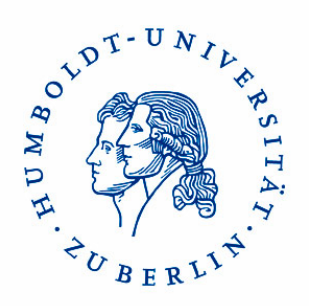
Stochastik–Praktikum
Lineare Modelle

Thorsten Dickhaus

Humboldt-Universität zu Berlin

06.10.2010



Übersicht

Übersicht

Modell der einfachen linearen Regression

Zielgröße (response): Y, Einflussgröße (erklärende Variable): X
Modellbildung: Lineare Abhängigkeit zwischen Y und X,
Beobachtungen gestört durch additives i. i. d. Rauschen
$\varepsilon_i \sim \mathbf{N}(0, \sigma^2)$ mit konstanter Varianz:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \ldots, n.$$

Kleinste–Quadrate–Schätzung und ML–Schätzung:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum_i (x_i - \bar{x}_n)^2}, \quad \hat{\beta}_0 = \bar{y}_n - \hat{\beta}_1 \bar{x}_n.$$

Satz (Gauß–Markov)

Der KQ–Schätzer ist bester linearer unverzerrter Schätzer und eindeutig.

Residualvarianz (erwartungstreue Form):

$$s^2 = \hat{\sigma}^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-2}$$

Geschätzte Standardabweichungen der geschätzten Regressionskoeffizienten:

$$\hat{SE}(\hat{\beta}_0) = \frac{s}{\sqrt{n}} \sqrt{\frac{\sum_i x_i^2}{SSX}}, \hat{SE}(\hat{\beta}_1) = \frac{s}{\sqrt{n}} \sqrt{SSX^{-1}}, SSX = \sum_{i=1}^{n} (x_i - \bar{x}_n)^2.$$

$\hat{\beta}_i / \hat{SE}(\beta_i)$, $i = 0, 1$ (als Zufallsvariable aufgefasst) ist t_{n-2}–verteilt $\Rightarrow$ Konfidenzintervalle:

$$I = \hat{\beta}_i \pm t_{n-2;1-\alpha/2} \cdot \hat{SE}(\hat{\beta}_i), \ i = 0, 1.$$

Ein Test für die Globalhypothese $H_0 : \beta_0 = \beta_1 = 0$ kann basierend auf der folgenden F-Statistik durchgeführt werden:

$$F = \frac{(n-2)\sum_i(\hat{Y}_i - \bar{Y})^2}{\sum_i(Y_i - \hat{Y}_i)^2} \underset{H_0}{\sim} F_{1,n-2}.$$

Bestimmtheitsmaß:

$$R^2 = 1 - \frac{\sum_i(y_i - \hat{y}_i)^2}{\sum_i(y_i - \bar{y})^2} = 1 - \frac{s^2}{(n-2)SSY}$$

