- Korrelation und Unabhängigkeit
- Lineare Regression
- Weitere Regressionsverfahren
- Zufallszahlen
- Clusteranalyse

6. Multivariate Verfahren

Übersicht (nicht alle werden behandelt)

Korrelation und Unabhängigkeit

Unabhängigkeit und Unkorreliertheit, Wdh.

Die Zufallsvariablen $X_1, \ldots, X_N$ heißen unabhängig, falls für alle $x_1, \ldots, x_N \in \mathbb{R}$

$$P(X_1 < x_1, \ldots, X_N < x_N) = P(X_1 < x_1) \cdots P(X_N < x_N)$$

Die Zufallsvariablen $X_1, \ldots, X_N$ heißen unkorreliert, falls

$$\mathbf{E}(X_1 \cdots X_N) = \mathbf{E}(X_1) \cdots \mathbf{E}(X_N).$$

Unabhängigkeit $\Rightarrow$ Unkorreliertheit

$\not\Leftarrow$

Unabhängigkeit $\Leftrightarrow$ Unkorreliertheit falls $X_i \sim \mathcal{N}$

Korrelation und Unabhängigkeit

Fall a) Stetige (metrische) Merkmale

Seien (X_i, Y_i), $i = 1, \ldots, N$ unabhängige bivariate Zufallsvariablen. Wir testen

H_0 : X und Y sind unabhängig (unkorreliert) gegen

H_1 : X und Y sind linear abhängig (korreliert)

Pearson-Korrelation

$$r_{XY} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 \sum (Y_i - \overline{Y})^2}}$$

$$T = \sqrt{N-2} \cdot \frac{r_{XY}}{\sqrt{1 - r_{XY}^2}} \sim t_{N-2}$$

wird in SAS zur Berechnung der p-Werte verwendet.

Korrelation und Unabhängigkeit

Fall a) Stetige (metrische) Merkmale

H_0 : X und Y sind unabhängig (unkorreliert) gegen

H_1 : X und Y sind monoton abhängig

Spearman-Rangkorrelationskoeffizient

$$r_S = \frac{\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_i (R_i - \overline{R})^2 \sum_i (S_i - \overline{S})^2}}$$

Weitere Korrelationskoeffizienten: Kendall.
Wenn keine Normalverteilung vorliegt, so Spearman oder Kendall nehmen!

Korrelation und Unabhängigkeit

a) Metrisch skalierte Merkmale

```
PROC CORR PEARSON SPEARMAN KENDALL;
   VAR vars;
RUN;
```

b) Ordinal oder nominal skalierte Merkmale

```
PROC FREQ;
   TABLES var1*var2 / CHISQ;
RUN;
```

`Descr_Scatter.sas`
`Descr_Scatter_Heroin.sas`

Korrelation und Unabhängigkeit

Ordinal oder nominal skalierte Merkmale

Frage: Bestehen Abhängigkeiten?

Geschlecht - Studienfach
Studiengang - Note
Geburtsmonat - IQ

Antwort: χ^2 - Unabhängigkeitstest (Pearson, 1908)

Annahme:
X hat Ausprägungen $a_1, \ldots, a_m$
Y hat Ausprägungen $b_1, \ldots, b_l$

(sind die Daten metrisch, so wird automatisch eine Klasseneinteilung vorgenommen.)

$$P(X = a_i) = p_{i.} \qquad P(Y = b_j) = p_{.j}$$

$$P(X = a_i, Y = b_j) = p_{ij}$$

Unabhängigkeitstests

Häufigkeitstabelle (= Kontingenztafel)

$X \mid Y$	b_1	b_2	$\cdots$	b_j	$\cdots$	b_l	
a_1	h_{11}	h_{12}	$\cdots$	h_{1j}	$\cdots$	h_{1l}	$h_{1.}$
a_2	h_{21}	h_{22}	$\cdots$	h_{2j}	$\cdots$	h_{2l}	$h_{2.}$
$\cdots$							
a_i	h_{i1}	h_{i2}	$\cdots$	h_{ij}	$\cdots$	h_{iN}	$h_{i.}$
$\cdots$							
a_m	h_{m1}	h_{m2}	$\cdots$	h_{mj}	$\cdots$	h_{ml}	$h_{m.}$
	$h_{.1}$	$h_{.2}$	$\cdots$	$h_{.j}$	$\cdots$	$h_{.l}$	$h_{..}$=N

h_{ij}: Häufigkeiten

Unabhängigkeitstests

Die Häufigkeiten h_{ij} werden verglichen mit den theoretischen Häufigkeiten np_{ij}.

$$\begin{aligned} H_0: \quad & p_{ij} = p_{i.} \cdot p_{.j}, \quad i = 1, ..., m, j = 1, ...l \\ H_1: \quad & p_{ij} \neq p_{i.} \cdot p_{.j}, \quad \text{für ein Paar} (i, j) \end{aligned}$$

H_0: X und Y sind unabhängig.
H_1: X und Y sind abhängig.

Betrachten zunächst die Stichprobenfunktion

$$\widetilde{T} = \sum_i \sum_j \frac{(h_{ij} - np_{ij})^2}{np_{ij}}$$

Unabhängigkeitstests

Konstruktion der Teststatistik

Problem: $p_{i.}$ und $p_{.j}$ sind unbekannt. Sie müssen also geschätzt werden,
das sind $m + l - 2$ Parameter ($\sum p_{i.} = \sum p_{.j} = 1$)

$$\hat{p}_{i.} = \frac{h_{i.}}{N} \qquad \hat{p}_{.j} = \frac{h_{.j}}{N}$$

$$h_{i.} = \sum_{j=1}^{l} h_{ij} \qquad h_{.j} = \sum_{i=1}^{m} h_{ij}$$

Unabhängigkeitstests

Einsetzen der Schätzungen in $\widetilde{T}$ (unter H_0)

$$\begin{aligned} Q_P &= \sum_i \sum_j \frac{(h_{ij} - n\hat{p}_{i.}\hat{p}_{.j})^2}{n\hat{p}_{i.}\hat{p}_{.j}} \\ &= n \sum_i \sum_j \frac{(h_{ij} - \frac{h_{i.}h_{.j}}{n})^2}{h_{i.}h_{.j}} \\ &\sim \chi^2_{(m-1)(l-1)} \quad \text{approx. unter} H_0 \end{aligned}$$

Die Anzahl der Freiheitsgrade ergibt sich aus:

$$m \cdot l - 1 - \underbrace{(m + l - 2)}_{\#\text{geschätzte Werte}}$$

H_0 ablehnen, falls

$$Q_P > \chi^2_{(m-1)(l-1)}, \quad \text{bzw. falls p-Wert} < \alpha$$

Korrelation und Unabhängigkeit

Faustregel für die Anwendung des χ^2-Unabhängigkeitstests:

- alle $h_{ij} > 0$.
- $h_{ij} \geq 5$ für mindestens 80% der Zellen,

sonst Klassen zusammenfassen.

