Statistical Data Analysis

Dr. Jana de Wiljes

30. November 2021

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 $\mathbf{A} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top$

Gauss-Markov

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

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

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

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

LS

Proof

Proof: $\mathbb{E}[\hat{\beta}^L] = \mathbb{E}[\underbrace{b + Ay}_{=\hat{\beta}^L}] = \mathbb{E}[b] + \mathbb{E}[A(X\beta + \varepsilon)]$

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

$= b + AX\beta \overset{!}{=} \beta \quad \forall \beta \in \mathbb{R}^{p+1}$

For the special $\beta = 0$: $b + \underbrace{AX \cdot 0}_{=0} = 0 \quad (=\beta) \Rightarrow b = 0$

$\Rightarrow$ For $\hat{\beta}^L$ unbiased $\Rightarrow \boxed{b=0}$

$\forall \beta \in \mathbb{R}^{p+1}$

$\boxed{AX\beta = \beta}$

$\Leftrightarrow (AX - I_{p+1})\beta = 0$

$\Leftrightarrow \boxed{AX = I_{p+1}}$

$B(X) = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$

$f: V \xrightarrow{B} W, \; x \mapsto Bx$

if B has full rank then $x = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$ is only solution

- I_{p+1} has full rank: $rk(I_{p+1}) = p+1$
- X has full rank: $rk(I_{p+1})$
- condition: $rk(A) = p+1$

$rk(AX) = \min(rk(X), rk(A)) = rk(I_{p+1}) = p+1$

$\rightarrow$

Proof

Let the matrix without loss of generality be of the form

$$A = (X^T X)^{-1} X^T + B$$

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

$$I_{p+1} = AX = \underbrace{(X^T X)^{-1} X^T X}_{Id} + BX = I_{p+1} + \underbrace{BX}_{0}$$

$$\rightsquigarrow BX = 0$$

Note: $Cov(\hat{\beta}^L) = \sigma^2 A A^T$ (proof too similar to one problem sheet)

$$Cov(\hat{\beta}^L) = \sigma^2 A A^T$$

$$= \sigma^2 \left((X^T X)^{-1} X^T + B \right) \left(X (X^T X)^{-1} + B^T \right)$$

$$= \sigma^2 \Big(\underbrace{(X^T X)^{-1} X^T X}_{Id} (X^T X)^{-1} + (X^T X)^{-1} \underbrace{X^T B^T}_{=(BX)^T} + \underbrace{BX}_{=0} (X^T X)^{-1} + BB^T \Big)$$

$$= \sigma^2 (X^T X)^{-1} + \sigma^2 BB^T = Cov(\hat{\beta}) + \sigma^2 BB^T$$

Proof

We know BB^T is positive semi-definite

$$\mathrm{Cov}(\hat{\beta}^L) - \mathrm{Cov}(\hat{\beta}) = \sigma^2 BB^T \geq 0$$

$$c^T \mathrm{Cov}(\hat{\beta}^L) c - c^T \mathrm{Cov}(\hat{\beta}) c \geq 0 \quad \forall c \in \mathbb{R}^{p+1}$$

$$\Rightarrow \mathrm{Var}(c^T \hat{\beta}^L) \geq \mathrm{Var}(c^T \hat{\beta})$$

$$\mathrm{Var}(c^T \hat{\beta}^L) = c^T \mathrm{Cov}(\hat{\beta}^L) c \quad \text{and} \quad \mathrm{Var}(c^T \hat{\beta}) = c^T \mathrm{Cov}(\hat{\beta}) c$$

As c is arbitrary we can choose for each $j = 0, \dots, p$

$$c = (0, \dots, 1, \dots, 0)^T$$

$$\mathrm{Var}(\hat{\beta}_j^L) = \mathrm{Var}(c^T \beta^L) \geq \mathrm{Var}(c^T \hat{\beta}) = \mathrm{Var}(\hat{\beta}_j)$$