Anteil der durch das Modell erklärten Varianz der Zielgröße

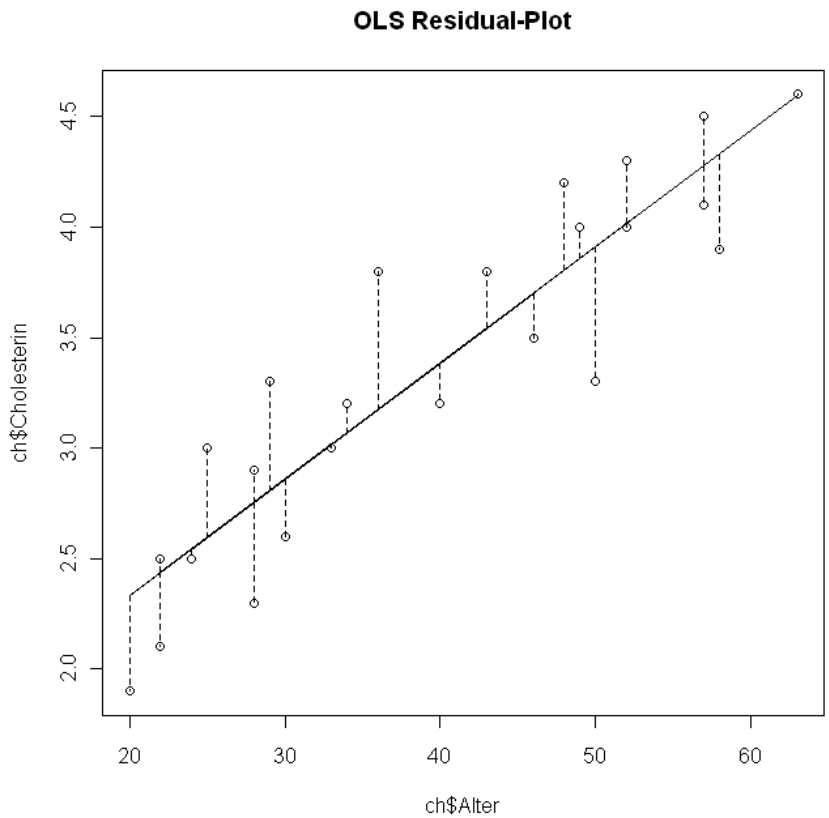
R–Code: lineare Regression Cholesterin–Beispiel

```
> ch<-read.table(file="cholesterin.csv",header=TRUE,dec=",")

> summary(lm(ch$Cholesterin~ch$Alter))

Call:
lm(formula = ch$Cholesterin ~ ch$Alter)

Residuals:
     Min       1Q   Median       3Q      Max
-0.6111 -0.2151 -0.0058  0.2297  0.6256

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.279868   0.215699   5.934 5.69e-06 ***
ch$Alter    0.052625   0.005192  10.136 9.43e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.334 on 22 degrees of freedom
Multiple R-squared: 0.8236,     Adjusted R-squared: 0.8156
F-statistic: 102.7 on 1 and 22 DF,  p-value: 9.428e-10
```

```
> confint(lm(ch$Cholesterin~ch$Alter))
                2.5 %     97.5 %
(Intercept) 0.83253668 1.72720003
ch$Alter    0.04185806 0.06339175
```

Residualanalyse

```
> plot(ch$Alter,ch$Cholesterin,main="OLS Residual-Plot")
> lines(ch$Alter,fitted(lm(ch$Cholesterin~ch$Alter)))
> segments(ch$Alter,fitted(lm(ch$Cholesterin~ch$Alter)),
+ ch$Alter,ch$Cholesterin,lty=2)
```



Übersicht

Zielgröße hängt linear von $p \geq 1$ erklärenden Variablen $X_1, \ldots, X_p$ ab. Mit Designmatrix X ergibt sich Modellgleichung:

$$(Y_i)_{i=1}^n =: Y = X\beta + \varepsilon = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Kleinste–Quadrate–Schätzer: $\hat{\beta} = (X^t X)^{-1} X^t Y$.

Beste Lineare Vorhersage (in X) für Y:

$$\hat{Y} = X\hat{\beta} = X\left(XX^t\right)^{-1} X^t Y =: HY$$

mit der sogenannten Hut–Matrix H.

Mit den Residuen $\hat{\varepsilon} = Y - \hat{Y}$ folgt für die Residualvarianz:

$$s^2 = \hat{\sigma}^2 = \frac{\hat{\varepsilon}^t \hat{\varepsilon}}{n - p - 1} .$$

Geschätzte Standardabweichung der geschätzten Regressionskoeffizienten:

$$\hat{SE}(\hat{\beta}_i) = \sqrt{\frac{s^2}{(X^t X)_{ii}}}$$

Die Hypothese $\beta_i = 0 \; \forall i \in \{1, \ldots, p\}$ kann basierend auf der folgenden F-Statistik geprüft werden:

$$F = \frac{\left((Y - \bar{Y})^t (Y - \bar{Y}) - \hat{\varepsilon}^t \hat{\varepsilon} \right) / p}{\hat{\varepsilon}^t \hat{\varepsilon} / (n - p - 1)} \underset{H_0}{\sim} F_{p, n-p-1}$$