`Descr_Freq_Heroin_Unabhaengigkeitstest`

Korrelation und Unabhängigkeit

Weitere Unabhängigkeitstests (1)

- LQ-χ^2- Unabhängigkeitstest

$$G^2 = 2 \sum_i \sum_j h_{ij} \ln \frac{h_{ij}}{h_{i.} h_{.j}} \sim \chi^2_{(m-1)(l-1)}$$

- Continuity Adjusted χ^2 (bei SAS nur: 2x2-Tafel)

$$Q_c = N \sum_i \sum_j \frac{\max(0, |h_{ij} - \frac{h_{i.} h_{.j}}{N}| - 0.5)^2}{h_{i.} h_{.j}} \sim \chi^2_{(m-1)(l-1)}$$

- Mantel-Haenszel (r_{XY}: Pearson-Korrelation)

$$Q_{MH} = (N-1) r^2_{XY} \sim \chi^2_1$$

- Phi-Koeffizient

$$\Phi = \begin{cases} \frac{h_{11} h_{22} - h_{12} h_{21}}{\sqrt{h_{1.} h_{2.} h_{.1} h_{.2}}} & m = l = 2 \\ \sqrt{Q_p / n} & sonst \end{cases}$$

Weitere Unabhängigkeitstests (2)

- Kontingenzkoeffizient

$$P = \sqrt{\frac{Q_P}{Q_P + N}}$$

- Fishers Exact Test (bei 2x2-Tafeln)

durch Auszählen aller Tafel-Möglichkeiten bei gegebenen Rändern.
(gilt als etwas konservativ.)

- Cramers V

$$V = \begin{cases} \Phi & \text{falls} \quad 2\text{x}2 \text{ Tafel} \\ \sqrt{\frac{Q_P/N}{\min(m-1, l-1)}} & \text{sonst} \end{cases}$$

Weitere Unabhängigkeitstests (2)

Anmerkungen

- Mantel- Haenszel Test verlangt ordinale Skalierung, vgl. $(N-1)r_{XY}^2$

‘gut’ gegen lineare Abhängigkeit.

- Der χ^2 Unabhängigkeitstest testet gegen allg. Unabhängigkeit.
- Der LQ-Test G^2 ist plausibel und geeignet.
- Der LQ-Test G^2 und der χ^2 Unabhängigkeitstest sind asymptotisch äquivalent.

Unabhängigkeitstests

Φ-Koeffizient (2x2 Tafel)

Y X	Sportler	Nichtsportler	Summe
w	p_{11}	p_{12}	$p_{1.}$
m	p_{21}	p_{22}	$p_{2.}$
Summe	$p_{.1}$	$p_{.2}$	1

$$X \sim Bi(1, p_{.2}) \qquad Y \sim Bi(1, p_{2.})$$

$$\begin{aligned} \mathbf{E}(X) &= p_{.2} \quad var(X) = p_{.2}(1 - p_{.2}) = p_{.2}p_{.1} \\ \mathbf{E}(Y) &= p_{2.} \quad var(Y) = p_{2.}(1 - p_{2.}) = p_{2.}p_{1.} \end{aligned}$$

$$cov(X, Y) = \mathbf{E}(X \cdot Y) - \mathbf{E}(X)\mathbf{E}(Y) = p_{22} - p_{.2}p_{2.}$$

Unabhängigkeitstests

Korrelationskoeffizient in einer 2x2 Tafel

$$\rho = \frac{p_{22} - p_{.2}p_{2.}}{\sqrt{p_{.2}p_{1.}p_{2.}p_{.1}}} = \frac{p_{11}p_{22} - p_{12}p_{21}}{\sqrt{p_{.2}p_{2.}p_{1.}p_{.1}}}$$

$$\begin{aligned} p_{22} - p_{2.}p_{.2} &= p_{22} - (p_{21} + p_{22})(p_{12} + p_{22}) \\ &= p_{22} - (p_{21}p_{12} + p_{22}p_{12} + p_{21}p_{22} + p_{22}^2) \\ &= p_{22}(1 - p_{12} - p_{21} - p_{22}) - p_{21}p_{12} \\ &= p_{22}p_{11} - p_{21}p_{12} \end{aligned}$$

Für $m = l = 2$ ist der Phi-Koeffizient eine Schätzung des Korrelationskoeffizienten.

6.2 Lineare Regression

Einfache lineare Regression (vgl. Kap. 4.7)

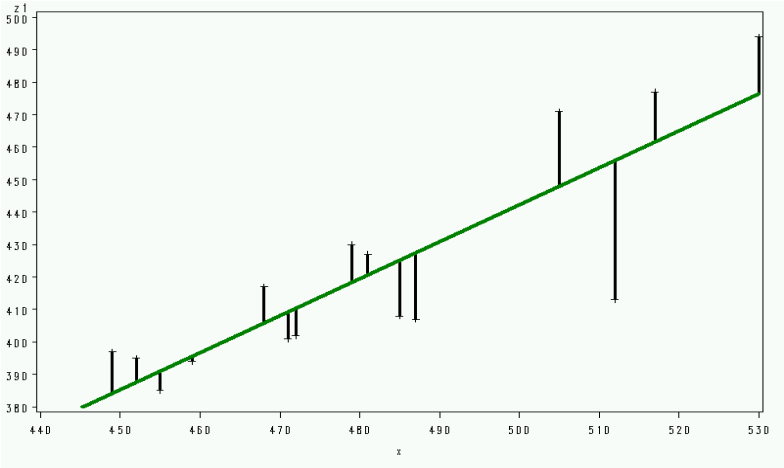
$$Y_i = \theta_0 + \theta_1 X_i + \epsilon_i \qquad \epsilon_i \sim (0, \sigma^2)$$

$$\begin{aligned} \hat{\theta}_1 &= \frac{S_{XY}}{S_X^2} \\ \hat{\theta}_0 &= \frac{1}{n}\left(\sum Y_i - \hat{\theta}_1 \sum X_i\right) \end{aligned}$$

als Lösung der Minimumaufgabe

$$\sum_{i=1}^{n} (Y_i - \theta_1 X_1 - \theta_0)^2 \rightarrow \text{ min.}$$

Lineare Regression



Die Summe der Quadrate der Länge der Streckenabschnitte soll minimal werden.

$$S_{XY} = \frac{1}{n-1}\sum_i (X_i - \overline{X})(Y_i - \overline{Y})$$

$$S_X^2 = \frac{1}{n-1}\sum_i (X_i - \overline{X})^2$$

`Regression_Venusmuscheln`
`Regression_Plot`

Lineare Regression

Multiple lineare Regression

Modell

$$\begin{aligned} Y_i &= \theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i} + \ldots + \theta_m x_{mi} + \epsilon_i \\ Y_i &= \theta_0 + \theta_1 X_{1i} + \theta_2 X_{2i} + \ldots + \theta_m X_{mi} + \epsilon_i \end{aligned}$$

Y_i, ϵ_i Zufallsvariablen, unabh., $\epsilon_i \sim (0, \sigma^2)$, $i = 1 \ldots n$
$\theta_0 \ldots \theta_m, \sigma$: Modellparameter $\Rightarrow$ zu schätzen