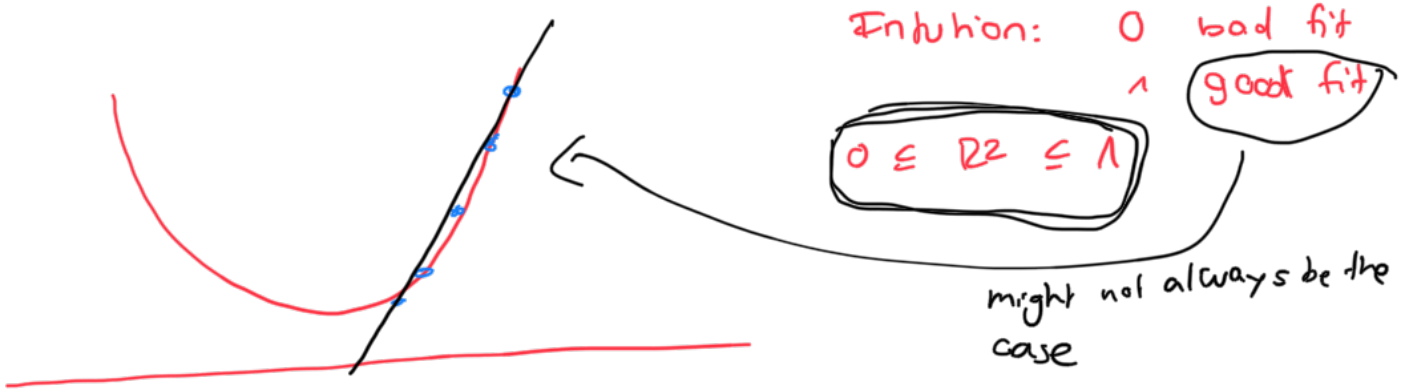
□

Coefficient of determination

Def: The coefficient of determination is defined by

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad \bar{\hat{y}} = \bar{y} \tag{5}$$

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.,

$$\sum_{i=1}^{n} \hat{\epsilon}_i = 0 \quad \text{and} \quad \bar{\hat{\epsilon}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i = 0 \tag{6}$$

(iv) • The mean of the estimated values

$$\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i = \bar{y} \tag{7}$$

Proof

Proof of (i) will be using $H = X(X^TX)^{-1}X^T$, $\hat{y} = X\hat{\beta}$, $\hat{\varepsilon} = y - X\hat{\beta} = y - \hat{y}$

$$\hat{y}^T\hat{\varepsilon} = (X(X^TX)^{-1}X^Ty)^T \cdot (y - X\hat{\beta})$$

($\hat{\beta} = (X^TX)^{-1}X^Ty$)

$$= y^T X (X^TX)^{-1} X^T \cdot (y - Hy)$$

($H = X(X^TX)^{-1}X^T$)

$$= y^T H (Id - H) y = y^T H y - y^T H H y$$

($H = H$)

$$= 0 \quad \square$$

$\hat{y} = Y$

Proof of (ii) $X^T \cdot \hat{\varepsilon} = X^T(y - \hat{y}) = X^Ty - X^THy$

$$= X^Ty - X^TX(X^TX)^{-1}X^Ty$$

($X^TX(X^TX)^{-1} = Id$)

$$= X^Ty - X^Ty = 0$$

Proof

Proof (iii) Remind: $X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}$

$\rightsquigarrow (x^0) = (1, \ldots, 1)^T$

$0 = (x^0)^T \hat{\varepsilon} = \mathbf{1}^T \hat{\varepsilon} = \sum_{i=1}^{n} \hat{\varepsilon}_i$ □

↑ true (ii)

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

$$\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} (y_i - \hat{\varepsilon}_i) = \sum_{i=1}^{n} y_i - \underbrace{\sum_{i=1}^{n} \varepsilon_i}_{=0} = \sum_{i=1}^{n} y_i$$

$\frac{1}{n}$ $\hat{y}_i = X_i \hat{\beta}_i$ $\quad \bar{\hat{y}} = \bar{y}$

$y_i = X_i \hat{\beta}_i - \hat{\varepsilon}_i$ □

Coefficient of determination

Lemma: The following decomposition holds:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}\hat{\epsilon}_i^2 \tag{8}$$

Proof

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

$$C = I_n - \frac{1}{n} 1 1^T$$

$$= I_n - \frac{1}{n} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}_{n \times 1} \underbrace{(1 \dots 1)}_{1 \times n}$$

Note that: • $C \cdot C = \left(I_n - \frac{1}{n} 1 1^T\right)\left(I_n - \frac{1}{n} 1 1^T\right)$

$$= I_n - 2\frac{1}{n} I_n 1 1^T + \frac{1}{n^2} 1 \underbrace{1^T 1}_{1 \times n \cdot n \times 1 \in \mathbb{R}} 1^T$$

$$= \left(I_n - \frac{1}{n} 1 1^T\right) = C$$

(the $1^T 1$ term equals n)

