

1. Tutorium Generalisierte Regression

- Lineare Modelle -

Cynthia Huber:

14.10.2014 und 21.10.2014

Michael Hanselmann:

16.10.2014 und 23.10.2014

Institut für Statistik, LMU München

Gliederung

- 1 Das klassische lineare Modell
- 2 Lineare Modelle mit R
- 3 Varianzanalyse
- 4 Umgang mit Faktoren

Gliederung

1 Das klassische lineare Modell

2 Lineare Modelle mit R

3 Varianzanalyse

4 Umgang mit Faktoren

Einfaches lineares Modell

Modellierung des Zusammenhangs zwischen einer Einflussgröße und einer **skalaren** Zielgröße:

$$y_i = \beta_0 + x_i \beta_1 + \varepsilon_i$$

$i = 1, \dots, n$ (Beobachtungen)

Annahmen:

- $\mathbb{E}(\varepsilon_i) = 0 \quad \forall i$ **Exogenität**
- $\text{Var}(\varepsilon_i) = \sigma^2 \quad \forall i$ **Homogenität**
- $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i \neq j$ **Unkorreliertheit**
- [Für Inferenz: $\varepsilon_i \sim N(0, \sigma^2) \quad \forall i$ **Normalverteilung**]

Multiple lineares Modell

Modellierung des Zusammenhangs zwischen **p** Einflussgrößen und einer **skalaren** Zielgröße:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

bzw.

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} + \varepsilon_i$$

$i = 1, \dots, n$ (Beobachtungen)

$j = 1, \dots, p$ (Einflussgrößen ohne Intercept)

Annahmen:

- $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0} \quad \forall i$ **Exogenität**
- $\text{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$ **Homogenität, Unkorreliertheit**
- [Für Inferenz: $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$ **Multivariate Normalverteilung**]

Interpretation des Modells

$$\mathbb{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$$

Interpretation

- Steigt x_j um eine Einheit, so steigt die Zielgröße im Erwartungswert um β_j Einheiten, wenn alle anderen Kovariablen festgehalten werden
- Linearer Zusammenhang von $\mathbb{E}(Y)$ und x_j bei Festhalten der übrigen Kovariablen

Gliederung

- 1 Das klassische lineare Modell
- 2 Lineare Modelle mit R**
- 3 Varianzanalyse
- 4 Umgang mit Faktoren

Modellbildung

`lm(formula=Modellformel, data=Daten, ...)`

Beispiele für Modellformeln:

- *abh.Variable* ~ .

Erklärung der abhängigen Variable durch alle übrigen Größen und einem Interceptterm

- *abh.Variable* ~ *Variable i* + *Variable j*

Erklärung der abhängigen Variable durch die Variablen i und j und einem Interceptterm

- *abh.Variable* ~ *Variable i* + *Variable j* - 1

Erklärung der abhängigen Variable durch die Variablen i und j (ohne Intercept)

Modellanalyse

- `coef()` oder `coefficients()`:
Ausgabe der Koeffizientenschätzer
- `summary()`:
u.a. Ausgabe der 5-Punkte-Zusammenfassung für die Residuen, der Koeffizientenschätzer und ihrer t-Teststatistiken, des Bestimmtheitsmaßes und der Teststatistik des Overall-F-Tests
- `plot()`
Erzeugung von sog. Diagnoseplots (z.B. Residuenplots)
- ...

Gliederung

- 1 Das klassische lineare Modell
- 2 Lineare Modelle mit R
- 3 Varianzanalyse**
- 4 Umgang mit Faktoren

Varianzanalyse I

Gegeben: Lineares Modell M_1

Frage: Wird ein ausreichend großer Anteil der Gesamtstreuung der abhängigen Variable durch die Regression erklärt oder sind die Residuen zu hoch?

$$\begin{aligned} SQ_{Total} &= SQ_{Regression} + SQ_{Residual} \\ \sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (\hat{y}_i - y_i)^2 \end{aligned}$$

Lösung in R: `anova( $M_1$ )`

Varianzanalyse II

Gegeben: M_1 und M_2 , wobei M_2 nur auf Basis eines Teiles der Regressoren von M_1 gebildet wurde

Frage: Ist die Residualstreuung von M_2 annähernd gleich klein wie die von M_1 ?

$$SQ_{Residual}(M_2) = SQ_{Residual}(M_1) + SQ_{Residual}(M_2|M_1)$$

Lösung in R: `anova( $M_2, M_1$ )`

[Mehr zur Theorie des Tests: s. Vorlesung "lineare Modelle"]

Gliederung

- 1 Das klassische lineare Modell
- 2 Lineare Modelle mit R
- 3 Varianzanalyse
- 4 Umgang mit Faktoren**

Dummy- und Effektkodierung

Gegeben.: $k=1,\dots,K$ Kategorien ; K als Referenzkategorie

k	x_{i_1}	x_{i_2}	$\dots$	$x_{i_{k-1}}$
1	1	0	$\dots$	0
2	0	1	$0\dots$	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$K-1$	0	$\dots$	$\dots 0$	1
K	0	$\dots$	$\dots 0$	0

Dummykodierung

k	x_{i_1}	x_{i_2}	$\dots$	$x_{i_{k-1}}$
1	1	0	$\dots$	0
2	0	1	$0\dots$	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$K-1$	0	$\dots$	$\dots 0$	1
K	-1	$\dots$	$\dots -1$	-1

Effektkodierung

- Standard in R: Verwendung von Dummykodierung bei der Erzeugung linearer Modelle

Faktorisierung in R

- `as.factor(Variablen bzw. Objekt)`

Einfache Faktorisierung ohne Zusatzoptionen

[Referenzkategorie ist dabei automatisch die erste Kategorie]

- `factor(Variablen bzw. Objekt, levels = Faktorstufen,
labels = "Faktorbezeichnungen", ...)`

Faktorisierung mit Zusatzoptionen, wie z.B. der Benennung der einzelnen Faktorstufen

- `relevel(Faktorvariable, ref = "neue Referenzkategorie")`

Veränderung der Referenzkategorie