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)

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

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=Dateiname PRED RESID;
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)

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

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

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

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

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