

Vorlesung: Lineare Modelle

Prof. Dr. Helmut Küchenhoff

Institut für Statistik, LMU München

SoSe 2015

Homepage: <https://www.elab.moodle.elearning.lmu.de/course/view.php?id=692>

Vorlesung:

Prof. Helmut Küchenhoff

Mittwoch, 10:00 - 12:00 Uhr Geschw.-Scholl-Pl. 1 E 004

Donnerstag, 14:00 - 16:00 Uhr Schellingstr. 3 (S)Vg.-004

Übung:

Andreas Bender, Veronika Deffner

Gruppe 1: Montag, 10 - 12 Uhr Geschw.-Scholl-Pl. 1 A017

Gruppe 2: Dienstag, 14 - 16 Uhr Geschw.-Scholl-Pl. 1 M001

Tutorium

Felix Günther

Literatur

L. Fahrmeir, Th. Kneib, S. Lang:

Regression

Springer-Verlag Deutsche und englische Fassung

J. Faraway:

Linear Models with R

Extending the Linear Model with R Englische Fassung

0 Einführung und Beispiele

1 Das einfache lineare Regressionsmodell

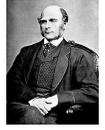
3 Das multiple lineare Regressionsmodell

3 Quadratsummenzerlegung und statistische Inferenz im multiplen linearen Regressionsmodell

5 Diskrete Einflußgrößen: Dummy- und Effektkodierung, Mehrfaktorielle Varianzanalyse

Sir Francis Galton

Sir Francis Galton (1822-1911) „Regression toward mediocrity in hereditary stature“



Gibt es eine Midlife - Crisis ?

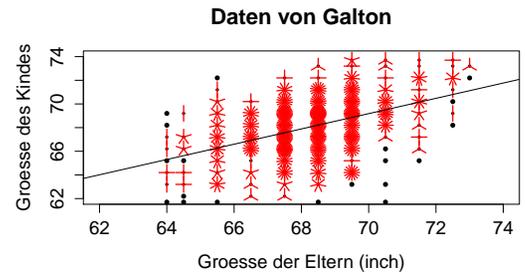
Daten über die Zufriedenheit von Personen

$$E(Y_t) = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 \log(x_{3t}) + \beta_4 x_{4t} + \dots + \beta_{A1} AI + \beta_{A2} AI^2 + \beta_{A3} AI^3$$

x-Variablen: Gehalt, Geschlecht, Gesundheitsstatus etc.

→ multiple lineare Regression

Der Begriff Regression



Zusammenhang zwischen Größe der Eltern und ihrer Kinder

$$E(Y) = \bar{Y} + 0.65(X - \bar{X})$$
$$E(Y) = 24 + 0.65X$$

Y: Größe des Kindes X: Größe der Eltern

Botanik: Wachstum von Pflanzen

In einem Versuchswald wird bei einem Teil der Bäume eine Kalkung vorgenommen. Annahme von linearem Wachstum.

$$E(Y_{ijt}) = \beta_0 + \beta_j(t - t_0)$$

Y_{ijt} : Baumhöhe von Baum i in der Gruppe j zum Zeitpunkt t
 β_j : Wachstumsparameter in Gruppe j (j=1: Kalk, j=2: kein Kalk)

→ Kovarianzanalyse

Zusammenhang zwischen Lesefähigkeit und Lehrmethode

$$E(Y_{ij}) = \beta_0 + \beta_1 x_{1ij} + \alpha_j$$

x_{1ij} : Zahl der Förderstunden
 α_j : Klasseneffekt

→ Gemischtes lineares Modell

Zusammenhang zwischen dem Auftreten von Chronischer Bronchitis (CBR) und Staubbelastung am Arbeitsplatz.
Studie: 1246 Arbeitnehmer

$$P(Y_i = 1) = G(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 IR_i)$$

Variablen:

Y : Indikator für CBR
 x_1 : Staubkonzentration am Arbeitsplatz
 x_2 : Dauer der Exposition
 IR : Rauchstatus (ja/nein)