• C is symmetric

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

$$a^T C a = \sum_{i=1}^{n} (a_i - \bar{a})^2 \quad (1)$$

$1 \times n \quad n \times n \quad n \times 1 \quad \in \mathbb{R}$

$y = \hat{y} + \hat{\varepsilon} \mid \cdot C$

$Cy = C\hat{y} + C\hat{\varepsilon}$

$C\hat{\varepsilon} = \begin{pmatrix} \hat{\varepsilon}_1 - \bar{\hat{\varepsilon}} \\ \vdots \\ \hat{\varepsilon}_n - \bar{\hat{\varepsilon}} \end{pmatrix}$ $\quad \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$

$\Rightarrow Cy = C\hat{y} + \hat{\varepsilon} \mid^T$

$y^T C = \hat{y}^T C + \hat{\varepsilon}^T$

Proof

Using that result we obtain:

$$y^T C C y = (\hat{y}^T C + \hat{\varepsilon}^T)(C\hat{y} + \hat{\varepsilon})$$

$$= \hat{y}^T C C \hat{y} + \hat{y}^T \cdot C\hat{\varepsilon} + \hat{\varepsilon}^T C \hat{y} + \hat{\varepsilon}^T \hat{\varepsilon}$$

$C\hat{c} = \hat{\varepsilon}$

$CC = C$

$$= \hat{y}^T C \hat{y} + \underbrace{\hat{y}^T \hat{\varepsilon}}_{=0} + \underbrace{\hat{\varepsilon}^T \hat{y}}_{=0} + \hat{\varepsilon}^T \hat{\varepsilon}$$

- orthogonality follows from previous Lemma part (i)

① and $CC = C$

$$\rightarrow y^T C C y = \underbrace{y^T}_{a} C \underbrace{y}_{a} = \sum_{i=1}^{n} (y - \bar{y})^2$$

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

$$\hat{y}^T C \hat{y} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

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

□

Proof

$Ax = \lambda x$

$rk(A \cdot B) = \min(rk(A), rk(B))$

$A - Id(\lambda)$

$+\lambda$

$\begin{pmatrix} \lambda & & ? \\ & \ddots & \\ & & \lambda \end{pmatrix}$

$\begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_2 \end{pmatrix}$

Coefficient of determination

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

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{\hat{\beta}^\top \mathbf{X}^\top \mathbf{y} - n\bar{y}^2}{\mathbf{y}^\top \mathbf{y} - n\bar{y}^2} \tag{9}$$

Proof

Proof of second equality : Consider nommator and denominator separately

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n} \hat{y}_i \hat{y}_i - 2\sum_{i=1}^{n} \hat{y}_i \bar{y} + \sum_{i=1}^{n} \bar{y}^2 = \hat{y}^T \hat{y} - n\bar{y}^2$$