Man unterscheidet Fälle:
$x_i = (x_{1i}, \ldots, x_{mi})$ fest, und $X_i = (X_{1i}, \ldots, X_{mi})$ zufällig
oder auch gemischt.
Matrix-Schreibweise:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$$

Lineare Regression

Multiple lineare Regression

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$$

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \dots \\ Y_n \end{pmatrix} \qquad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \dots \\ \epsilon_n \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & X_{11} & \dots & X_{1m} \\ . & . & \dots & . \\ 1 & X_{n1} & \dots & X_{nm} \end{pmatrix}$$

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_0 \\ \dots \\ \theta_m \end{pmatrix}$$

Methode der kleinsten Quadrate: Bestimme $\hat{\boldsymbol{\theta}}$ so daß

$$(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}}) = \min_{\boldsymbol{\theta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\theta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\theta})$$

Lineare Regression

Multiple lineare Regression

Kleinste Quadrat-Schätzung

Vor.: rg$(\mathbf{X}'\mathbf{X}) = m$ (voll)

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

wenn $(\mathbf{X}'\mathbf{X})$ nicht regulär: verallg. Inverse (Moore-Penrose)

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}$$

Lineare Regression

Multiple lineare Regression

Kleinste Quadrat-Schätzung, Spezialfall $m = 1$ (1)

$$
\begin{aligned}
(\mathbf{X}'\mathbf{X})^{-1} &= \left[\begin{pmatrix} 1 & 1 & \dots & 1 \\ X_{11} & . & \dots & X_{n1} \end{pmatrix} \begin{pmatrix} 1 & X_{11} \\ . & \dots \\ 1 & X_{n1} \end{pmatrix} \right]^{-1} \\
&= \begin{pmatrix} n & \sum_i X_i \\ \sum_i X_i & \sum_i X_i^2 \end{pmatrix}^{-1} \qquad (X_i = X_{1i}) \\
&= \frac{1}{n \sum X_i^2 - (\sum X_i)^2} \begin{pmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & n \end{pmatrix}
\end{aligned}
$$

Lineare Regression

Multiple lineare Regression

Kleinste Quadrat-Schätzung, Spezialfall $m = 1$ (2)

$$
\begin{aligned}
\mathbf{X}'\mathbf{Y} &= \begin{pmatrix} 1 & 1 & \ldots & 1 \\ X_1 & . & \ldots & X_n \end{pmatrix} \cdot \begin{pmatrix} Y_1 \\ \ldots \\ Y_n \end{pmatrix} = \begin{pmatrix} \sum Y_i \\ \sum X_i Y_i \end{pmatrix} \\
\hat{\boldsymbol{\theta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\
&= \frac{1}{n\sum X_i^2 - (\sum X_i)^2} \begin{pmatrix} \sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i \\ -\sum X_i \sum Y_i + n \sum X_i Y_i \end{pmatrix}
\end{aligned}
$$

Lineare Regression

Multiple lineare Regression

Schätzung für $\mathbf{Y}$: $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\theta}}$

Vergleiche mit $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$

Einsetzen von $\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$:

$$\begin{aligned} \hat{\mathbf{Y}} &= \underbrace{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'}_{\mathbf{H}}\,\mathbf{Y} \\ &= \mathbf{H}'\mathbf{Y} \end{aligned}$$

$\mathbf{H}$: Hat-Matrix

Aus dem Beobachtungsvektor $\mathbf{Y}$ wird der geschätzte Beobachtungsvektor $\hat{\mathbf{Y}}$.

Lineare Regression

Multiple Lineare Regression (Fortsetzung)

Quadratsummenaufspaltung:

$$\underbrace{\sum (Y_i - \overline{Y})^2}_{SST} = \underbrace{\sum (\hat{Y}_i - \overline{Y})^2}_{SSM} + \underbrace{\sum (Y_i - \hat{Y}_i)^2}_{SSE}$$

$MST = \frac{1}{n-1} SST$: Schätzung für σ^2.

$MSE = \frac{1}{n-m-1} SSE = \hat{\sigma}^2$. (e-treu)

$MSM = \frac{1}{m} SSM$ (m+1 Einflussvariablen)

Bestimmtheitsmaß (wie bei der Varianzanalyse)

$$R^2 = \frac{SSM}{SST}.$$

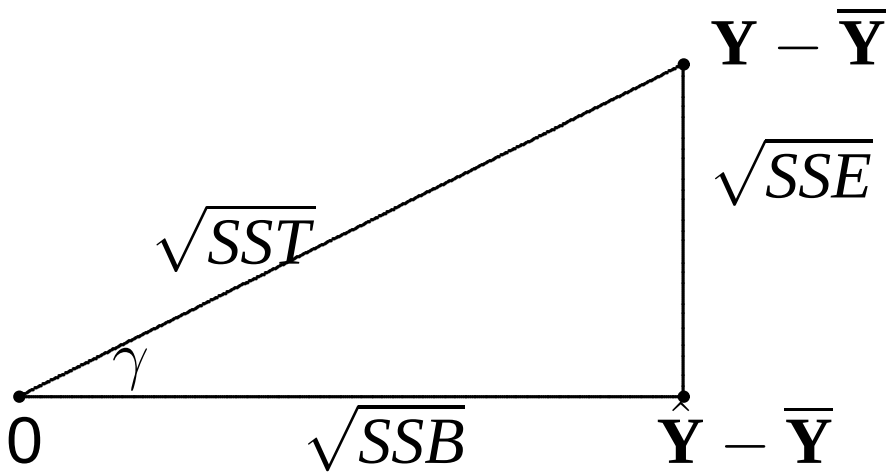
Geometrische Veranschaulichung

zur Multiplen Linearen Regression

$\mathbf{Y} = (Y_{11}, \ldots, Y_{kn_k})$ Dimension N

$\hat{\mathbf{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_1)$

$\overline{\mathbf{Y}} = (\underbrace{\overline{Y}, \ldots, \overline{Y}}_{\text{n mal}})$, $\overline{Y} = \frac{1}{N} \sum_{i,j} Y_{ij}$



$$SSB + SSW = SST \qquad R^2 = cos^2\gamma$$

$$||\hat{\mathbf{Y}} - \overline{Y}||^2 + ||\mathbf{Y} - \hat{\mathbf{Y}}||^2 = ||\mathbf{Y} - \overline{Y}||^2$$

Lineare Regression

Multiple Lineare Regression (Fortsetzung)

$H_0 : \theta_1 = \theta_2 = ... = \theta_m = 0$

Unter der Annahme $\epsilon_i \sim$ Normal

$$F = \frac{SSM}{SSE} \cdot \frac{n-m-1}{m} \sim F_{m,n-m-1}$$

PROC REG; MODEL y = x1 x2 x3 / Optionen; TEST x2=0 x3=0; /*zusaetzl. Hypothesen*/ **RUN**;

```
Regression_Tibetan
Regression_Phosphor
```

Zusätzliche Hypothesen, z.B.:

$H_{0a} : \theta_1 = 0 \; , \quad H_{1a} : \theta_1 \neq 0$

$H_{0b} : \theta_k = 0 \; , \quad H_{1b} : \theta_k \neq 0$

$H_{0c} : \theta_1 = \theta_2 = 0 \; , \quad H_{1b} : \theta_1 \neq 0 \vee \theta_2 \neq 0$

Lineare Regression

Multiple Lineare Regression (Fortsetzung)

R^2-adjustiert für Anzahl p der Parameter im Modell

$$Adj\_R^2 = 1 - \frac{n-i}{n-p}(1-R^2)$$

$$\begin{cases} i = 0 & \text{ohne intercept} \\ i = 1 & \text{mit intercept} \end{cases}$$

Dependent Mean: Mittelwert der abhängigen Variable

StdError MeanPredict: Standardfehler für vorhergesagten Erwartungswert

Lineare Regression

Multiple Lineare Regression (Fortsetzung)

Optionen (Auswahl)

XPX:	Ausgabe der Matrizen $\mathbf{X}'\mathbf{X}, \mathbf{X}'\mathbf{Y}, \mathbf{Y}'\mathbf{Y}$
I:	Ausgabe der Inversen von $\mathbf{X}'\mathbf{X}$
COVB:	Schätzung der Kovarianzmatrix der Schätzung = $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$
CLM, CLI:	Konfidenzbereiche (s.u.)
CLB:	Konfidenzintervall für Parameter θ
R:	studentisierte Residuen (s.u.)
DW:	Durbin-Watson “Test” auf Autokorrelation (s.u.)

Lineare Regression

Multiple Lineare Regression (Fortsetzung)

Output Statistics (Optionen CLI, CLM, R)

Dependent Variable	Y_i
Predicted Value	$\hat{Y}_i = \hat{\boldsymbol{\theta}}\mathbf{X}$
StdErrorMeanPredict	$\hat{\sigma}_{\hat{Y}_i}$
95% CL Mean (s.u.)	nur Variablität in Parameter-schätzung berücksichtigt
95% CL Predict (s.u.)	Variablilität im Fehlerterm mit berücksichtigt
Residual	$e_i = Y_i - \hat{Y}_i$
StdErrorResidual	s.u., $s\sqrt{1 - h_{ii}}$
Student Residual	r_i
Cooks D_i	s.u.
Predicted Residual SS	s.u.

Lineare Regression

Multiple Lineare Regression (Fortsetzung)

Konfidenzintervalle für allg. Parameter ϑ_i:

$$\frac{\hat{\vartheta}_i - \vartheta_i}{S_{\hat{\vartheta}_i}} \sim t_{n-1} \qquad \underline{\text{Vor.}}\ \epsilon_j \sim N(0, \sigma^2) \quad \text{u.a.}$$

KI: $[\hat{\vartheta}_i - t_{1-\frac{\alpha}{2},n-1} \cdot S_{\hat{\vartheta}_i}, \hat{\vartheta}_i + t_{1-\frac{\alpha}{2},n-1} \cdot S_{\hat{\vartheta}_i}]$

95% Konfidenzintervall für $\underline{E(Y_i)}$

($\vartheta_i = \mathbf{E}(Y_i)$, Option CLM)
Nur die Variabilität in der Parameterschätzung wird berücksichtigt.

Lineare Regression

Multiple Lineare Regression (Fortsetzung)

95% Konfidenzintervall für Vorhersagen $\underline{Y_i}$

($\vartheta_i = Y_i$, Option CLI)
Die Variabilität im Fehlerterm wird mit berücksichtigt.

95% Konfidenzintervall für $\boldsymbol{\theta}$

($\vartheta_i = \theta_j$, Option CLB)

Darstellung von Konfidenzbereichen bei der einfachen Regressionsanalyse

```
SYMBOL I=RLCLI95;
PROC GPLOT;
```

Multiple Lineare Regression (Forts.)

Residualanalyse (1)

Studentisierte Residuen (Option R)

$$r_i = \frac{e_i}{s\sqrt{1 - h_{ii}}}$$

$e_i = y_i - \hat{y}_i$ (Residuen) sind korreliert,
var $e_i = \sigma^2(1 - h_{ii})$ $s = \hat{\sigma}$

Cook's D_i

$$D_i = \frac{(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{(i)})'(\mathbf{X}'\mathbf{X})(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{(i)})}{(m+1)S^2}, \qquad i = 1...n$$

beschreibt den Einfluß der i-ten Beobachtung auf die Parameterschätzung
$\hat{\boldsymbol{\theta}}_{(i)}$: KQS von $\boldsymbol{\theta}$ ohne Beobachtung i.
Faustregel: $D_i > 1 \rightarrow$ 'starker' Einfluß

Multiple Lineare Regression (Forts.)

Residualanalyse (2)

Predicted Residual SS (PRESS)

$$\sum (y_i - \hat{y}_{i(i)})^2$$

$\hat{y}_{i(i)}$: i-te Beobachtung weggelassen.

“Test” auf Autokorrelation: Durbin-Watson-Test (Option DW)

$$DW = \frac{\sum_{i=1}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

DW=2: Unkorreliertheit der Residuen

Multiple Lineare Regression (Forts.)

Residualanalyse (3)

Weitere Bewertung der Residuen

Kommando PLOT in der Prozedur REG

PLOT rstudent.*obs.;
PLOT residual.*y residual.*predicted.;
OUTPUT OUT=dateiname RESIDUAL=;
und evtl. Test auf Normalverteilung.

rstudent.	:	studentisierte Residuen
residual.	:	Residuen
obs	:	Beobachtungsnummer
y	:	beobachteter Wert von Y
predicted.	:	geschätzter Wert von Y: $\hat{Y}$

Lineare Regression

Multiple Lineare Regression (Fortsetzung)

Modellwahl in der linearen Regression

SELECTION=

BACKWARD: Die Variablen mit größten p-Wert werden herausgenommen (min. p-Wert: SLSTAY [=0.1])

FORWARD: Start ohne Variablen, die Var. mit kleinstem p-Wert kommt hinzu (max. p-Wert: SLENTRY[= 0.5])

STEPWISE: Start ohne Variable, 1.Schritt wie bei FORWARD (Standard: SLENTRY = 0.15), Variablen können wieder eliminiert werden (Standard: SLSTAY=0.1)

MAXR: Fr jeweils eine feste Anzahl von Variablen wird das Modell mit max. R^2 ausgegeben.

Werte in [ ] sind Standardwerte

Lineare Regression

Multiple Lineare Regression (Fortsetzung)

a) Wenn $rg(\mathbf{X}'\mathbf{X})$ nicht voll ($< m+1$)

$\Rightarrow (\mathbf{X}'\mathbf{X})^{-}$ und Anmerk. in Output

b) Condition number

$\sqrt{\frac{\lambda max}{\lambda min}}$ $\lambda_{max}, \lambda_{min}$ größter u. kleinster Eigenwert von $\mathbf{X}'\mathbf{X}$ (ohne 1-Spalte).
große Konditionszahl (etwa $>$ 30): schlechte Kondition ($\approx$ lineare Abhängigkeit)

c) C(p): Mallows (1973) Kriterium für die Modellwahl

$$C(p) = \frac{SSE_p}{MSE} - n + 2p$$

SSE_p: SSE im Modell mit p Parametern

Multiple Lineare Regression (Forts.)