Mit den Residuen $\hat{\varepsilon} = Y - \hat{Y}$ folgt für die Residualvarianz:

$$s^2 = \hat{\sigma}^2 = \frac{\hat{\varepsilon}^t \hat{\varepsilon}}{n - p - 1} .$$

Geschätzte Standardabweichung der geschätzten Regressionskoeffizienten:

$$\hat{SE}(\hat{\beta}_i) = \sqrt{\frac{s^2}{(X^t X)_{ii}}}$$

Für einzelne Koeffizienten kann die Hypothese $H_j : \beta_j = 0$ mit einem t–Test untersucht werden:

$$T_j = \frac{\hat{\beta}_j}{\hat{SE}(\hat{\beta}_j)} \underset{H_j}{\sim} t_{n-p-1}.$$

Übersicht

Einfaktorielle Varianzanalyse

ANOVA ist eng verwandt mit linearem Regressionsmodell.

Gegensatz zur linearen Regression:
Erklärende Variablen sind dichotom bzw. kategoriell.

Einfaktorielle Varianzanalyse (ANOVA1):
Untersuche Einfluss eines Faktors auf die Zielgröße.

Das Modell ist bestimmt durch:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \sim \mathbf{N}\left(\mu_i, \sigma^2\right), \mu_i = \mu + \alpha_i.$$

Dabei indiziert $i = 1, \ldots, p$ die Faktorgruppen und $j = 1, \ldots, n_i$ die unabhängigen Wiederholungen in den Faktorgruppen.

Einfaktorielle Varianzanalyse

Ziel der Varianzanalyse:
Test auf Gleichheit der Mittelwerte in den p Faktorgruppen, also $H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_p$ versus $H_1 : \complement H_0$.

Vorgehen:

1. Schätze Varianzen in den Gruppen.
2. Schätze Varianzen zwischen den Gruppen.
3. Schätze totale Varianz.
4. Je größer der Anteil der Varianz zwischen den Gruppen an der Gesamtvarianz im Vergleich zur Varianz innerhalb der Gruppen, desto mehr Evidenz für H_1 liegt vor.

Einfaktorielle Varianzanalyse

Ziel der Varianzanalyse:

Test auf Gleichheit der Mittelwerte in den p Faktorgruppen, also $H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_p$ versus $H_1 : \complement H_0$.

Vorgehen:

1. Schätze Varianzen in den Gruppen.
2. Schätze Varianzen zwischen den Gruppen.
3. Schätze totale Varianz.
4. Je größer der Anteil der Varianz zwischen den Gruppen an der Gesamtvarianz im Vergleich zur Varianz innerhalb der Gruppen, desto mehr Evidenz für H_1 liegt vor.

Einfaktorielle Varianzanalyse

Ziel der Varianzanalyse:

Test auf Gleichheit der Mittelwerte in den p Faktorgruppen, also $H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_p$ versus $H_1 : \complement H_0$.

Vorgehen:

1. Schätze Varianzen in den Gruppen.
2. Schätze Varianzen zwischen den Gruppen.
3. Schätze totale Varianz.
4. Je größer der Anteil der Varianz zwischen den Gruppen an der Gesamtvarianz im Vergleich zur Varianz innerhalb der Gruppen, desto mehr Evidenz für H_1 liegt vor.

