Formelsammlung

Multivariate Statistik

1. Korrelationsanalyse

Korrelationskoeffizient (Bravais-Pearson)

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{n} y_i^2}} = \frac{s_{xy}}{\sqrt{s_x \cdot s_y}}, \qquad -1 \le r \le +1$$

Statistischer Test

Soll statistisch getestet werden, ob ein linearer Zusammenhang zwischen den Variablen x und y für die Grundgesamtheit besteht, also die Hypothese geprüft werden, ob der unbekannte Korrelationskoeffizient der Grundgesamtheit ρ sich signifikant von Null unterscheidet, so bedarf es spezieller *Annahmen*:

i) Die gemeinsame (bivariate) Verteilung der Variablen ist normalverteilt,
ii) Die vorliegende Stichprobendatei ist per Zufall zustande gekommen.

Unter diesen Annahmen kann der Korrelationskoeffizient nach folgenden Schritten getestet werden:

Schritte des statistischen Tests:

1. Formuliere die Null- und die Alternativhypothese

   $H_0: \rho = 0$ $\qquad$ $H_1: \rho \neq 0$

   (Die Alternativhypothese wird dabei je nach Erwartung über die Richtung des Zusammenhangs unterschiedlich formuliert. Hat man keinerlei Erwartung über die Richtung des Zusammenhangs, so gilt: $H_1: \rho \neq 0$. Es handelt sich dann um einen zweiseitigen Test. Erwartet man, dass die Variablen sich in gleicher Richtung verändern, so wird der positive Zusammenhang mit $H_1: \rho > 0$ formuliert. Bei Erwartung eines negativen Zusammenhangs gilt entsprechend $H_1: \rho < 0$. In diesen Fällen handelt es sich um einseitige Tests.)

2. Berechne die Prüfgröße:

$$t_{prüf} = r\sqrt{\frac{n-2}{1-r^2}}$$

3. Suche $t_{n-2;\alpha}$ in der Tabelle der t-Verteilung.
4. Entscheidung:

   Lehne H_0 ab, falls $|t_{prüf}| > t_{n-2;\alpha}$ gilt; sonst gibt es keinen Grund, H_0 abzulehnen.
   (Der letzte Fall heißt: es liegt ein statistisch gesicherter Zusammenhang zwischen den beiden Merkmalen.)

2. Regressionsanalyse

Einfache lineare Regression

$$y^* = a_0 + a_1 x.$$

Normalgleichungen:

$$\begin{cases} n \cdot a_0 + a_1 \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ a_0 \cdot \sum_{i=1}^{n} x_i + a_1 \cdot \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i \cdot y_i \end{cases}$$

Multiple lineare Regression

$$y^* = a_0 + a_1 x + a_2 x_2 + \ldots + a_n x_n$$

Lösung:

$$a = \left(X^T X \right)^{-1} X^T y$$

mit

$$X^T X = \begin{pmatrix} n & \sum_i x_{i1} & \cdot & \cdot & \cdot & \sum_i x_{im} \\ \sum_i x_{i1} & \sum_i x_{i1}^2 & \cdot & \cdot & \cdot & \sum_i x_{i1} x_{im} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \sum_i x_{im} & \sum_i x_{im} x_{i1} & \cdot & \cdot & \cdot & \sum_i x_{im}^2 \end{pmatrix}; \qquad X^T y = \begin{pmatrix} \sum_i y_i \\ \sum_i x_{i1} y_i \\ \cdot \\ \cdot \\ \cdot \\ \sum_i x_{im} y_i \end{pmatrix}.$$

Speziell *für zwei unabhängige Variablen* löst man folgendes *Normalgleichungssystem*:

$$\begin{cases} n \cdot a_0 + a_1 \sum_{i=1}^{n} x_{i1} + a_2 \sum_{i=1}^{n} x_{i2} = \sum_{i=1}^{n} y_i \\ a_0 \sum_{i=1}^{n} x_{i1} + a_1 \sum_{i=1}^{n} x_{i1}^2 + a_2 \sum_{i=1}^{n} x_{i1} x_{i2} = \sum_{i=1}^{n} x_{i1} y_i \\ a_0 \sum_{i=1}^{n} x_{i2} + a_1 \sum_{i=1}^{n} x_{i1} x_{i2} + a_2 \sum_{i=1}^{n} x_{i2}^2 = \sum_{i=1}^{n} x_{i2} y_i \end{cases}$$

