

Eine Einführung in R: Lineare Regression

Katja Nowick, Markus Kreuz
(basierend auf Vorarbeiten von Verena Zuber und Bernd Klaus)

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

http://www.nowick-lab.info/?page_id=365

13. Januar 2015

I. Ergänzungen zu Übung 2

- `chisq.test`: χ^2 -Test
- `fisher.test`: Fisher-Test
- `binom.test`: Binomial-Test
- `cor.test`: Korrelationstest
- `kruskal.test`: Kruskal-Wallis-Test
- `ks.test`: Kolmogorov-Smirnov-Test
- `shapiro.test`: Shapiro-Wilk-Test

Wenn die theoretische Verteilung der interessierenden Statistik nicht bekannt ist, können Bootstrapverfahren eingesetzt werden. Mögliche Anwendungen:

- Bootstrap Konfidenzintervalle
- Bootstrap Tests

Vorgehen:

Aus der Originalstichprobe werden B Bootstrap-Stichproben der gleichen Größe mit zurücklegen gezogen: $x_b = (x_1^*, \dots, x_n^*)$, $b = 1, \dots, B$. Dies entspricht einer Ziehung aus der empirischen Verteilungsfunktion. Für jede der B Stichproben kann die interessierende Statistik T berechnet werden. Dies ermöglicht die Abschätzung der Verteilung von T und damit die Schätzung von Quantilen und p-Werten.

Bootstrap Beispiel: Konfidenzintervall

```
x<-rnorm(100)
mean(x)
```

Fragestellung: Bestimme das 95% Konfidenzintervall für die Schätzung des Mittelwertes. `t<-rep(NA,N)`

```
for (i in 1:N){
t[i]<-mean(sample(x,length(x),replace=T))
}
quantile(t,c(0.05,0.95))
```

Lineare Einfachregression

- **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 $(y_1, x_1), \dots, (y_n, x_n)$

Ziel der linearen Regression

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

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

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

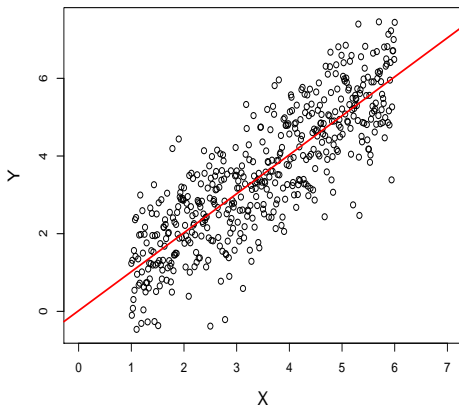
Für $i = 1, \dots, n$ Beobachtungen:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, \dots, n$$

- 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 $i = 1, \dots, n$
- Die Fehler ε_i , mit $i = 1, \dots, n$, 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 Schätzer:

MLQ

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

Dies führt zu folgenden Schätzungen für β_0 , β_1 und der gefitteten Wert \hat{Y} (Regressionsgerade):

Schätzungen

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\text{cov}(X, Y)}{\text{var}(X)}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

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

Zusammenhang zwischen β_1 und $cor(X, Y)$

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

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 Streuungszersetzung:

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

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

Liegen die Punkte $(y_1, x_1), \dots, (y_n, x_n)$ **alle auf einer Geraden**, so ist **SQR** = 0 und die Gesamtstreuung wäre gleich der erklärten Streuung. Das Bestimmtheitsmaß R^2 ist gegeben durch:

Zerlegung des R^2

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

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

Multiple Regression

- **Fragestellung:** Wie ist der Einfluss mehrerer Variablen X_1, \dots, X_p auf eine Zielgröße Y ?
- **Realisierungen:** $(y_1, x_{11}, \dots, x_{1p}), \dots, (y_n, x_{n1}, \dots, x_{np})$
- Modell der **multiplen linearen Regression** mit p erklärenden Größen $X = X_1, \dots, X_p$:

Modell der multiplen linearen Regression

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

$$y_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} + \varepsilon_i \quad i = 1, \dots, n, j = 1, \dots, p$$

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)

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

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

Der Least-Squares Schätzer ergibt sich nach Umformen zu:

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

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:**

→ 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

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.

