Statistical Data Analysis

Dr. Jana de Wiljes

1. Dezember 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 $\hat{\beta}^L$

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

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

Proof

Proof

Coefficient of determination

Def: The coefficient of determination is defined by

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

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:

- The fitted values $\hat{\mathbf{y}}$ are orthogonal to the residuals $\hat{\epsilon}$, i.e., $\hat{\mathbf{y}}^\top \hat{\epsilon} = 0$.
- The columns of $\mathbf{X}$ are orthogonal to the residuals $\hat{\epsilon}$, i.e., $\mathbf{X}^\top \hat{\epsilon} = 0$
- The residuals are zero on average, i.e.,

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

- 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

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

Proof

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

Coefficient of determination

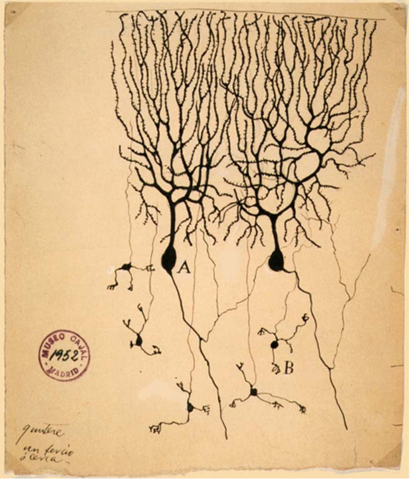
Def: The corrected coefficient of determination $\bar{R}^2$ is defined by

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

Connection to Neural Networks

Neural Networks

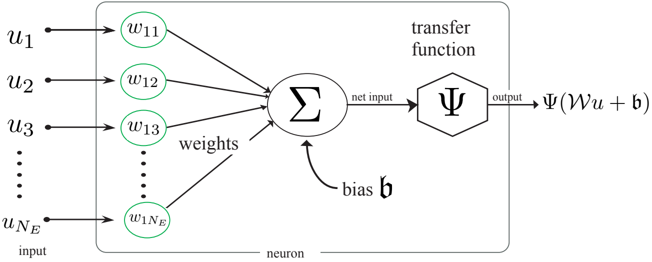
Motivation from biology



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