Modellwahl in der linearen Regression

$$R^2 = \frac{SSM}{SST}.$$

$$C(p) = \frac{SSE_p}{MSE} - n + 2p$$

SSE_p: SSE im Modell mit p Parametern
Ziel: R^2 groß, $C(p)$ nahe p

Idee von $C(p)$: Wenn die Wahl von p Parametern gut, dann

$$MSE \approx MSE_p = \frac{SSE_p}{n-p} \quad \Rightarrow \quad C(p) \approx n - p - n + 2p = p$$

`Regression_Tibetan_Modellwahl`

Lineare Regression

Multiple Lineare Regression (Fortsetzung)

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$$

Einfache Varianzanalyse als Spezialfall

$$\begin{pmatrix} Y_{11} \\ Y_{21} \\ \dots \\ Y_{n_1 1} \\ Y_{12} \\ \dots \\ Y_{n_2 2} \\ \dots \\ \dots \\ Y_{1k} \\ \dots \\ Y_{n_k k} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ . & . & . & & . \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ . & . & . & & . \\ 1 & . & 1 & \dots & 0 \\ . & . & 0 & \dots & 0 \\ . & . & . & & . \\ 1 & . & . & 0 & 1 \\ . & . & . & & \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \dots \\ \alpha_k \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ \epsilon_{n_k k} \end{pmatrix}$$

Lineare Regression

Multiple Lineare Regression (Fortsetzung)

$$\begin{pmatrix} Y_1 \\ \dots \\ \dots \\ Y_N \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & \dots & X_{1p} \\ . & \dots & & \dots \\ . & \dots & & \dots \\ 1 & X_{N1} & \dots & X_{Np} \end{pmatrix} \begin{pmatrix} \mu \\ \theta_1 \\ \dots \\ \theta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \dots \\ \dots \\ \epsilon_N \end{pmatrix} \Leftrightarrow$$

$$\mathbf{Y} = X\boldsymbol{\theta} + \boldsymbol{\epsilon}$$

Weitere Regressionsverfahren

Mögliche Probleme bei der linearen Regression

Probleme	Lösungsansätze
• Ausreißer	Robuste Lineare Regression
• keine Normalverteilung	Datentransformation, (L_1-Regression)
• kein linearer Zusammenhang	Nichtlineare Regression Nichtparametrische Regr.
• Zielvariable nicht stetig	Logistische Regression

Robuste Lineare Regression (Skizze)

Ausreißer können auftreten in

- Y-Richtung
- X-Richtung(en) (Leverage points)
- Y- und X- Richtungen

Fall: Ausreißer(verdacht) in Y-Richtung:
es werden nicht die Abstandsquadrate minimiert, sondern (z.B.) die Gewichtsfunktion (Bisquare Biweight, Huber)

$$W(x,c) = \begin{cases} 1 - \left(\frac{x}{c}\right)^2 & \text{falls} \quad |x| < c \\ 0 & \text{sonst.} \end{cases}$$

verwendet.

Robuste Lineare Regression (2)

Außerdem wird der Skalenparameter σ wird nicht durch s sondern durch den MAD geschätzt.

```
PROC ROBUSTREG;
    MODEL y=x1 x2 x3/DIAGNOSTICS LEVERAGE;
RUN;
```

`Regression_Phosphor`

Robuste Lineare Regression (3)

Diagnosestatistiken

Ausreißer: standardis. robust residual $>$ cutoff (outlier)

Leverage Point: robuste MCD-Distanz $>$ cutoff (Leverage)

Mahalanobis-Distanz

$\approx$ mit Kovarianzmatrix gewichteter mittlerer quadratischer Abstand von $\overline{X}$.

Robust MCD Distance:

anstelle von $\overline{X}$: robuste multivariate Lokationsschätzung (MCD)

Goodness of fit: zum Modellvergleich

je größer R^2, je kleiner AICR, BICR desto besser.

Nichtlineare Regression

Modell, f wird als bekannt angenommen

$$\begin{aligned} Y &= f(x, \theta) + \epsilon \\ \mathbf{Y} &= \mathbf{F}(\mathbf{X}, \boldsymbol{\theta}) + \boldsymbol{\epsilon} \end{aligned}$$

$$L(\boldsymbol{\theta}) = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = \sum_i (Y_i - \mathbf{F}(\mathbf{X}_i, \boldsymbol{\theta}))^2 \longrightarrow min_\theta$$

Dazu werden Iterationsverfahren verwendet.

```
PROC NLIN METHOD = MARQUARDT;
   MODEL abh.Var = Ausdruck;
   PARMS Anfangswerte;
RUN;
```

Nichtparametrische Regression

Modell: f unbekannt, aber ”glatt”

$$Y_i = f(\mathbf{x}_i) + \epsilon_i$$

$\epsilon_i \sim (0, \sigma^2)$ da x_i fest oder zufällig

$$\min_{f \in C^2} \sum_{i=1}^{n} (Y_i - f(x_i))^2 + \lambda \int (f''(x))^2 dx$$

- $\int (f'')^2$, Strafterm
- λ Glättungsparameter
  - $\lambda \to 0$: Interpolierender Spline
  - $\lambda \to \infty$: lineare Regression

Lösung der Minimumaufgabe: natürlicher kubischer Spline

Nichtparametrische Regression

```
PROC TPSPLINE;
   MODEL abh.Var = (unabhaengige Variablen);
   OUTPUT OUT=Datei1 PRED RESID;
RUN;
```

Wahl der Glättungsparameter

Es kann eine ganze Liste abgearbeitet werden mit der Option LOGNLAMBDA in der MODEL-Anweisung, z.B.

```
   MODEL y = (x) /LOGNLAMBDA=-4 to -2 by 0.1;
```

Visualisierung

```
PROC GPLOT DATA=Datei1;
   PLOT pred*x;
RUN;
```

Logistische Regression

Y: Binäre Zielgröße, $P(Y=1)=p, P(Y=0)=1-p$,
$Y \sim B(1,p)$

Wenn wir lineare Regression machen würden:

$$\begin{aligned} y_i &= \alpha + \beta x_i + \epsilon_i \\ \mathbf{E}Y_i &= \alpha + \beta x_i, \quad \mathbf{E}\epsilon_i = 0 \\ p_i &= \alpha + \beta x_i \end{aligned}$$

Problem: Wahrscheinlichkeiten sind beschränkt, lineare Funktionen aber nicht.

Ausweg: Odds ratio $OR := \frac{p}{1-p}$

nach oben unbeschränkt, aber nicht nach unten

Logistische Regression (2)

Logit

$$Logit(p) := \ln\Big(\frac{p}{1-p}\Big)$$

ist auch nach unten unbeschränkt.

