(\epsilon_n) = \sigma^2\mathbf{I}_n \tag{46}$$

with the following assumption being fulfilled:

$$\lim_{n\to\infty} \frac{1}{n}\mathbf{X}_n^\top\mathbf{X}_n = \mathbf{V} \tag{47}$$

where $\mathbf{V}$ is positive definite. Then

- The LS-estimator $\hat{\beta}_n$ for β as well as the ML- and REML-estimators $\hat{\sigma}_n^2$ for σ^2 are consistent. ($\text{MSE}_\theta(\hat{\theta}) \to 0$ $n \to \infty$)
- The LS-estimator $\hat{\beta}_n$ for β is asymptotically normally distributed:

$$\sqrt{n}(\hat{\beta}_n - \beta) \to \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{V}^{-1}) \text{ (in distribution)} \tag{48}$$

Asymptotic Properties of the LS-Estimator

Proposition: Hence, for sufficiently large n it follows that $\hat{\beta}_n$ is approximately normally distributed with

$$\hat{\beta}_n \to \mathcal{N}(\beta, \sigma^2 \mathbf{V}^{-1}/n) \text{ (almost surely)} \tag{49}$$

Proposition:

- Similar to the error terms, also the residuals have expectation zero.
- In contrast to the error terms, the residuals are not uncorrelated.

Asymptotic Properties of the LS-Estimator

Proposition: Beside the usual assumptions, additionally assume that the error terms are normally distributed. Then the following properties hold:

- The distribution of the squared sum of residuals is given by:

$$\frac{\hat{\epsilon}^\top \hat{\epsilon}}{\sigma^2} = (n - p - 1)\frac{\hat{\sigma}^2}{\sigma^2} \tag{50}$$

- The squared sum of residuals $\hat{\epsilon}^\top \hat{\epsilon}$ and the LS-estimator $\hat{\beta}$ are independent.

Prediction

Proposition:

1. The expected prediction error is zero i.e., $\mathbb{E}[\hat{\mathbf{y}}_0 - \mathbf{y}_0] = 0$, i.e., $\mathbb{E}[\hat{\mathbf{y}}_0 - \mathbf{y}_0] = 0$
2. Prediction error covariance matrix is given by:

$$\mathbb{E}[(\hat{\mathbf{y}}_0 - \mathbf{y}_0)(\hat{\mathbf{y}}_0 - \mathbf{y}_0)^\top] = \sigma^2(\mathbf{X}_0(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_0^\top + \mathbf{I}_{T_0}) \qquad (51)$$

Proof

Proof of (i): The true value is given by $y_0 = X_0\beta + \epsilon_0$. For the prediction error $\hat{y}_0 - y$ one obtains

$$\mathbb{E}[\hat{y}_0 - y_0] = \mathbb{E}[X_0\hat{\beta} - X_0\beta - \epsilon_0] \tag{52}$$

$$= \mathbb{E}[X_0(\hat{\beta} - \beta) - \epsilon] \tag{53}$$

$$= X_0 \underbrace{\mathbb{E}[\hat{\beta} - \beta]}_{\mathbb{E}[\hat{\beta}]-\beta} - \underbrace{\mathbb{E}[\epsilon_0]}_{=0} = 0 \tag{54}$$

Proof of (ii): For the prediction error variance are obtains

$$\begin{aligned}
\mathbb{E}[(\hat{y}_0 - y_0)^\top(\hat{y}_0 - y_0)] &= \mathbb{E}[(X_0(\hat{\beta} - \beta - \epsilon))(X_0(\hat{\beta} - \beta - \epsilon)^\top] \\
&= X_0\mathbb{E}[(\hat{\beta} - \beta - \epsilon)(\hat{\beta} - \beta - \epsilon)^\top]X_0^\top + \mathbb{E}[\epsilon_0\epsilon_0^\top] \\
&- X_0\mathbb{E}[(\hat{\beta} - \beta)\epsilon_0^\top] - \underbrace{\mathbb{E}[\epsilon_0(\hat{\beta} - \beta)^\top]}_{\epsilon_0 \text{ and } (\hat{\beta}-\beta) \text{ are independent}} X_0^\top \\
&= \sigma^2(X_0(X^\top X)^{-1}X_0^\top + I)
\end{aligned}$$

□