→ Logistisches Regressionsmodell



- 0 Einführung und Beispiele
- 1 Das einfache lineare Regressionsmodell**
- 3 Das multiple lineare Regressionsmodell
- 3 Quadratsummenzerlegung und statistische Inferenz im multiplen linearen Regressionsmodell
- 5 Diskrete Einflußgrößen: Dummy- und Effektkodierung, Mehrfaktorielle Varianzanalyse

Das einfache lineare Regressionsmodell

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i; \quad i = 1, \dots, n \quad (1.1)$$

$$E(\varepsilon_i) = 0 \quad (1.2)$$

$$V(\varepsilon_i) = \sigma^2 \quad (1.3)$$

$$\{\varepsilon_i | i = 1, \dots, n\} \quad \text{stoch. unabhängig} \quad (1.4)$$

$$\varepsilon_i \sim N(0, \sigma^2) \quad (1.5)$$

Y_i : Zielgröße (Zufallsgröße), abhängige Variable
 x_i : **feste** bekannte Einflussgröße, unabhängige Variable
 ε_i : Zufallsfehler
 $\beta_0, \beta_1, \sigma^2$: unbekannte Parameter
 n : Anzahl der Beobachtungen

Wir betrachten Modell (1.1).
Dann heißt:

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg \min_{\beta_0, \beta_1} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2 \quad (1.6)$$

KQ-Schätzer (Schätzer nach der Methode der kleinsten Quadrate)

$$\hat{\varepsilon}_i := Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \quad \text{heien **Residuen**.} \quad (1.7)$$

Interpretation der Modellparameter I

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i; \quad i = 1, \dots, n$$
$$E(Y|X) = \beta_0 + \beta_1 x$$

Systematische Komponenten: $\beta_0 + \beta_1 x_i$

- Zentrale Groe ist der Steigungsparameter β_1
Wenn X um eine Einheit steigt, dann steigt Y **im Durchschnitt** um β_1 Einheiten
Die Groe von β_1 hangt von der Einheit von X ab. Sie kann also nur in Zusammenhang mit X interpretiert werden.
- Interpretation des Achsenabschnitts β_0 :
Erwartungswert von Y bei X=0. In vielen Fallen irrelevant

Der KQ-Schatzer **existiert** und ist **eindeutig**, (falls $\sum_{i=1}^n (x_i - \bar{x})^2 \neq 0$):

$$\hat{\beta}_1 = \frac{S_{xY}}{S_x^2} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (1.8)$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}. \quad (1.9)$$

Beweis:

Durch Differenzieren von (1.6) erhalt man: $(\hat{\beta}_0, \hat{\beta}_1)$ sind Losung der **Normalgleichungen**

$$\sum_{i=1}^n \hat{\varepsilon}_i = 0 \quad (1.10)$$

$$\sum_{i=1}^n \hat{\varepsilon}_i x_i = 0 \quad (1.11)$$

Interpretation der Modellparameter II

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i; \quad i = 1, \dots, n$$
$$E(Y|X) = \beta_0 + \beta_1 x$$

Stochastische Komponente : Storterms ε

- Interpretation des Parameters σ :
Standardabweichung des Storterms.
Ma fur die durchschnittliche Abweichung der Y-Werte von der Regressionsgeraden.

Eigenschaften des KQ-Schätzers I

Gegeben sei Modell (1.1) mit Annahme (1.2).

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i; \quad i = 1, \dots, n \tag{1.1}$$

$$E(\varepsilon_i) = 0 \tag{1.2}$$

1 Dann ist $(\hat{\beta}_0, \hat{\beta}_1)$ ein **erwartungstreuer** Schätzer für (β_0, β_1) :

$$E(\hat{\beta}_0, \hat{\beta}_1) = (\beta_0, \beta_1). \tag{1.12}$$

Eigenschaften des KQ-Schätzers III

3 Unter der NV-Annahme (1.5)

$$\varepsilon_i \sim N(0, \sigma^2) \tag{1.5}$$

ist der KQ-Schätzer $(\hat{\beta}_0, \hat{\beta}_1)$ **ML-Schätzer**.