Modell

$$\begin{aligned} Logit(p_i) &= \ln\Big(\frac{p_i}{1-p_i}\Big) \\ &= \alpha + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} = \boldsymbol{\beta}'\mathbf{x}_i, \end{aligned}$$

$i = 1, \ldots, n, p_i = P(Y_i = 1)$.
$\mathbf{x}_i' = (1, x_{i1}, \ldots, x_{ik}), \boldsymbol{\beta}' = (\alpha, \beta_1, \ldots, \beta_k)$.

Umstellen der letzten Gleichung liefert

Logistische Regression (3)

$$p_i = \frac{e^{\beta' \mathbf{x}_i}}{1 + e^{\beta' \mathbf{x}_i}}.$$

Gegeben sind Beobachtungen: $(y_i, \mathbf{x}_i)$.
Unbekannt sind p_i.

Frage: Wie schätzen wir β ?

Methode: Maximum-Likelihood

```
PROC LOGISTIC;
   MODEL Y=X1 X2 /Optionen;
RUN;
```

```
Logistic_banknote
Logistic_tibetan
Logistic_water
```

Kurze Übersicht Regressionsverfahren

a) Lineare Regression

Modell:

$$Y_i = \theta_0 + \sum_{j=1}^{m} \theta_j X_{ij} + \epsilon_i$$

$\epsilon_i \sim (0, \sigma^2)$, $i = 1, \ldots, n$
Y_i, X_i, ϵ_i zufällig (X_i kann auch fest sein)
$\theta_0 \ldots \theta_m; \sigma$: Modellparameter

```
PROC REG <Optionen>;
   MODEL abh.Variable = unabh.Variable(n)
       </Optionen>;
RUN;
```

Kurze Übersicht Regressionsverfahren (2)

b) Robuste Lineare Regression

robuste Abstandsfunktion
MAD statt s als Skalenschätzung.

```
PROC ROBUSTREG;
   MODEL abh.Variable = unabh.Variable(n)
         / diagnostics leverage;
RUN;
```

Kurze Übersicht Regressionsverfahren (3)

c) Nichtlineare Regression

Modell:

$$Y_i = f(X_{1i}, ..., X_{mi}, \theta_1, ..., \theta_p) + \epsilon_i$$

f: bekannt (i.A. nichtlinear)

```
PROC NLIN;        <METHOD = MARQUARDT>
   MODEL abh.Variable = Ausdruck;
   PARMS Parameter = Anfangswert;
RUN;
```

Kurze Übersicht Regressionsverfahren (4)

d) Nichtparametrische Regression

Modell:

$$Y_i = f(X_{1i}, ..., X_{mi}) + \epsilon_i$$

f unbekannt, aber ”glatt”, z.B. $f \in C^2$.

```
PROC TPSPLINE;
   MODEL abh.Var. = (unabh. Var);
RUN;
```

Regression_Phosphor_Uebersicht.sas

Kurze Übersicht Regressionsverfahren (5)

e) Logistische Regression

Y: binäre Zielgröße

$$p_i = P(Y_i = 1) = \frac{e^{\beta' \mathbf{x}_i}}{1 + e^{\beta' \mathbf{x}_i}}.$$

Parameter: β.
Odds ratio: $\frac{p_1}{1-p_1}$

```
proc logistic;
   model binaere Variable = abh. Variablen
run;
```

4. Zufallszahlen

- werden nach einem determinist. Algorithmus erzeugt $\Rightarrow$ Pseudozufallszahlen

- wirken wie zufäll. Zahlen (sollen sie jedenfalls)

Algorithmus:

Startwert x_0 $\quad x_{n+1} = f(x_n)$ (z.B. Kongruenzen)

Der Generator von SAS

$$x_{n+1} = \underbrace{397204094}_{2 \cdot 7 \cdot 7 \cdot 4053103} x_n \text{mod}(2^{31} - 1)$$

liefert gleichverteilte ganze Zufallszahlen auf $(0, 2^{31} - 1)$.

Zufallszahlen

auf $(0,1)$ gleichverteilte Zufallsgrössen, $U_n \sim R(0,1)$

$$U_n = \frac{x_n}{2^{31}-1}$$

```
seed = -1;  /* zufaelliger Startwert */
x=ranuni(seed) /*auf (0,1) gleichvert.ZZ. */
x=rannor(seed) /*Standard-Normal-ZZ.*/
```

Der interne Startwert wird dann durch x_1 ersetzt, der folgende Aufruf von rannor(seed) liefert eine neue Zufallszahl.

Zufallszahlen

vorgegebene stetige Verteilung

wird z.B. aus gleichverteilter Zufallsvariable U_i mittels Quantilfunktion ($F^{-1}(U_i)$) gewonnen.

diskrete Verteilungen

werden erzeugt durch Klasseneinteilung des Intervalls $(0, 1)$ entsprechend der vorgegebenen Wahrscheinlichkeiten p_i, also

$$(0, p_1], (p_1, p_1 + p_2], (p_1 + p_2, p_1 + p_2 + p_3],$$

$$\ldots, (p_1 + \cdots + p_{k-1}, 1)$$

Zufallszahlen

Wünschenswerte Eigenschaften

- Einfacher Algorithmus, wenig Rechenzeit.
- möglichst viele verschieden Zufallszahlen sollen erzeugbar sein

$\Rightarrow$ lange Periode.

- k-Tupel $(U_1, \ldots, U_k) \sim R(0,1)^k$, $k \leq 10$

$\Rightarrow$ Test auf Gleichverteilung.

- “Unabhängigkeit”

Test auf Autokorrelation (z.B. Durbin-Watson Test, vgl. Regression)
Plot der Punkte (U_i, U_{i+k}), $k = 1, 2...$
es sollten keine Muster zu erkennen sein.

`Zufallszahlen_test.sas`
`Zufallszahlen_Dichte.sas`

Clusteranalyse

Ziel: Zusammenfassung von
- “ähnlichen” Objekten zu Gruppen (Clustern),
- unähnliche Objekte in verschiedene Cluster.

Cluster sind vorher nicht bekannt.

20 Patienten, Blutanalyse

Merkmale: Eisengehalt X_1, alkalische Phosphate X_2

Umweltverschmutzung in verschiedenen Städten

Merkmale: Schwebeteilchen, Schwefeldioxid

Byzantinische Münzen

Lassen sich gesammelte Münzen verschiedenen Epochen zuordnen?

Clusteranalyse

Wir unterscheiden:

partitionierende Clusteranalyse

Zahl der Cluster ist vorgegeben
(MAXCLUSTERS=)
PROC FASTCLUS (k-means),
PROC MODECLUS (nichtparam. Dichteschätzung)

hierarchische Clusteranalyse

PROC CLUSTER, gefolgt von
PROC TREE und evtl.
PROC GPLOT

Fuzzy Clusteranalyse

Clusteranalyse

Abstandsdefinitionen (p: # Merkmale)

Euklidischer Abstand (das ist Standard)

$$d_E^2(x,y) = \sum_{i=1}^{p} (x_i - y_i)^2$$

City-Block Abstand (Manhattan-Abstand)

$$d_C(x,y) = \sum_{i=1}^{p} |x_i - y_i|$$

Tschebyscheff-Abstand

$$d_T(x,y) = \max_i |x_i - y_i|$$

Clusteranalyse

- Nichteuklidische Abstände müssen selbst berechnet werden.