Einfaktorielle Varianzanalyse

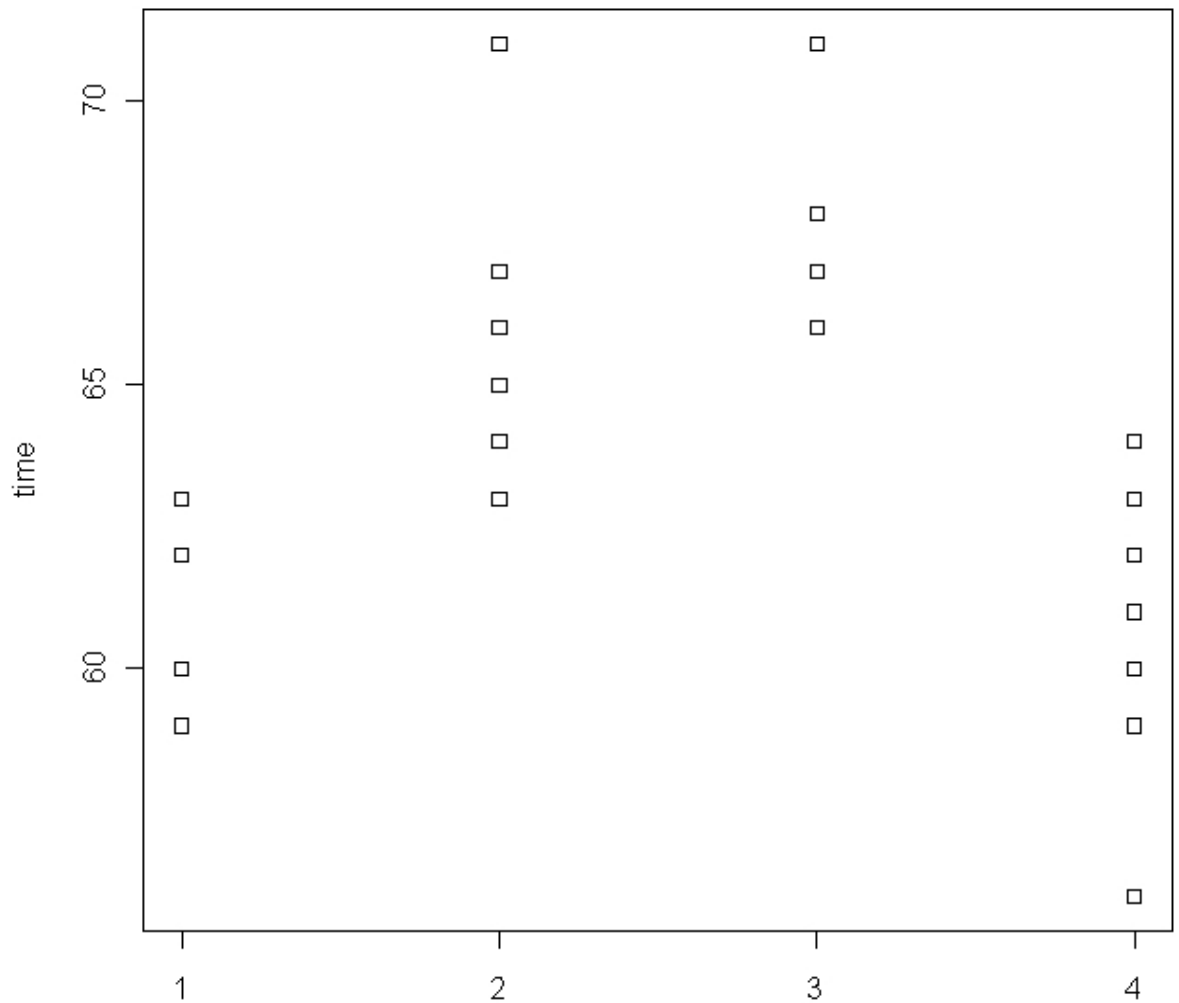
Ziel der Varianzanalyse:

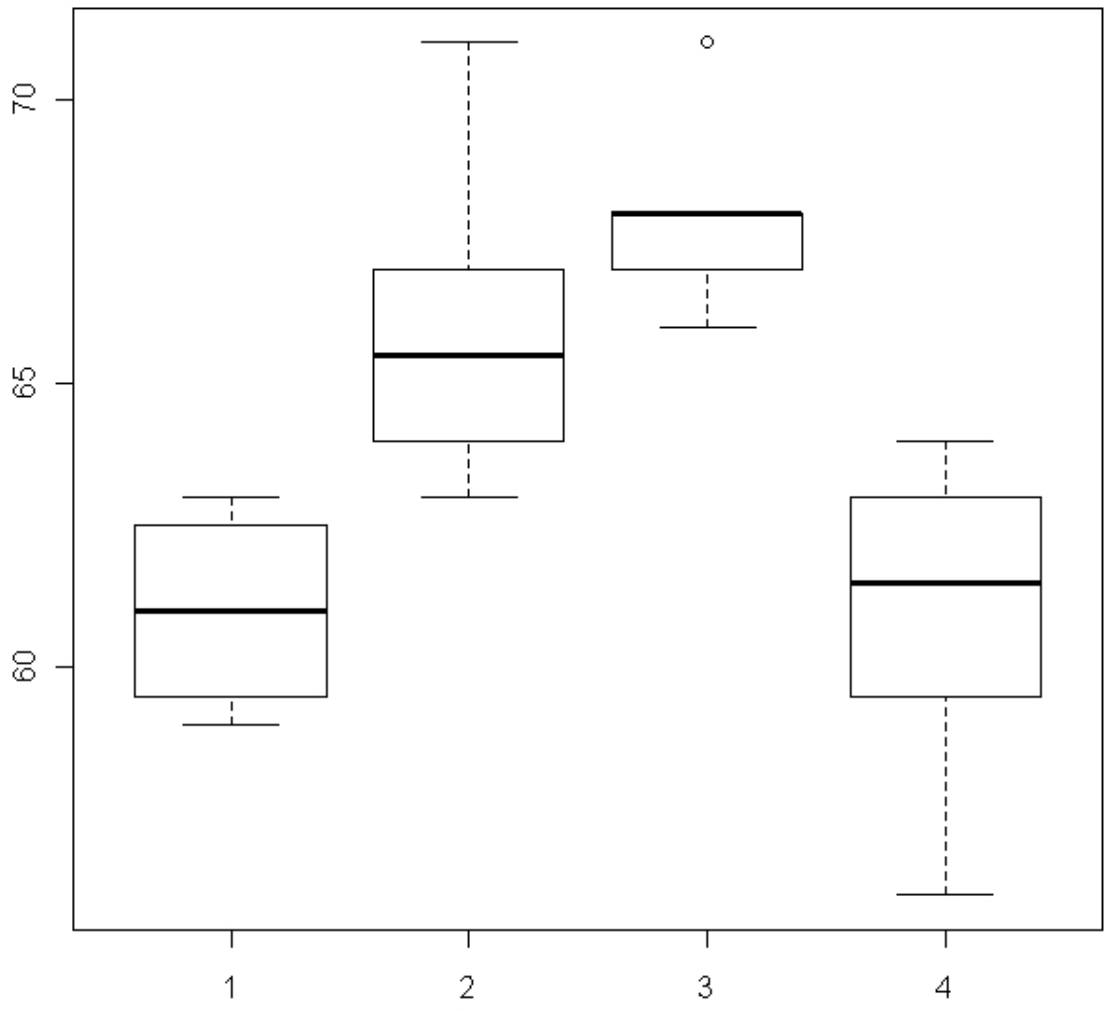
Test auf Gleichheit der Mittelwerte in den p Faktorgruppen, also $H_0 : \alpha_1 = \alpha_2 = \ \ldots \ = \alpha_p$ versus $H_1 : \complement H_0$.

Vorgehen:

1. Schätze Varianzen in den Gruppen.
2. Schätze Varianzen zwischen den Gruppen.
3. Schätze totale Varianz.
4. Je größer der Anteil der Varianz zwischen den Gruppen an der Gesamtvarianz im Vergleich zur Varianz innerhalb der Gruppen, desto mehr Evidenz für H_1 liegt vor.