Eigenschaften des KQ-Schätzers II

2 Für die **Varianzen** von $(\hat{\beta}_0, \hat{\beta}_1)$ gilt unter den Annahmen (1.3), (1.4):

$$V(\varepsilon_i) = \sigma^2 \tag{1.3}$$

$$\{\varepsilon_i | i = 1, \dots, n\} \quad \text{stoch. unabhängig} \tag{1.4}$$

$$V(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sigma^2 / nS_x^2 \tag{1.13}$$

$$V(\hat{\beta}_0) = \sigma^2 \left[\frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \right] = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{nS_x^2} \right] \tag{1.14}$$

Beweis des Satzes I

1 Es ist zu beachten, dass die x_i fest sind und dass die einzige stochastische Komponente des Modells ε_i ist.

$$\begin{aligned}
E(\hat{\beta}_1) &= E \left[\frac{\sum_{i=1}^n (Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \\
&= \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \sum_{i=1}^n E[(\beta_0 + \beta_1 x_i + \varepsilon_i - \beta_0 - \beta_1 \bar{x} - \bar{\varepsilon})(x_i - \bar{x})] \\
&= \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \sum_{i=1}^n [(\beta_0 + \beta_1 x_i + E[\varepsilon_i] - \beta_0 - \beta_1 \bar{x} - E[\bar{\varepsilon}])(x_i - \bar{x})] \\
&= \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \sum_{i=1}^n \beta_1 (x_i - \bar{x})(x_i - \bar{x}) = \beta_1 \\
E(\hat{\beta}_0) &= E[\bar{Y} - \hat{\beta}_1 \bar{x}] = E[\beta_0 + \beta_1 \bar{x} + \bar{\varepsilon}] - \beta_1 \bar{x} = \beta_0
\end{aligned}$$

2 Übung

Beweis des Satzes II

3 Die Likelihood von Beobachtungen (y_i, x_i) lautet:

$$L(y_1, \dots, y_n, x_1, \dots, x_n) = \prod_{i=1}^n \left(\sqrt{2 * \pi * \sigma^2} \right)^{-1} \exp \left[-\frac{[\varepsilon_i(\beta_0, \beta_1)]^2}{2\sigma^2} \right]$$

$$\ln L(y_1, \dots, y_n, x_1, \dots, x_n) = -n/2 * \ln(\sigma^2 * 2 * \pi) - \frac{1}{2\sigma^2} \sum_{i=1}^n \varepsilon_i^2$$

Da die Parameter β_0 und β_1 nur in $\sum_{i=1}^n \varepsilon_i^2$ vorkommen, entspricht die Maximierung von $\ln L(y_i, x_i)$ der Minimierung von $\sum_{i=1}^n \varepsilon_i^2$. Damit entspricht die KQ-Methode der ML-Methode.

Als ML-Schätzung von σ^2 ergibt sich nach Einsetzen von $\hat{\varepsilon}_i := \varepsilon_i(\hat{\beta}_0, \hat{\beta}_1)$:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i^2$$

Dieser Schätzer ist nicht erwartungstreu. Daher wird er selten verwendet.

Schätzung von σ^2 und Konfidenzintervalle für β_0 und β_1 II

2 Unter der Normalverteilungsannahme (1.5) gilt:

$$\frac{1}{\sigma^2} \sum_{i=1}^n \hat{\varepsilon}_i^2 \sim \chi_{n-2}^2 \tag{1.16}$$

$$(\hat{\beta}_0, \hat{\beta}_1) \text{ und } \hat{\sigma}^2 \text{ stochastisch unabhängig} \tag{1.17}$$

Schätzung von σ^2 und Konfidenzintervalle für β_0 und β_1 I

Gegeben sei das Modell (1.1) bis (1.4).

1 Dann ist

$$\hat{\sigma}^2 := \frac{1}{n-2} \sum_{i=1}^n \hat{\varepsilon}_i^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 \tag{1.15}$$

ein **erwartungstreuer** Schätzer für σ^2 .