Macro %DISTANCE

- Abstandsmatrix kann in der DATA-Anweisung angegeben werden.

DATA=name (TYPE=DISTANCE)

- Die Variablen sollten i.A. vor der Analyse standardisiert werden, da Variablen mit großer Varianz sonst großen Einfluß haben (Option STANDARD oder die Prozedur ACECLUS zuvor laufen lassen).

davor: Ausreißer beseitigen.

Hierarchische Clusteranalyse

Methoden (1)

Die Methoden unterscheiden sich durch die Definition der Abstände $D(C_i, C_j)$ zwischen Clustern C_i und C_j.

Single Linkage

$$D_S(C_i, C_j) = \min\{d(k,l), k \in C_i, l \in C_j)\}$$

Complete Linkage

$$D_C(C_i, C_j) = \max\{d(k,l), k \in C_i, l \in C_j)\}$$

Centroid

$D_{CE}(C_i, C_j) = d(\overline{X}_i, \overline{X}_j)$ Abstände der Schwerpunkte

Hierarchische Clusteranalyse

Methoden (2)

Average Linkage

$D_A(C_i, C_j) = \frac{1}{n_i n_j} \sum_{k \in C_i, j \in C_j} d(k, l)$

Ward

ANOVA-Abstände innerhalb der Cluster minimieren, außerhalb maximieren. Nach Umrechnen erhält man
$D_W(C_i, C_j) = \frac{n_i n_j}{n_i + n_j} D_{CE}(C_i, C_j)$.

Density Linkage

beruht auf nichtparametrischer Dichteschätzung (DENSITY, TWOSTAGE)

Hierarchische Clusteranalyse

Tendenzen

WARD:	Cluster mit etwa gleicher Anzahl von Objekten
AVERAGE:	ballförmige Cluster
SINGLE:	große Cluster, “Ketteneffekt”, langgestreckte Cluster
COMPLETE:	kompakte, kleine Cluster

Im Mittel erweisen sich **Average Linkage** und **Ward** sowie die nichtparametrischen Methoden als die geeignetsten Methoden.

Hierarchische Clusteranalyse

Agglomerative Verfahren

1. Beginne mit der totalen Zerlegung, d.h.
   $Z = \{C_1, ..., C_n\}, C_i \cap C_j = \varnothing \quad C_i = \{O_i\}$
2. Suche C_r, C_l : $\quad d(C_r, C_l) = \min_{i \neq j} d(C_i, C_j)$
3. Fusioniere C_r, C_l zu einem neuen Cluster:
   $C_r^{new} = C_r \cup C_l$
4. Ändere die r-te Zeile und Spalte der Distanzmatrix durch Berechnung der Abstände von C_r^{new} zu den anderen Clustern! Streiche die l-te Zeile und Spalte!
5. Beende nach n-1 Schritten, ansonsten fahre bei 2. mit geänderter Distanzmatrix fort!

Hierarchische Clusteranalyse

- Alle von SAS angebotenen hierarchischen Methoden sind agglomerativ.
- Es gibt auch divisive Methoden.
- Fall großer Datensätze:

PROC FASTCLUS: Vorclusteranalyse mit großer Anzahl von Clustern
PROC CLUSTER: mit diesen Clustern.

Hierarchische Clusteranalyse

zu WARD:

ANOVA Abstände innerhalb eines Clusters i

$$D_i = \frac{1}{n_i} \sum_{l \in C_i} d^2(O_l, \overline{X}_i)$$

Fusioniere die Cluster C_i und C_j, wenn

$$D_{CE}(C_i, C_j) - D_i - D_j \longrightarrow \min_{i,j}$$

Clusteranalyse

Durchführung

```
PROC CLUSTER /*hierarchische Clusteranalyse*/
   METHOD=methode
   STANDARD /*Standardisierung*/
   OUTREE=datei;/*Eingabedatei fuer Proc Tree*/
RUN;
PROC TREE DATA=datei
   OUT=out /*Ausgabedatei z.B.f. PROC GPLOT*/
   NCLUSTERS=nc /*Anz. Cluster*/
   COPY vars /*vars in die Ausgabedatei*/
RUN;
PROC GPLOT;
   PLOT variablen=cluster;/*Symbol-Anweis.
         vorher definieren*/
RUN;
```

Hierarchische Clusteranalyse

Die Ausgabedatei OUTTREE=

NAME	Bezeichnung der Cluster
	≥ 2 Beobachtungen: CLn
	1 Beobachtung: OBn
NCL	Anzahl der Cluster
FREQ	Anzahl der Beobachtungen im jeweiligen Cluster

n: Clusternummer (CLn) oder
Beobachtungsnummer (OBn = _N_)

```
Cluster_Air.sas
Cluster.sas
Cluster_Banknoten.sas
Cluster_Muenzen.sas
```

3D-Darstellung von Datenpunkten

```
PROC G3D;
   SCATTER y*x = z;
RUN;

/*Wertetabelle erstellen,
vgl. z.B. Texashut.sas*/
PROC G3D;
   PLOT y*x = z;
RUN;
```

Glatte 3D-Darstellung, Kontur-Plot

Glatte 3D-Darstellung

```
PROC G3GRID;
   GRID var1*var2=y/SPLINE SMOOTH=Wert;
      AXIS1=von TO bis BY Schrittweite;
      AXIS2=von TO bis BY Schrittweite;
RUN;
```

Kontur-Plot

```
PROC GCONTOUR;
   PLOT var1*var2 = y /LLEVEL=1;
RUN;
```

Erläuterung dazu siehe Programm

`Npar_Banknote.sas`

7. Zusammenfassung

Zusammenfassung

Basiswissen

- Klassifikation von Merkmalen
- Wahrscheinlichkeit
- Zufallsvariable
- Diskrete Zufallsvariablen (insbes. Binomial)
- Stetige Zufallsvariablen
- Normalverteilung
- Erwartungswert, Varianz
- Gesetz der großen Zahlen,
  Zentraler Grenzwertsatz

Beschreibende Statistik

(Robuste) Lage- und Skalenschätzungen

```
PROC UNIVARIATE TRIMMED=Zahl
  ROBUSTSCALE; RUN;
```

Boxplots

```
PROC BOXPLOT;  PLOT Variable*Faktor
        /BOXSTYLE=SCHEMATIC;  RUN;
```

Häufigkeitsdiagramme:

```
PATTERN1 ...;
PROC GCHART; VBAR Variable; RUN;
```

Scatterplots, Regressionsgerade:

```
SYMBOL1 ...;
PROC GPLOT; PLOT y*x=1 / REGEQN; RUN;
```

Zusammenfassung Statistische Tests
und Multivariate Verfahren

Testproblem: Nullhypothese - Alternative, z.B.

