

Eine Einführung in R: Das Lineare Modell II

Bernd Klaus, Verena Zuber

Institut für Medizinische Informatik, Statistik und Epidemiologie (IMISE),
Universität Leipzig

13. Januar 2009

I. Modelldiagnose

II. Interpretation der Koeffizienten

Metrische erklärende Variablen

Kategoriale erklärende Variablen

Testen der Regressionskoeffizienten

Interaktionen

III. Prädiktion

I. Modelldiagnose

Wiederholung: Residuenanalyse

Frage: Sind die Voraussetzungen für das lineare Modell erfüllt?

Zu untersuchen sind:

1. Anpassung des Modells an die Daten:
→ Residuen gegen gefittete Wert \hat{Y}
2. Normalverteilung des Fehlers:
→ QQ-Plot: Quantile der Residuen gegen die theoretische NV
3. Homoskedastizität des Fehlers:
→ Standardisierte Residuen gegen gefittete Wert \hat{Y} ,
wenn die geeignet mit H standardisierten Residuen abhängig
von \hat{Y} sind, deutet dies auf ungleiche Varianzen der Fehler hin

Beispiele: Simulationen

```
h1<-seq(1,6,0.01)
```

```
X<-h1+rnorm(length(h1), mean=0, sd=0.1)
```

1. Kein linearer, sondern quadratischer Zusammenhang:

```
epsilon1<-rnorm(length(X), mean=0, sd=1)
```

```
Y1<-X*X+epsilon1
```

2. Kein Normal-, sondern gleichverteilter Fehler:

```
epsilon2<-runif(length(X), min=-1, max=1)
```

```
Y2<-X+epsilon2
```

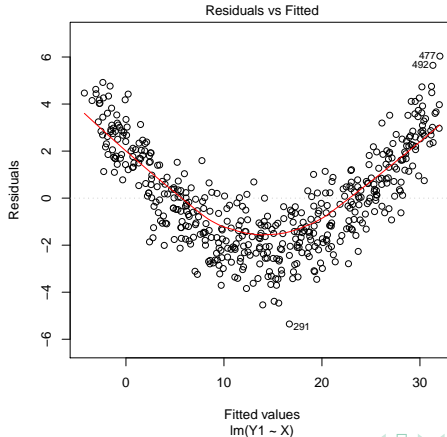
3. Die Fehler haben unterschiedliche Varianz,
bzw sind abhängig von Y :

```
epsilon3<-rnorm(length(X),  
mean=rep(0,length(X)), sd=X)
```

```
Y3<-X+epsilon3
```

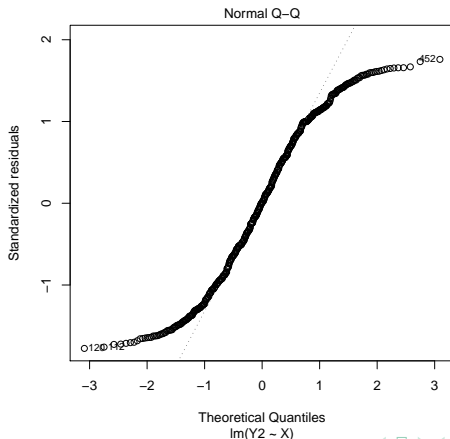
Modelldiagnose in R I: Residuen gegen gefittete Werte

- ▶ Residuen gegen gefittete Werte \hat{Y} zur Untersuchung der Anpassung des Modells an die Daten



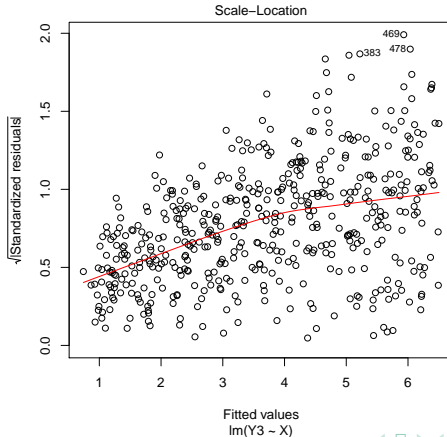
Modelldiagnose in R II: Residuen-QQ

- ▶ Plot der studentisierten gegen die theoretischen (NV) Residuen zur Untersuchung der Normalverteilung des Fehlers



Modelldiagnose in R III: Standardisierte Residuen gegen \hat{Y}

- ▶ Standardisierte, absolute Residuen gegen gefittete Werte \hat{Y} zur Untersuchung der Homoskedastizität des Fehlers



R^2 oder das adjustierte R^2

In der multiplen Regression wird zum R^2 meist auch das adjustierte R^2 ausgegeben:

$$R^2_{adjust} = R^2 - \frac{p - (1 - R)}{n - p - 1}$$

Dann können auch Modelle verglichen werden, die eine unterschiedliche Zahl p an Variablen besitzen.

