Eine Einführung in R: Das Lineare Modell

Bernd Klaus, Verena Zuber

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

6. Januar 2009

I. Lineare Einfachregression

Einleitung

- Ziel der Regressionsanalyse:
  Welchen Einfluss hat eine Größe X auf eine andere Zufallsvariable Y?
  - Y : metrische Zielvariable, zu erklärende Variable, Regressand
  - X : erklärende Variable, Regressor (zufällig oder deterministisch)
- Daten:
  n Realisierungen $(y_1, x_1), \ldots, (y_n, x_n)$

Die Lineare Regression untersucht, ob ein linearer Zusammenhang zwischen X und Y besteht.

Modell der Linearen Regression

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i = 1, \ldots, n$$

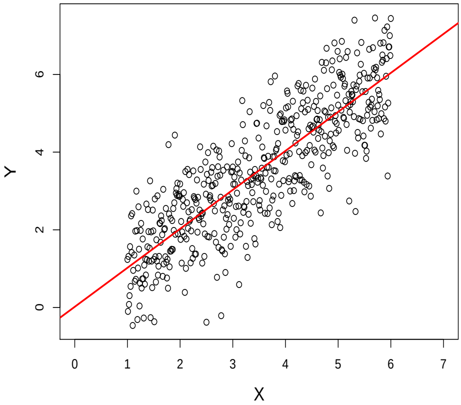
- y_i : Zielvariable, zu erklärende Variable, Regressand
- x_i : erklärende Variable, Regressor
- ε_i : unbeobachtbare Fehlervariable, unabhängig und identisch verteilt (in der Regel als $N(0, \sigma)$)
- zu schätzende Koeffizienten des Models: β_0, β_1
- β_0 : Intercept
- β_1 : Regressionskoeffizient der Variable X

Annahmen: Lineare Regression

- Es besteht ein linearer Zusammenhang zwischen X und Y
- Y ist metrisch und normalverteilt
  (Kategorial: Logit Regression; Allgemeinere Verteilungen: GLM's)
  - $E(y_i) = \beta_0 + \beta_1 x_i$
  - $Var(y_i) = \sigma^2$
- Homoskedastizität, d.h. die Fehler ε_i haben die gleiche Varianz:
  $Var(\varepsilon_i) = \sigma^2$ für alle $i = 1, ..., n$
- Die Fehler ε_i, mit $i = 1, ..., n$, sind unabhängig
  (GegenBsp: Zeitreihendaten)
- Die Fehler ε sind unabhängig vom Wert der Zielvariable Y

Beispiel: Simulierte Daten

```
X<-seq(1,6,0.01)
epsilon<-rnorm(length(X), mean=0, sd=1)
Y<-X+epsilon
```



Schätzung der β_i

β_0 und β_1 können durch Minimierung der Summe des Quadratischen Fehlers geschätzt werden (Kleinste Quadrate Schätzers):

$$\mathrm{MLQ} = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2 \rightarrow \min!$$

Dies führt zu folgenden Schätzungen für β_0 und β_1:

$$\beta_0 = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)$$

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Testen des β-Koeffizienten

Der Regressionskoeffizient β_1 der Variable X ist ein Indikator für den linearen Zusammenhang von X und Y. Es gilt:

$$\beta_1 = cor(X, Y)\frac{\sigma_Y}{\sigma_X}$$

Daraus folgt:

- $\beta_1 < 0$: negativer (linearer) Zusammenhang
- $\beta_1 = 0$: kein (linearer) Zusammenhang
- $\beta_1 > 0$: positiver (linearer) Zusammenhang

Es gibt einen einfachen Test, der angibt, ob β_1 signifikant ungleich Null ist, d.h. ob ein signifikanter Zusammenhang zwischen X und Y besteht.

Zerlegung der Gesamtstreuung

Die Maßzahl R^2 dient als Hinweis darauf, wie gut ein Regressionsmodell zu den Daten passt. Die Idee hinter diesem Maß ist die sogenannte Streuungszerlegung:

$$\text{SQT} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{SQR}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{SQE}}$$

- SQT: Sum of Squares Total, die Gesamtstreuung ($Var(Y)$)
- SQE: Sum of Squares Explained, die durch das Modell erklärte Streuung
- SQR: Sum of Squares Residuals, die Rest- oder Residualstreuung

Bestimmtheitsmaß R^2

Liegen die Punkte $(y_1, x_1), \ldots, (y_n, x_n)$ alle auf einer Geraden, so ist SQR$= 0$ und die Gesamtstreuung wäre gleich der erklärten Streuung. Wir definieren

$$R^2 = \frac{\text{SQE}}{\text{SQT}} = 1 - \frac{\text{SQR}}{\text{SQT}} \quad \in [0, 1]$$