Wir laden den Datensatz “airquality”

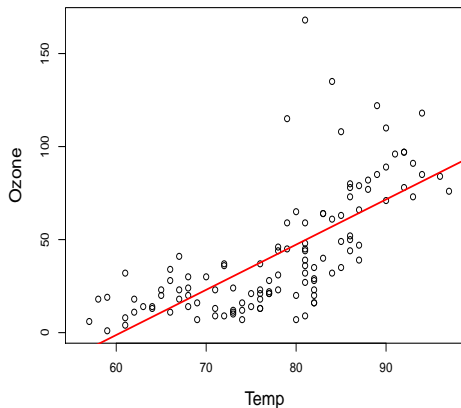
- `data(“airquality”)`
- Wir untersuchen das Modell:
- $\text{Ozone}_i = \beta_0 + \beta_1 \cdot \text{Temp}_i + \varepsilon_i$
- ... also die Abhängigkeit des Ozons von der Temperatur
- Aufruf der Funktion `lm()`
- `test <- lm(formula= Ozone ~ Temp, data= airquality)`
- test ist ein Objekt der Klasse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

Ausgabe in R:

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

Scatterplot: Ozone ~ Temp

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



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

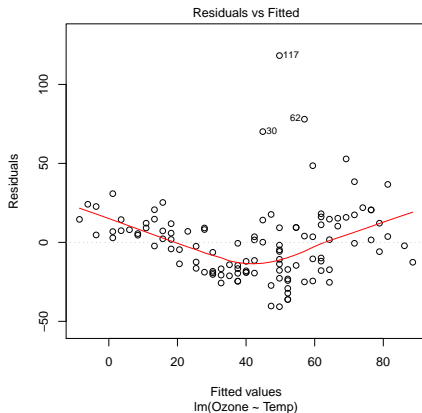
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-146.9955	18.2872	-8.038	9.37e-13
Temp	2.4287	0.2331	10.418	< 2e-16

Multiple R-squared: 0.4877, Adjusted R-squared: 0.4832

- 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)

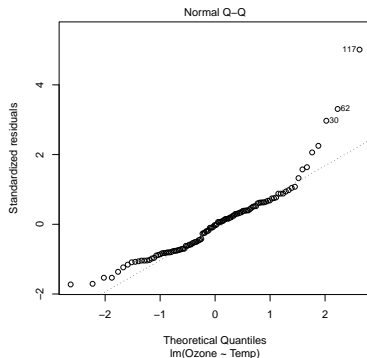
Modelldiagnose in R I: Residuen gegen gefittete Werte

- Residuen gegen gefittete Werte \hat{Y} zur Untersuchung der Anpassung des Modells an die Daten
- Keine systematische Abweichung, z.B. Trend oder U-Form

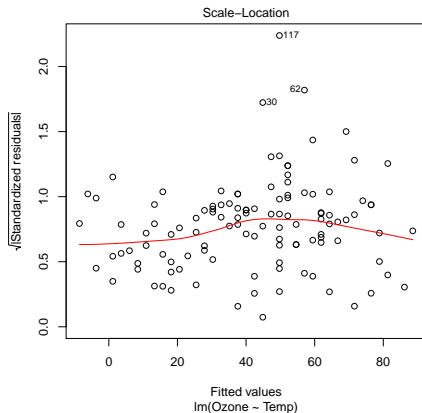


Modelldiagnose in R II: Residuen-QQ

- Plot der studentisierten (besondere Standardisierung) gegen die theoretischen (NV) Residuen zur Untersuchung der Normalverteilung des Fehlers
- Wenn die Residuen normalverteilt sind, sollten sie auf der gestrichelten Geraden liegen



- Standardisierte, absolute Residuen gegen gefittete Werte \hat{Y} zur Untersuchung der Homoskedastizität des Fehlers
- Keine systematische Abweichung, z.B. ansteigende Varianz



Multiple Regression in R

- Wir untersuchen nun das Modell:
- $\text{Ozone}_i = \beta_0 + \beta_1 \cdot \text{Temp}_i + \beta_2 \cdot \text{Solar.R}_i + \varepsilon_i$
- ... also die Abhängigkeit des Ozons von der Temperatur und der Sonneneinstrahlung
- Aufruf der Funktion `lm()`
- `model2 <- lm(formula= Ozone ~ Temp + Solar.R, data= airquality)`

Ausgabe in R:

Coefficients:

(Intercept)	Temp	Solar.R
-145.70316	2.27847	0.05711

Ausgabe von `summary(model2)`:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-145.70316	18.44672	-7.899	2.53e-12
Temp	2.27847	0.24600	9.262	2.22e-15
Solar.R	0.05711	0.02572	2.221	0.0285

Multiple R-squared: 0.5103, Adjusted R-squared: 0.5012

Interpretation:

- Solar.R besitzt ein β , das signifikant von Null verschieden ist (p Wert $0.0285 < 0.05$)
- Das β der Variable Temp verändert sich nur leicht durch die Aufnahme von Solar.R: von 2.4287 zu 2.27847
- Das R^2 wird durch die Aufnahme von Solar.R nur noch leicht verbessert: von 0.4832 zu 0.5012
- Durch die beiden Variablen Solar.R und Temp kann die Hälfte der Streuung der Ozonmessungen erklärt werden.

```
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~ X1 + X2 + ... + Xp`
- Beispiel: `formula= Ozone ~ Temp + Solar.R`