$\hat{y}^T \hat{y}$

$n\bar{y}^2$

$-2\bar{y} \sum_{i=1}^{n} \hat{y}_i$

$n \bar{\hat{y}}$

$\bar{y}$

$-2\bar{y}^2 n$

$$= \hat{\beta} X^T X \hat{\beta} - n\bar{y}^2$$

$$= \hat{\beta} \underbrace{X^T X (X^T X)^{-1}}_{Id} X^T y - n\bar{y}^2$$

$$= \hat{\beta} X^T y - n\bar{y}^2$$

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = y^T y - n\bar{y}^2$$

□

Proof

Coefficient of determination

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

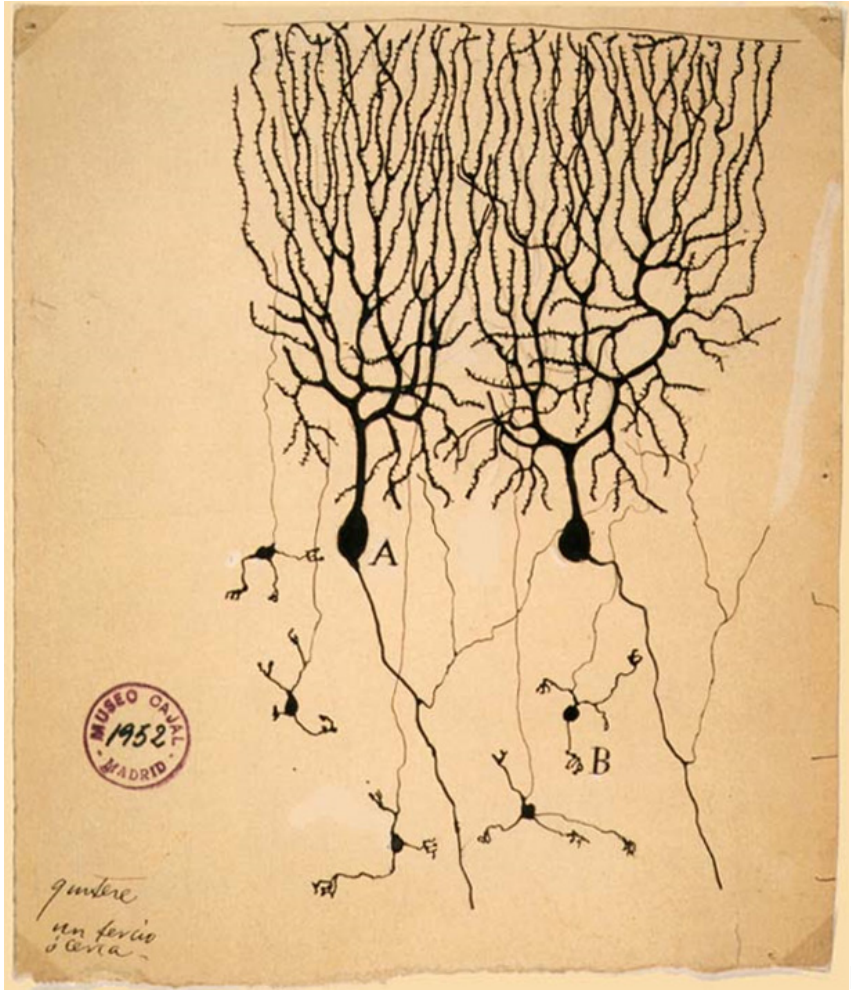
$$\bar{R}^2 = 1 - \left(\frac{n-1}{n-p-1}\right)(1-R^2) \tag{10}$$

Remark:
- $0 \leq R^2 \leq 1$
- R^2 increases with p

Connection to Neural Networks

Neural Networks

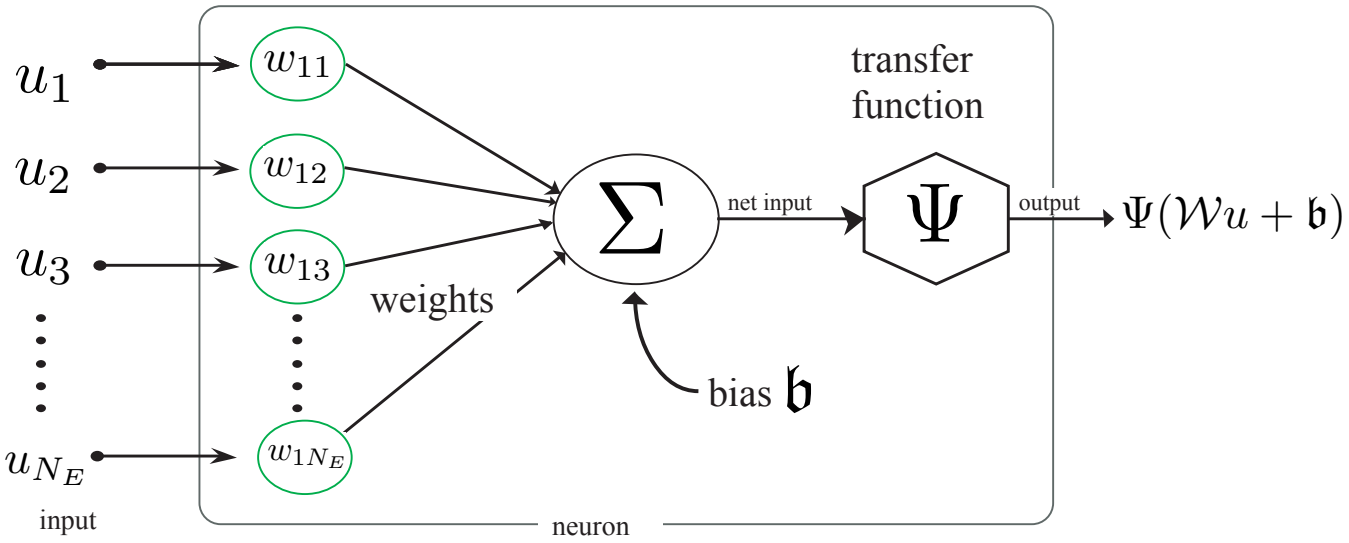
Motivation from biology



By Santiago Ramn y Cajal in 1899 see

https://de.wikipedia.org/wiki/Santiago_Ramn_y_Cajalfordetails

Neuron


u_1
u_2
u_3
u_{N_E}
input
w_{11}
w_{12}
w_{13}
w_{1N_E}
weights
Σ
net input
transfer function
Ψ
output
$\Psi(\mathcal{W}u + \mathfrak{b})$
bias $\mathfrak{b}$
neuron