Je größer also das R^2 ist, desto besser passt das Modell zu den Daten. Dabei bedeuten:

- $R^2 = 0$: Die erklärte Streuung ist 0, d.h. das Modell ist extrem schlecht; X und Y sind nicht linear abhängig
- $R^2 = 1$: Die erklärte Streuung entspricht der Gesamtstreuung, das Modell passt perfekt
- $R^2 \in (0, 1)$: Es gibt unerklärte Streuung. Als Faustregel für ein passables Modell kann ein $R^2 \geq 0.4$ gelten

II. Multiple Regression

Mehrere erklärende Variablen

- Ziel: Ist man an dem Einfluss mehrerer Variablen $X_1, ..., X_p$ auf eine Zielgröße Y, kann man in der multiplen linearen Regression p erklärende Größen $X = X_1, ..., X_p$ aufnehmen.
- Realisierungen: $(y_1, x_{11}, ..., x_{1p}), \ldots, (y_n, x_{n1}, ..., x_{np})$
- Modell:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i$$

$$Y = X\beta + \varepsilon$$

  Dabei ist $X = (x_{ij})$ die sogenannte Designmatrix.
- Vorteil zur einfachen Regression:
  β_j beschreibt den Zusammenhang der j.ten Variable zu Y bedingt auf alle übrigen $j - 1$ Variablen (Kontrolle von ungewollten oder Scheineffekten)

Least-Squares Schätzer

$\beta_0, \beta_1, ..., \beta_p$ können (analog zur einfachen linearen Regression) durch Minimierung der Summe des Quadratischen Fehlers geschätzt werden (Kleinste Quadrate oder Least-Squares):

$$\text{MLQ} = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{pi})\right)^2 \rightarrow \text{min!}$$

Der Least-Squares Schätzer ergibt sich nach Umformen zu:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Hat-Matrix

Die Matrix

$$H := (X^T X)^{-1} X^T$$

bezeichnet man auch als "Hat"-Matrix, da sie die beobachteten Daten Y in geschätzte Werte $\hat{Y} = HY$ verwandelt.
Es gilt für den Residuenvektor:

$$r = \hat{Y} - Y = HY - Y = (I_n - H)Y$$
$$r \sim \; N\big(0, (I_n - H)\sigma\big)$$

Die Residuen besitzen also die Varianz / Kovarianz

$$\mathrm{Var}(r_i) = \sigma^2(1 - h_{ii}) \qquad \text{und} \qquad \mathrm{Cov}(\hat{r}_i, \hat{r}_j) = -\sigma^2(1 - h_{ij}) \; , i \neq j$$

Residuenanalyse

Da die Residuen alle unterschiedliche Varianz besitzen, skaliert man sie auf einheitliche Varianz:

$$r_{i,\text{stud}} = \frac{r_i}{\hat{\sigma} \cdot \sqrt{1 - h_{ii}}} \sim N(0, \sigma)$$

Frage: Sind die Voraussetzungen für das lineare Modell erfüllt?
Zu untersuchen sind:

1. Anpassung des Modells an die Daten:
   $\rightarrow$ Residuen gegen gefittete Wert $\hat{Y}$
2. Normalverteilung des Fehlers:
   $\rightarrow$ QQ-Plot: Quantile der Residuen gegen die theoretische NV
3. Homoskedastizität des Fehlers:
   $\rightarrow$ 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

III. Umsetzung in R

Beispieldaten: “airquality”

- `Ozone`: *Mean ozone in parts per billion from 1300 to 1500 hours at Roosevelt Island*
- `Solar.R`: *Solar radiation in Langleys in the frequency band 4000-7700 Angstroms from 0800 to 1200 hours at Central Park*
- `Wind`: *Average wind speed in miles per hour at 0700 and 1000 hours at LaGuardia Airport*
- `Temp`: *Maximum daily temperature in degrees Fahrenheit at La Guardia Airport*

Mit diesen Daten kann untersucht werden welchen Einfluss Sonneneinstrahlung, Wind und Temperatur auf die Ozonwerte haben.