Weitere Kriterien sind AIC, BIC, ...

II. Interpretation der Koeffizienten

Metrische erklärende Variablen

- ▶ Der Regressionskoeffizient β_j einer Variable X_j gibt deren Einfluss auf die Zielgröße Y unter gleichzeitiger Kontrolle der anderen Variablen an.
- ▶ Bei einer metrischen Variable X_j gilt:
(Gegeben die übrigen $(p - 1)$ Variablen werden festgehalten)
Wenn sich X_j um eine Einheit erhöht, so verändert sich Y um β_j Einheiten.
- ▶ Beispiel: Datensatz “airquality”

$$Ozone_i = -145.7 + 2.27847 \cdot Temp_i + 0.05711 \cdot Solar.R_i + \varepsilon_i$$

- ▶ “Wenn die Temperatur (bei gegebener Sonneneinstrahlung) um eine Einheit steigt, steigt die Ozonkonzentration um ca. 2.3 Einheiten”

Kategoriale erklärende Variablen

- ▶ Warnung: Kategoriale Variablen X_j können nicht wie metrische interpretiert werden!
- ▶ Vorgehensweise:
 1. Wähle eine Kategorie als Referenz
 2. Führe binäre Variablen (“Dummyvariablen”) ein, die angeben, ob eine Beobachtung in die Referenzkategorie oder in eine andere Kategorie fällt
 3. Wenn k Kategorien vorliegen, müssen $k - 1$ Dummyvariablen konstruiert werden
 4. Interpretation:
(Gegeben die übrigen $(p - 1)$ Variablen werden festgehalten)
Wenn eine Beobachtung nicht in die Referenzkategorie fällt, so verändert sich Y um β_j Einheiten
- ▶ Deswegen ist es in R essentiell, kategoriale Variablen als Faktoren zu führen (dann berechnet R die Dummyvariablen automatisch)

Beispieldaten: “Work”

Untersuchung verschiedener Einflussfaktoren (COMP, RTW, PVT) auf den prozentualen Anteil der Beschäftigten im öffentlichen Sektor DENS, die in einer Gewerkschaft organisiert sind, in verschiedenen amerikanischen Bundesstaaten.

- ▶ Metrische Variablen:
 - ▶ DENS: *Percent of public sector employees in unions, 1982*
 - ▶ PVT: *Percent of private sector employees in unions, 1982*
- ▶ Kategoriale Variablen:
 - ▶ COMP: *State bargaining laws cover public employees (1) or not (0)* (Referenzkategorie: Keine Rechte)
 - ▶ RTW: *State right-to-work law (1) or not (0)*

Zunächst ist folgendes lineares Modell von Interesse:

$$\text{DENS}_i = \beta_0 + \alpha_{RTW_i} + \beta_1 \cdot \text{PVT}_i + \varepsilon_i$$

Beispieldaten: “Work”

- ▶ `Work <- read.table(“Work.csv”, header = TRUE)`
- ▶ Umwandlung von RTW in einen Faktor
`Work$RTW <- as.factor(Work$RTW)`
- ▶ `test <- lm(DENS ~ RTW + PVT, data = Work)`

Ausgabe in R:

```
Coefficients:  
(Intercept)    RTW1      PVT  
35.3881      -10.8599    0.1418
```

Interpretation von α_{RTW} : Gibt es ein “Recht auf Arbeit”, so verringert sich der Anteil der im öffentlichen Dienst in einer Gewerkschaft organisierten Beschäftigten um ca. 11%

Testen der Regressionskoeffizienten

Der standardisierte Regressionskoeffizient ist t -verteilt mit einer Freiheitsgradzahl, die sich aus dem Stichprobenumfang n und der Variablenzahl p bestimmt:

$$T_j = \frac{\hat{\beta}_j}{\text{SD}(\hat{\beta}_j)} \sim t(n - p - 1)$$

$\text{SD}(\hat{\beta}_j)$: Standardabweichung von $\hat{\beta}_j$

- ▶ $H_0 : \beta = 0$ ablehnen, falls $|T_j| > t_{1-\alpha/2}(n - p - 1)$
- ▶ $H_0 : \beta > 0$ ablehnen, falls $T_j < t_\alpha(n - p - 1)$
- ▶ $H_0 : \beta < 0$ ablehnen, falls $T_j > t_{1-\alpha}(n - p - 1)$

R gibt in der `summary` sowohl die β 's (estimate), deren Standardabweichung (Std. Error) und t -Statistik (t value) und p -Wert (`Pr(>|t|)`) an.