Bestimmtheitsmaß:

$$B_{y;1,2,...,m} = \frac{\sum_{i=1}^{n} (y_i^* - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \qquad 0 \le B_{y;1,2,...,m} \le 1$$

Im Falle einer <u>einfachen</u> linearen Regression besser folgende Formel benutzen:

$$r_{yx} := \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{\sqrt{\left(n \cdot \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} x_i \right) \cdot \left(n \cdot \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} y_i \cdot \sum_{i=1}^{n} y_i \right)}}$$

$$-1 \le r_{yx} \le +1$$

Korrigiertes Bestimmtheitsmaß :

$$B^*_{y;1...m} = 1 - (1 - B_{y;1...m}) \frac{n-1}{n-m-1}$$

Varianz bzw. der Standardfehler der Residuen

$$s_u^2 = \frac{\sum_{i=1}^{n} u_i^{*2}}{n-(m+1)} == \frac{1}{n-(m+1)} u^{*T} u$$

Varianz bzw. der Standardfehler der Regressionsparameter

$$s_{a_i} = s_u \sqrt{x^{(ii)}}, \quad i = 0,1,...,n.$$

Dabei ist x^{ii} das $i-$te Diagonalelement der Matrix $(X^T X)^{-1}$.

Relative Standardfehler der Regressionsparameter

$$s'_{a_i} = \frac{s_{a_i}}{a_i}, \quad i = 0,1,...,n.$$

Konfidenzintervall der Regressionsparameter
Sei

$$Y = A_0 + A_1 x_1 + ... + A_m x_m$$

die Regressionsgleichung der Grundgesamtheit.

Das *Konfidenzintervall* für den Regressionsparameter $A_i,\ i = 1,2,...,m,$ mit dem Signifikanzniveau α wird folgendermaßen bestimmt:

$$\left[a_i - t_{n-m-1} \cdot s_{a_i},\ a_i + t_{n-m-1} \cdot s_{a_i} \right],\ i = 0,1,...,m.$$

Test der Positivität von a_1
Wir beschränken uns auf den Fall, dass die Standardabweichung des Störfaktors in der Grundgesamtheit σ_u nicht bekannt ist.
Dann sind folgende Schritte durchzuführen:

Schritt 1:
Formuliere der Nullhypothese H_0 und der Alternativhypothese H_1.

Schritt 2:
Wende die $t-$Verteilung an.

Schritt 3:
Bestimme die Annahme und die Ablehnungsbereiche auf Grund des Freiheitsgrades $df = n - m - 1$ und α. Die kritische Grenze sei mit $t_{\alpha,n-m-1}$ bezeichnet.

Schritt 4:
Berechne

$$s_{stat} = \frac{a_1 - A_1}{s_{a_1}}.$$

Schritt 5:
Entscheide über die Annahme oder der Ablehnung der Nullhypothese durch Verglich von $t_{\alpha,n-m-1}$ und s_{stat}.

Nichtlineare multiple Regression

Zur Bestimmung der Regressionsgleichung der Cobb-Douglas-Funktion:

$$y^* = \gamma A^{\alpha} K^{\beta},\ \alpha, \beta, \gamma > 0$$

lauten die *Normalgleichungen:*

$$\begin{cases} n \cdot \lg \gamma + \alpha \sum_{i=1}^{n} \lg A_i + \beta \sum_{i=1}^{n} \lg K_i = \sum_{i=1}^{n} \lg y_i \\ \lg \gamma \sum_{i=1}^{n} \lg A_i + \alpha \sum_{i=1}^{n} (\lg A_i)^2 + \beta \sum_{i=1}^{n} \lg A_i \cdot \lg K_i = \sum_{i=1}^{n} \lg A_i \cdot \lg y_i \\ \lg \gamma \sum_{i=1}^{n} \lg K_i + \alpha \sum_{i=1}^{n} \lg A_i \cdot \lg K_i + \beta \sum_{i=1}^{n} (\lg K_i)^2 = \sum_{i=1}^{n} \lg K_i \cdot \lg y_i \end{cases}$$

3. Varianzanalyse

Schritt 1 (Formulierung der Hypothesen)

$H_0: \mu_1 = \mu_2 = ... = \mu_k$; $H_1: \mu_i \neq \mu_j$, für mindestens ein $i \neq j,\ i, j = 1, 2, ..., k,\ s \geq 3.$