Beispiel in R

Wir laden den Datensatz "airquality"

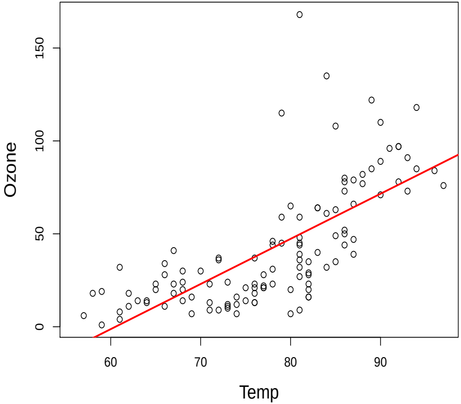
- `data("airquality")`
- Wir untersuchen das Modell:
- $\texttt{Ozone}_i = \beta_0 + \beta_1 \cdot \texttt{Temp}_i + \varepsilon_i$
- ... also die Abhängigkeit des Ozons von der Temperatur
- Aufruf der Funktion `lm()`
- `test <- lm( formula= Ozone ~ Temp, data= airquality)`

Ausgabe in R:

```
 Coefficients:
(Intercept)    Temp
-146.995       2.429
```

Scatterplot: Ozone ∼ Temp

```
plot(Temp,Ozone)
abline(test$coefficients, col="red")
```



Modelldiagnose

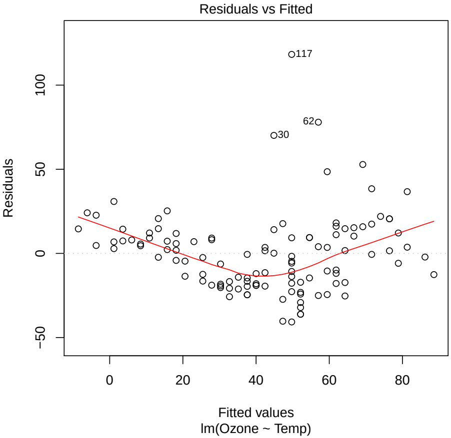
- R^2 und andere Maße des Modells : `summary(test)`

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-146.9955	18.2872	-8.038	9.37e-13
Temp	2.4287	0.2331	10.418	< 2e-16

- Koeffizienten: `test$coefficients`
- Gefittete Werte $\hat{Y}$: `test$fitted.values`
- Studentisierte Residuen: `ls.diag(test)$std.res`
- Hat-Matrix: `ls.diag(test)$hat`
- Verschiedene Diagnoseplots: `plot(test)` oder `plot.lm(test)` (u.a. Residuenanalyse)

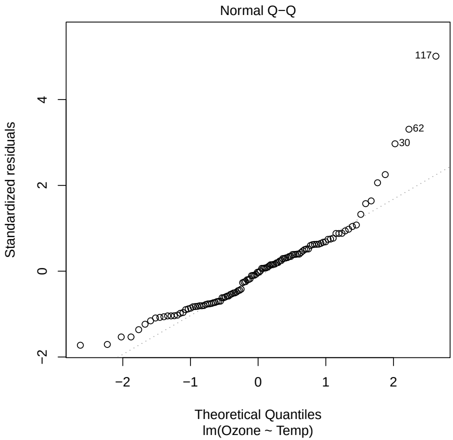
Modelldiagnose in R I: Residuen gegen gefittete Werte

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



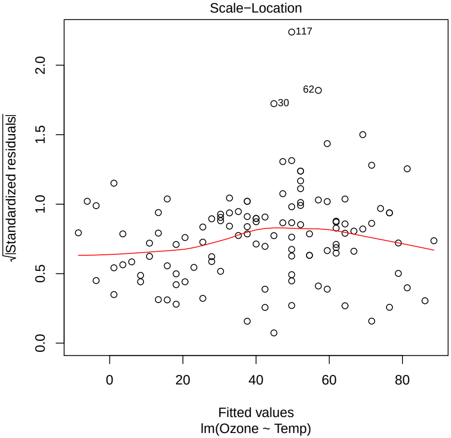
Modelldiagnose in R II: Residuen-QQ

- 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



Multiple Regression in R

- Wir untersuchen nun das Modell:
- $\texttt{Ozone}_i = \beta_0 + \beta_1 \cdot \texttt{Temp}_i + \beta_2 \cdot \texttt{Solar.R}_i + \varepsilon_i$
- ... also die Abhängigkeit des Ozons von der Temperatur und der Sonneneinstrahlung
- Aufruf der Funktion `lm()`
- `test <- lm( formula= Ozone ~ Temp + Solar.R, data= airquality)`

Ausgabe in R:

```
 Coefficients:
(Intercept)    Temp       Solar.R
-145.70316     2.27847    0.05711
```

Spezifikation der Regressionsvariablen

`lm(formula, ...)`

- `formula`: Hier muss das Modell bzw die Variablen des Modelles spezifiziert werden.

- Allgemeiner Aufbau der linearen Einfachregression
  `formula= Y~X`
- Beispiel: `formula= Ozone ~ Temp`
- Allgemeiner Aufbau der multiplen linearen Regression
  `formula= Y~` $X_1 + X_2 + \ldots + X_p$
- Beispiel: `formula= Ozone ~ Temp + Solar.R`