$H_0 : \mu = \mu_0 \qquad H_1 : \mu \neq \mu_0$

Entscheidung für H_0/gegen H_0: anhand einer

Teststatistik, z.B.

$$T = \frac{\overline{X} - \mu_0}{S} \cdot \sqrt{n}$$

Entscheidung

$|t| > t_{krit} \Rightarrow H_0$ ablehnen, $\quad P(|T| > t_{krit}) = \alpha$

α : Fehler 1. Art, Signifikanzniveau (in der Regel vorgegeben)

Zusammenfassung Statistische Tests (2)

p-Wert (zweiseitg)

$P(|T| > t)$, wobei t: Realisierung von T

p-Wert $< \alpha \Rightarrow H_0$ ablehnen

p-Wert $\geq \alpha \Rightarrow H_0$ nicht ablehnen

Gütefunktion

P(H_0 abgelehnt$|\mu$ richtig) = $\beta(\mu)$
Fehler 2.Art: $1 - \beta(\mu)$

Wir betrachten Tests mit einer vergleichsweise hohen Gütefunktion.

Zusammenfassung Statistische Tests (3)

Einseitige Tests

Alternative geht in eine Richtung, (aus sachlichen Gründen kann es nur eine Richtung geben)

z.B. $\mu > \mu_0$

Zweiseitige Tests

Alternative geht in alle Richtungen,

z.B. $\mu \neq \mu_0$

Übersicht über Mittelwertvergleiche (1)

k	unverbunden	verbunden
1	Einstichproben t-Test,Vorzeichen-Wilcoxon-Test	
	PROC UNIVARIATE; o. PROC TTEST H0=Wert; VAR Variable; RUN	
2	t-Test	t-Test
	PROC TTEST; CLASS=Faktor; VAR Variable; RUN;	PROC TTEST; PAIRED Var1*Var2; RUN;
	Wilcoxon-Test	Vorzeichen-Wilcoxon-Test
	PROC NPAR1WAY WILCOXON; CLASS=Faktor; VAR Variable;RUN;	diff=a-b; PROC UNIVARIATE; VAR diff; RUN;

Übersicht über Mittelwertvergleiche (2)

einfache Varianzana. = einfaktorielle VA	einfaches Blockexperiment = zweifaktorielle VA
PROC ANOVA; CLASS Faktor; MODEL Y=Faktor; RUN; (PROC GLM)	PROC GLM; CLASS FaktorA FaktorB; MODEL Y=FaktorA FaktorB; RUN;
Kruskal-Wallis-Test	Friedman-Test
PROC NPAR1WAY Wilcoxon; CLASS Faktor; VAR var; RUN;	PROC FREQ; TABLES FaktorA*FaktorB*Y / CMH2 SCORES=RANK NOPRINT; RUN;

Anpassungstest auf Normalverteilung:

```
PROC UNIVARIATE NORMAL; VAR var; RUN;
```

Shapiro-Wilk-Test oder Anderson-Darling-Test

Anpassungstest auf Verteilung mit begrenzter Anzahl von Ausprägungen

```
PROC FREQ; TABLES Var1 /CHISQ NOPRINT
          TESTP=(p1,p2,...pk); RUN;
```

($p_1, \ldots, p_k$ vorher ausrechnen)

Test auf Korrelation (metrisch oder ordinal skalierte Merkmale)

```
PROC CORR PEARSON SPEARMAN KENDALL; RUN;
```

Test auf Unabhängigkeit (beliebig skalierte Merkmale):

```
PROC FREQ;
  TABLES Var1*Var2 /CHISQ NOPRINT; RUN;
```

Lineare Regression (1)

Parameterschätzung und Test

```
PROC REG;
  MODEL Y=Var1 Var2 ... Varn / CLI CLM R;
  TEST Var1=0 Var2=0;/*Zusaetzl.Hypothesen */
RUN;
```

Modellwahl

```
PROC REG;
  MODEL Y=Var1 Var2 ... Varn /
        SELECTION=backward; RUN;
```

Lineare Regression (2)

Residualanalyse

```
PROC REG;
  MODEL Y=Var1 Var2 ... Varn / R;
  PLOT rstudent.*obs.;  /*und/oder*/
  PLOT residual.*y; residual.*predicted.;
RUN;
```

und evtl. Test auf Normalverteilung.

Sonstige Regressionsverfahren, nur Übersicht

Robuste Lineare Regression
Nichtlineare Regression
Nichtparametrische Regression
Logistische Regression

Hierarchische Clusteranalyse:

```
PROC CLUSTER
  METHOD=Average
         (oder: CENTROID oder WARD)
  OUTTREE=baum; VAR Variablen; RUN;
PROC TREE DATA=baum
  NCLUSTERS = Anzahl der Cluster
              fuer GPLOT;
  OUT=Eingabedatei fuer Proc GPLOT;
RUN;
PROC GPLOT; PLOT VarA*VarB=cluster; RUN;
```

Konfidenzbereiche

für Parameter im Regressionsmodell

```
PROC REG;
   MODEL Y=var1...varn/ CLI CLM;
RUN;
```

Grafische Darstellung von Konfidenzbereichen bei der Regression

```
SYMBOL1 I=RLCLI95;
PROC GPLOT; PLOT y*x=1; RUN;
```

Wichtige Sprachelemente

Normalverteilte Zufallsvariable
mit zufälligem Startwert: seed=-1; RANNOR(seed);

Gleichverteilte Zufallsvariable
mit zufälligem Startwert: seed=-1; RANUNI(seed);

Wahrscheinlichkeitsverteilungen:

Verteilungsfunktion (Parameter)

`CDF('Verteilung',z,Parameterliste)`

Dichte oder Wahrscheinlichkeitsfunktion (Parameter)

`PDF('Verteilung',z,Parameterliste)`
z.B.: ('normal',z,0,1)
('binomial',z,n,p)

Quantile

Standardnormal: `PROBIT(u)`, $u \in (0,1)$.
`Quantile('Verteilung',z,Parameterliste)`

Übungen (1)

1. Folgen und Reihen, Potenzreihen
2. Differential- und Integralrechnung, Normalverteilung
3. Integrralrechnung, Rechnen mit Erwartungswerten
4. Berechnen von Erwartungswerten, Berechnen von robusten Lage- und Skalenschätzungen
5. Berechnen von Korrelationen
6. Korrelationen, Einfluss von Ausreißern, Minima von Funktionen zweier Veränderlicher
7. Aufgabenblatt 7, Regressionsmodel, Berechnen von t-Teststatistiken
8. Aufgabenblatt 8, t-Test und Varianzanalyse

Übungen (2)

9. Aufgabenblatt 9,
   Produkt von Matrizen, Eigenwerte, Eigenvektoren
10. Aufgabenblatt 10,
    Lineare Algebra, Matrizenrechnung, χ^2-Verteilung
11. Aufgabenblatt 11
12. Aufgabenblatt 12

Übungsaufgaben

7,8,9 Wahrscheinlichkeitsverteilungen

10,11 Statist. Maßzahlen, Boxplots

11 Histogramme, Dichteschätzung

11h-14,29,32,33,34 Korrelation, Unabhängigkeit, Lineare Regression

15-21,23-25 Lagetests, Anpassungstests

19,22 Varianzanalyse

26-28,30-31 Nichtparametrische Tests

35,36 Zufallszahlen

36 Clusteranalyse