Schritt 2 (Berechnung der Teststatistik)

Sei

x_{ij}: Wert der Beobachtung i für die Behandlung j

n_j: Anzahl der Beobachtungen für die Behandlung j

$\bar{x}_j$: Stichprobenmittelwert für die Behandlung j

s_j^2: Stichprobenvarianz für die Behandlung j

s_j: Standardabweichung der Stichprobe für die Behandlung.

$$\boxed{F = \frac{MSTR}{MSE}}$$

mit

$$MSTR := \frac{SSTR}{k-1}$$

(*Within-Tratments*: *Mean Square due to TReatment; Between Treatments: Mean Square due to Error)*

Geschätzte Varianz *innerhalb der Gruppen* (*within-Varianz, Fehlervarianz, „error“*):

Wie unterscheiden sich die einzelnen Werte in einer Stichprobe (oder Gruppe) von den übrigen Werten in der gleichen Gruppe?

Geschätzte Varianz *zwischen den Gruppen* (*between-Varianz*):Wie unterscheiden sich die Mittelwerte verschiedener Stichproben (oder Gruppen) voneinander?)

$$SSTR := \sum_{j=1}^{k} n_j (\bar{x}_j - \bar{\bar{x}})^2 \qquad (\underline{S}um\ of\ \underline{S}quares\ due\ to\ \underline{T}rea\underline{T}ments)$$

$$\bar{x}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j}$$

$$\bar{\bar{x}} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}}{n_T} \quad ; \quad n_T = n_1 + n_2 + ... + n_k . \qquad (Gesamtstichprobenmittelwert)$$

(Haben alle Stichproben den gleichen Umfang n, dann ist $n_T = k \cdot n$ und damit

$$\bar{\bar{x}} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}}{kn} = \frac{\frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}}{n}}{k} = \frac{\sum_{j=1}^{k} \bar{x}_j}{k}.)$$

(Haben alle Stichproben den gleichen Umfang n, dann ist $n_T = k \cdot n$ und damit

$$\bar{\bar{x}} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}}{kn} = \frac{\frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}}{n}}{k} = \frac{\sum_{j=1}^{k} \bar{x}_j}{k}.)$$

$$s_j^2 = \frac{\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}{n_j - 1}$$

$$SSE = \sum_{j=1}^{k} (n_j - 1) s_j^2 \qquad (\underline{S}um\ of\ \underline{S}quares\ due\ to\ \underline{E}rror)$$

$$MSE := \frac{\sum_{j=1}^{k} (n_j - 1) s_j^2}{n_T - k} = \frac{SSE}{n_T - k} \quad (\underline{M}ean\ \underline{S}quare\ due\ to\ \underline{E}rror)$$

Schritt 3 (Entscheidung)

$p-value-$Methode: Lehne H_0 ab, wenn $p-Value \le \alpha$

Critical-Value-Methode: Lehne H_0 ab, wenn $F \ge F_\alpha$

(F_α basiert auf der $F-$Verteilung mit dem Freiheitsgrad $k-1$ im Zähler und dem Freiheitsgrad $n_T - k$ im Nenner.)

Konfidenzintervall für die Mittelwerte in der Grundgesamtheit

$$\mu_j \in \left[\bar{x}_j - t_{\alpha/2} \cdot \frac{\sqrt{MSE}}{n_j},\ \bar{x}_j + t_{\alpha/2} \cdot \frac{\sqrt{MSE}}{n_j} \right], \quad j = 1,2,...,k,$$

(Dabei liegt ein Freiheitsgrad von $n_T - k$ vor.)

Paarweiser Vergleich (Fishers *LSD-Test*)

Schritt 1:
Formulierung der Hypothesen

$$H_0: \mu_i = \mu_j \qquad H_1: \mu_i \ne \mu_j \quad `$$

Schritt 2:
Berechnung der Teststatistik:

$$t = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{MSE \cdot \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}}$$

Schritt 3:
Entscheidung:

$p-value-Methode$: Lehne H_0 ab, wenn $p-Value \le \alpha$

Methode des kritischen Wertes: lehne H_0 ab, wenn $t \le -t_{\alpha/2}$ or $t \ge t_{\alpha/2}$

(Dabei liegt ein Freiheitsgrad von $n_T - k$ vor.)