Schätzung von σ^2 und Konfidenzintervalle für β_0 und β_1 III

3 Unter (1.5) gilt für die Schätzer $\hat{\beta}_1$ und $\hat{\beta}_0$:

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim t_{n-2} \text{ mit } \hat{\sigma}_{\hat{\beta}_1} := \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \tag{1.18}$$

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \sim t_{n-2} \text{ mit } \hat{\sigma}_{\hat{\beta}_0} := \sqrt{\hat{\sigma}^2 \left[\frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \right]} \tag{1.19}$$

- 4 Konfidenzintervalle zum Niveau $1 - \alpha$ für β_1 und β_0 unter Normalverteilungsannahme (1.5):

$$[\hat{\beta}_1 - \hat{\sigma}_{\hat{\beta}_1} t_{1-\alpha/2}(n-2); \hat{\beta}_1 + \hat{\sigma}_{\hat{\beta}_1} t_{1-\alpha/2}(n-2)] \quad (1.20)$$

$$[\hat{\beta}_0 - \hat{\sigma}_{\hat{\beta}_0} t_{1-\alpha/2}(n-2); \hat{\beta}_0 + \hat{\sigma}_{\hat{\beta}_0} t_{1-\alpha/2}(n-2)] \quad (1.21)$$

$t_{1-\alpha/2}(n-2)$: $1 - \alpha/2$ -Quantil der $t(n-2)$ -Verteilung.

Beispiel: Analyse von Daten aus Reihengräbern in Wenigumstadt

Fragestellung: Zusammenhang zwischen Alter (zum Todeszeitpunkt) und Knochenbälkchendicke des 4. Lendenwirbels

Im Zentrum der Interpretation stehen

- Parameterschätzung von β_1
- Schätzung der durchschnittlichen Abweichung von der Regressionsgeraden
- R^2

Probleme

- Messfehler
- sehr kleine Werte
- x_i sind zufällig

Teil 1 und Teil 2 später als Spezialfall im multiplen Regressionsmodell

Teil 3:

Def. t-Verteilung: $\left. \begin{matrix} X_1 \sim N(0; 1) \\ X_2 \sim \chi_n^2 \\ X_1, X_2 \text{ unabh.} \end{matrix} \right\} \frac{X_1}{\sqrt{\frac{X_2}{n}}} \sim t_n$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right), \text{ da } \hat{\beta}_1 = \sum \frac{x_i - \bar{x}}{\sum(x_i - \bar{x})^2} y_i \sim NV$$

(Summe von unabh. NV ZG)

Aus der Def. der t-Vert. und Teil 2 \Rightarrow Behauptung

Teil 4:

Standard Konstruktion von Konfidenzintervallen

Quadratsummenzerlegung I

Gegeben sei das Modell (1.1). Dann gilt:

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \underbrace{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}_{SSE} + \underbrace{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}_{SSM} \quad (1.22)$$

mit den **angepassten Größen** $\hat{Y}_i := \hat{\beta}_0 + \hat{\beta}_1 x_i$ (1.23)

SST:	Gesamtstreuung von Y	Sum of Squares Total
SSE:	Streuung der Residuen	Sum of Squares Error
SSM:	Streuung, die das Modell erklärt	Sum of Squares Model

2

$$R^2 = \frac{SSM}{SST} = 1 - \frac{SSE}{SST} \tag{1.24}$$

heißt **Bestimmtheitsmaß**.

Es gilt

$$R^2 = r_{xY}^2. \tag{1.25}$$

r_{xY} := Korrelationskoeffizient nach Bravais-Pearson.

Beweis

Beweis von 1)

$$\begin{aligned} \sum (y_i - \bar{y})^2 &= \sum (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\ &= \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2 + 2 \underbrace{\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})}_C \end{aligned}$$

$$C = \sum \hat{\varepsilon}_i (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y}) = 0$$

Zu den obigen Quadratsummen wird üblicherweise die Zahl der Freiheitsgrade angegeben. Sie bezeichnet die Anzahl der frei bestimmbar Summanden der obigen Quadratsummen (bei geg. x_i)

bei SST: $\sum (y_i - \bar{y}) = 0$
 \Rightarrow **df=n-1**

bei SSE: NGL liefern 2 Restriktionen
 \Rightarrow **df=n-2** ($\sum \hat{\varepsilon}_i = 0; \sum \varepsilon_i x_i = 0$)

bei SSM: $\sum (\hat{y}_i - \bar{y})^2 = \sum (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2 = \sum (\hat{\beta}_1 (x_i - \bar{x}))^2$
 \Rightarrow **df=1** (durch Wählen von 1 y-Wert liegt $\hat{\beta}_1$ fest)

