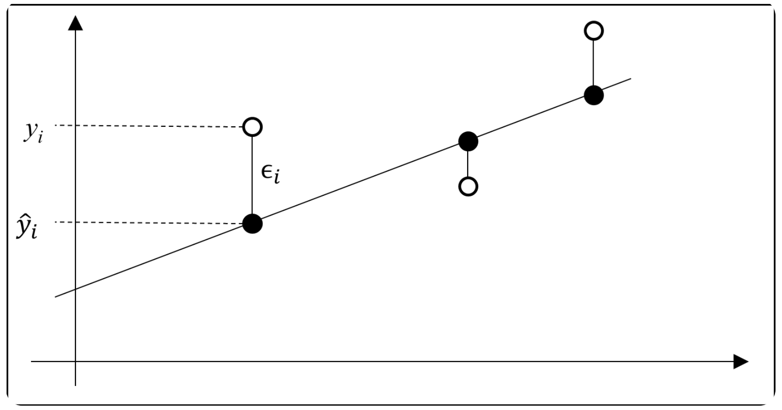
Lineare Regression

Definition

Gegeben sind Datenpunkte $(x_i; y_i)$ mit $1 \leq i \leq n$. Die *Residuen* oder *Fehler* $\epsilon_i = y_i - \underbrace{g(x_i)}_{=\hat{y}_i}$ dieser Datenpunkte sind die Abstände in y-Richtung zwischen den *beobachteten Werten* y_i und den durch die lineare Regression *prognostizierten Werten* $\hat{y}_i = g(x_i)$. Die *Ausgleichs-* oder *Regressionsgerade* $g(x) = m \cdot x + d$ (in y-Richtung) ist diejenige Gerade, für die die Summe der quadrierten Residuen $\sum_{i=1}^{n} {\epsilon_i}^2$ am kleinsten ist.



Zusammenfassung Regressionsgerade:

Die Regressionsgerade $g(x) = m \cdot x + d$ *mit den Parametern* m *und* d *ist die Gerade, für die die Residualvarianz* $\tilde{s}_\epsilon^2$ *minimal ist.*

Die Regressionsgerade hat die Steigung $m = \frac{\tilde{s}_{xy}}{\tilde{s}_x^2}$

und den y *-Achsenabschnitt* $d = \overline{y} - m \cdot \overline{x}$

Für die zugehörige (minimale) Residualvarianz gilt: $\tilde{s}_\epsilon^2 = \tilde{s}_y^2 - \underbrace{\frac{\tilde{s}_{xy}^2}{\tilde{s}_x^2}}_{=\tilde{s}_{\hat{y}}^2}$

mit:

Varianz der x_i-Werte: $\tilde{s}_x^2$

Varianz. der y_i -Werte (Totale Varianz) : $\tilde{s}_y^2$

Varianz der $\hat{y}_i$ -Werte (Prognostizierte Varianz): $\tilde{s}_{\hat{y}}^2$

Kovarianz: $\tilde{s}_{xy}$

arithmetische Mittelwerte $\overline{x}$ *und* $\overline{y}$

Hinweis*: Die Berechnungen können alternativ auch mit den korrigierten Varianzen* s_ϵ^2 , s_x^2, s_y^2, $s_{\hat{y}}^2$ *und der korrigierten Kovarianz* s_{xy} *erfolgen!*

Zusammenfassung Bestimmtheitsmass:

Die Totale Varianz setzt sich zusammen aus der Residualvarianz und der Varianz der prognostizierten Werte (*Varianzzerlegung*):

$$\tilde{s}_y^2 = \tilde{s}_\epsilon^2 + \underbrace{\tilde{s}_{\hat{y}}^2}_{=\frac{\tilde{s}_{xy}^2}{\tilde{s}_x^2}}$$

Das *Bestimmtheitsmass* R^2 beurteilt die globale Anpassungsgüte einer Regression über den Anteil der prognostizierten (erklärten) Varianz $\tilde{s}_{\hat{y}}^2$ an der totalen Varianz $\tilde{s}_y^2$:

$R^2 = \frac{\tilde{s}_{\hat{y}}^2}{\tilde{s}_y^2}$ bzw. $R^2 = \frac{s_{\hat{y}}^2}{s_y^2}$

Das *Bestimmtheitsmass* R^2 stimmt überein mit dem Quadrat *des Korrelations-koeffizienten (nach Bravais-Pearson):*

$R^2 = \frac{\tilde{s}_{xy}^2}{\tilde{s}_x^2 \cdot \tilde{s}_y^2} = r_{xy}^2$ bzw. $R^2 = \frac{s_{xy}^2}{s_x^2 \cdot s_y^2} = r_{xy}^2$

Bestimmung der Regressionsgerade als Matrizenproblem:

Die Parameter m und d werden mit der Matrix $A = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}$ aus den folgenden Normalengleichungen berechnet:

$$A^T \cdot A \cdot \binom{m}{q} = A^T \cdot \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Die Methode der kleinsten Quadrate (KQM)

Das lineare, inhomogene Gleichungssystem mit m Gleichungen und n Unbekannten $x_1, \ldots, x_n$

$$\left.\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \ldots & + & a_{1n}x_n & = & y_1 + \epsilon_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \ldots & + & a_{2n}x_n & = & y_2 + \epsilon_2 \\ \vdots & & \vdots & & & & \vdots & & \vdots \\ a_{m1}x_1 & + & a_{m2}x_2 & + & \ldots & + & a_{mn}x_n & = & y_m + \epsilon_m \end{array}\right\} \Leftrightarrow \underbrace{\begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix}}_{=A} \cdot \underbrace{\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}}_{=\vec{x}} = \underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}}_{=\vec{y}} + \underbrace{\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_m \end{pmatrix}}_{=\vec{\epsilon}}$$

hat immer (mind.) eine Lösung mit einem *Residuenvektor* $\vec{\epsilon}$ von minimalem Betrag. Diese Lösungen sind Lösungen der *Normalengleichungen* $A^T \cdot A \cdot \vec{x} = A^T \cdot \vec{y}$. Hat die Matrix A den Rang n , so ist die symmetrische $n \times n$ Matrix $A^T \cdot A$ invertierbar und die einzige Lösung erhält man aus der Gleichung $\vec{x} = (A^T \cdot A)^{-1} \cdot A^T \vec{y}$.

Für die Lösungen $\vec{x}$ gilt immer:

- $A \cdot \vec{x}$ und der Residuenvektor $\vec{\epsilon} = \vec{y} - A \cdot \vec{x}$ sind orthogonal zueinander.
- Es gilt der Satz von Pythagoras (Quadratsummenzerlegung):

$$|\vec{y}|^2 = |A \cdot \vec{x}|^2 + \underbrace{|\vec{y} - A \cdot \vec{x}|^2}_{=|\vec{\epsilon}|^2}$$

Zuletzt geändert: Montag, 11. Dezember 2023, 15:51

◄ 5. Zusammenfassung: Spezielle Verteilungen

Direkt zu:

7. und 8. Zusammenfassung: Schliessende Statistik ►

?