6.2 Lineare Regression

Einfache lineare Regression (vgl. Kap. 4.7)

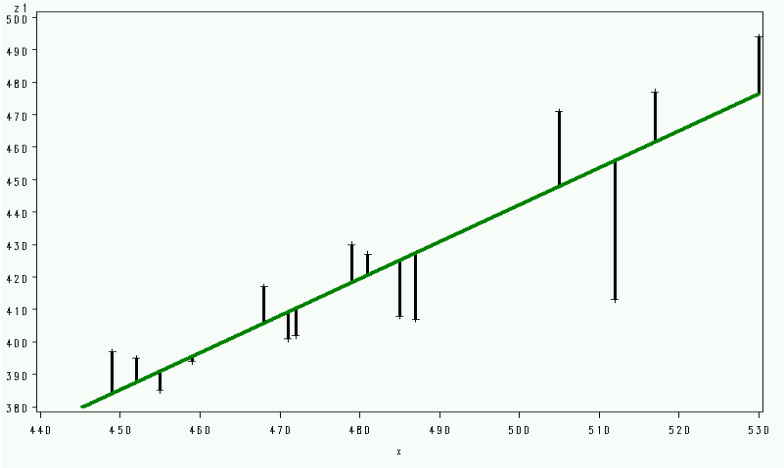
$$Y_i = \theta_0 + \theta_1 X_i + \epsilon_i \qquad \epsilon_i \sim (0, \sigma^2)$$

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

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

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

$\mathbf{H}$: 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)

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

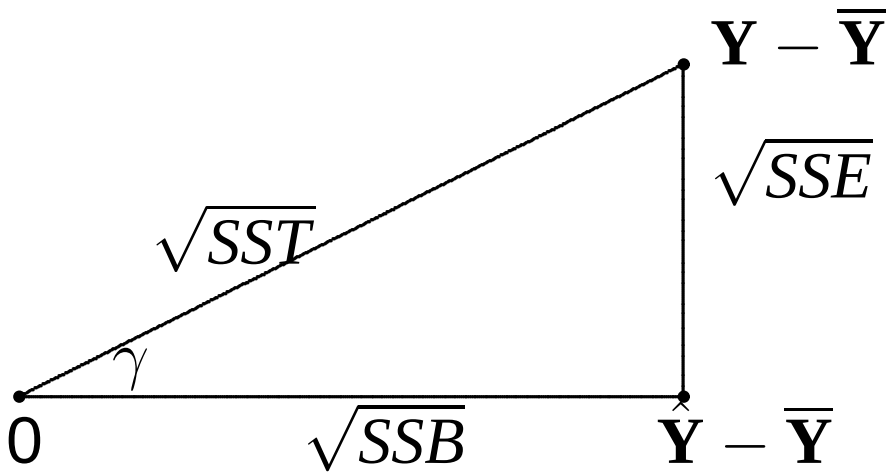
Geometrische Veranschaulichung

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 der 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

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 Beob. 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.;
OUTPUT OUT=dateiname RESIDUAL=;
PLOT residual.*y     residual.*predicted.;
```

und evtl. Test auf Normalverteilung.