Interaktionen

- ▶ Das Modell “test” besitzt nur ein R^2 von 0.25
- ▶ Wahrscheinlich sind wichtige Einflussfaktoren noch nicht berücksichtigt !
- ▶ Wir untersuchen daher das Modell:

$$DENS_i = \beta_0 + \alpha_{RTW_i} + \alpha_{COMP_i} + \alpha_{COMP_i * RTW_i} + \beta_1 \cdot PVT_i + \varepsilon_i$$

- ▶ Der Koeffizient $\alpha_{COMP_i * RTW_i}$ beschreibt eine multiplikative **Interaktion** der Faktoren COMP und RTW
- ▶ D.h. dieser Effekt besteht, wenn gleichzeitig Recht auf Arbeit UND Tarifverhandlungen der Gewerkschaftsmitglieder im öffentlichen Dienst gegeben sind.

Interaktionen - Umsetzung in R

- ▶ Das erweiterte Modell

$$DENS_i = \beta_0 + \alpha_{RTW_i} + \alpha_{COMP_i} + \alpha_{COMP_i * RTW_i} + \beta_1 \cdot PVT_i + \varepsilon_i$$

- ▶ Aufruf der Funktion `lm()`
- ▶ `testI <- lm(DENS ~ COMP*RTW + PVT, data = Work)`

Ausgabe in R:

```
Coefficients:  
(Intercept)      COMP1      RTW1      PVT      COMP1:RTW1  
  27.31371    14.92008   -0.58751    0.04727   -18.38713
```

Interaktionen - Interpretation der Ergebnisse

- ▶ COMP1: Gibt es ein Recht auf Tarifverhandlungen der Gewerkschaftsmitglieder im öffentlichen Dienst, aber KEIN Recht auf Arbeit, STEIGT der Anteil der öffentlich Beschäftigten, die gewerkschaftlich organisiert sind, um ca. 14.9%.
- ▶ RTW1: Gibt es ein Recht auf Arbeit, aber KEIN Recht auf Tarifverhandlungen der Gewerkschaftsmitglieder im öffentlichen Dienst, SINKT der Anteil der öffentlich Beschäftigten, die gewerkschaftlich organisiert sind, marginal um 0.6%.
- ▶ COMP1 :RTW1: Gibt es aber sowohl ein Recht auf Arbeit, als auch auf Tarifverhandlungen der Gewerkschaftsmitglieder im öffentlichen Dienst, so SINKT der Anteil der öffentlich Beschäftigten, die gewerkschaftlich organisiert sind, um ca. $18.4\% + 0.6\% - 14.9\% = 3.5\%$

III. Prädiktion

Vorhersage im Linearen Modell

- ▶ Gegeben: Lineares Modell mit Regressionskoeffizienten, die auf Basis bestehender Daten ermittelt wurden
- ▶ Neu: Eine neue Beobachtung X_{n+1} deren Zielgröße Y_{n+1} unbekannt ist
- ▶ Ziel: Vorhersage der unbekanntenen Zielgröße Y_{n+1}

Vorgehensweise zur Vorhersage:

- ▶ Bilde eine Vorhersageregeln (*prediction rule*) aus dem gegebenen Modell
- ▶ Setze die Werte der neuen Beobachtung X_{n+1} in diese Vorhersageregeln ein und berechne die Vorhersage \hat{Y}_{n+1}

Beispiel: *Airquality*-Daten

- ▶ Gegeben: Lineares Modell mit Regressionskoeffizienten aus dem Datensatz “airquality”

$$\text{Ozone}_i = -145.7 + 2.27847 \cdot \text{Temp}_i + 0.05711 \cdot \text{Solar.R}_i + \varepsilon_i$$

- ▶ Neu: 3 neue Beobachtungen X_{n+1} *newdata*, deren Zielgröße Y_{n+1} unbekannt sind

Ozone	Solar.R	Temp
?	80	110
?	80	112
?	80	114

- ▶ Ziel: Vorhersage der unbekanntenen Zielgröße Y_{n+1}

Beispiel: *Airquality*-Daten

- ▶ Berechnung der Vorhersageregeln `air`:
`air <- lm(formula= Ozone ~ Temp + Solar.R, data= airquality)`
- ▶ Vorhersage der Temperatur für `newdata` mit Hilfe des Modells `air` mit dem R-Aufruf: `predict(air,newdata)`
- ▶ Dies ergibt folgende Vorhersagen:

Ozone	Solar.R	Temp
109.4970	80	110
114.0539	80	112
118.6108	80	114