Interpretation von R^2

R^2 ist das zentrale Maß Anpassungsgüte der Regression

- Anteil der Varianz von Y, die durch X erklärt wird
- R ist skalenunabhängig
- R ist symmetrisch bzgl. X und Y
- Vorsicht: R^2 hängt auch von der Streuung von X in der Stichprobe ab

Prognose I

In Modell (1.1) mit (1.2) - (1.4) betrachten wir eine weitere Beobachtung x_{n+1} mit zugehörigem unbekanntem Y_{n+1} . Der Prognosewert von Y_{n+1} ist gegeben durch

$$\hat{Y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1} \tag{1.26}$$

Für den Erwartungswert und die Varianz des Prognosefehlers gilt:

$$E[\hat{Y}_{n+1} - Y_{n+1}] = 0 \tag{1.27}$$

$$V[\hat{Y}_{n+1} - Y_{n+1}] = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \tag{1.28}$$

Bemerkungen

- 1 Alle Aussagen gelten nur unter der zentralen Modellannahme des linearen Zusammenhangs von $E(Y)$ und x .
- 2 Transformationen sind grundsätzlich möglich. Zu beachten sind dann die geänderte Interpretation der Modellparameter und der Modellannahmen. Insbesondere ist

$$E[g(Y)] \neq g[E(Y)]$$

- 3 Das lineare Modell ist in vielen Beispielen eine sinnvolle Näherung

Prognose II

Prognoseintervall für y_{n+1} zum Niveau $1 - \alpha$:

$$[\hat{Y}_{n+1} - \hat{\sigma}_{\hat{Y}_{n+1}} t_{1-\alpha/2}(n-2); \hat{Y}_{n+1} + \hat{\sigma}_{\hat{Y}_{n+1}} t_{1-\alpha/2}(n-2)] \tag{1.29}$$

$$\text{mit } \hat{\sigma}_{\hat{Y}_{n+1}}^2 = \hat{\sigma}^2 \left[1 + 1/n + (x_{n+1} - \bar{x})^2 / \left(\sum_{i=1}^n (x_i - \bar{x})^2 \right) \right] \tag{1.30}$$

Lineares Modell als sinnvolle Annäherung I

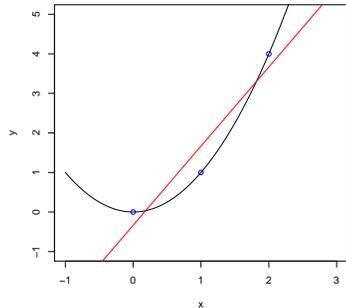
Gegeben sei ein quadratisches Modell:

$$Y_i = x_i^2 + \epsilon \text{ mit } E(\epsilon) = 0, \quad x_1 = 0, x_2 = 1, x_3 = 2 \text{ und } y_1, y_2, y_3$$

Angenommen wird aber ein lineares Modell: **KQ-Schätzung**

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon$$

$$\begin{aligned} E(\hat{\beta}_1) &= \frac{\sum_{i=1}^3 (x_i - \bar{x}) E(Y_i - \bar{Y})}{\sum_{i=1}^3 (x_i - \bar{x})^2} \\ &= \frac{1}{2} (-1E(Y_1 - \bar{Y}) + 1E(Y_3 - \bar{Y})) \\ &= \frac{1}{2} (-0^2 + 4) = 2 \\ E(\hat{\beta}_0) &= E(\bar{Y} - \hat{\beta}_1 \bar{x}) = \frac{5}{3} \cdot 2 \cdot 1 = -\frac{1}{3} \end{aligned}$$



Lineares Modell als sinnvolle Annäherung II

Je nachdem, wie die x -Werte (unter Annahme der Richtigkeit des quadratischen Modells) liegen, bekommt man eine andere Regressionsgerade (bei Annahme eines linearen Zusammenhangs).

- Näherung im Bereich $[0, 2]$ ist akzeptabel (abhängig von der Größe des Störterms).
- Prognose außerhalb von $[0, 2]$ ergibt falsche Ergebnisse. Auch das Prognose-Intervall ist vollkommen falsch: z.B. $x = -1$ ergibt eine Prognose von $\hat{y} = -1$, obwohl $y = 1$ richtig ist.