Beispiel: Diäten und Blutgerinnung




70
65
60
1
2
3
4

ANOVA–Tafel

Quelle der Varianz	Abk.	Berechnung	Freiheitsgrade
Faktor	SS_F	$SS_F = \sum_{i=1}^{p} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$	$p-1$
Gruppen	SS_E	$SS_E = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2$	$n-p$
total	SS_T	$SS_T = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{\cdot\cdot})^2$	$n-1$

Der (Gesamt–)Mittelwert $\bar{y}_{\cdot\cdot}$ und die Gruppenmittelwerte $\bar{y}_{i\cdot}$, $i = 1, \ldots, p$, sind ML–Schätzer.

Die Teststatistik für das Testen von H_0 ist

$$F = \frac{SS_F/(p-1)}{SS_E/(n-p)} \underset{H_0}{\sim} F_{p-1,n-p}.$$

Dabei werden wieder SS_F und SS_E als Zufallsvariablen aufgefasst.

Varianzanalyse liefert:

Test zur Überprüfung der Hypothese, dass alle Gruppenmittelwerte gleich sind.

Oft auch von Interesse:

- Alle paarweisen Vergleiche der Gruppenmittelwerte (MCA)
- Alle Vergleiche von Gruppenmittelwerten mit dem empirisch größten (MCB)
- Alle Vergleiche von Gruppenmittelwerten mit einer Referenzgruppe (MCC)

Dies sind klassische Multiple Testprobleme!
(vgl. z. B. Buch von Jason Hsu (1996)

Streiflicht: Zweifaktorielle Varianzanalyse

Modellgleichung der zweifaktoriellen Varianzanalyse:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}.$$

Fehler ε_{ijk}: Normalverteilt mit konstanter Varianz.

γ_{ij} ist Interaktionsterm, der Wechselwirkungen zwischen Faktoren modelliert.

$n_{i\cdot}$, $i \in \{1, \ldots, I\}$ Beobachtungen in Faktor–1–Gruppen,
$n_{\cdot j}$, $j \in \{1, \ldots, J\}$ Beobachtungen in Faktor–2–Gruppen.

Quelle der Varianz	Abk.	Berechnung	Freiheitsgrade
Faktor 1	SS_{F1}	$SS_{F1} = \sum_{i=1}^{I} n_{i\cdot} (\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot})^2$	$I - 1$
Faktor 2	SS_{F2}	$SS_{F2} = \sum_{j=1}^{J} n_{\cdot j} (\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdot\cdot\cdot})^2$	$J - 1$
Gruppen	SS_E	$SS_E = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k} (y_{ijk} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdot\cdot\cdot})^2$	$n - I - J + 1$
total	SS_T	$SS_T = \sum_{i=1}^{I} \sum_{j=1}^{J} (y_{ijk} - \bar{y}_{\cdot\cdot\cdot})^2$	$n - 1$

Übersicht

Verallgemeinerte lineare Modelle (GLM)

$\vec{X} := (X_1, \ldots, X_p)$, $\vec{x} := (x_1, \ldots, x_p)$
Die X_j können sowohl stetig also auch kategoriell sein.

Sei $Y \in \mathbb{R}$ und $\eta := g(\mathbb{E}[Y|\vec{X} = \vec{x}])$.
GLM's modellieren $\eta = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$.
g heißt dabei Link-Funktion.

Typische Beispiele:

Modell	Skalenniveau von Y	Link-Funktion g
ANCOVA	stetig	id. ($\vec{X}$ stetig)
ANOVA	stetig	id. ($\vec{X}$ kategoriell)
log.-linear	$Y \in (0, \infty)$	log
Poisson	$Y \in \mathbb{N}$	log
logistic	dichotom	$\text{logit}(x) = \ln(x/(1-x))$
Cox	Zeitspanne	$\eta = \log h(t)$

Übersicht

Sei T eine Zufallsvariable mit Werten in $\mathbb{R}_+$, die eine Lebenszeit beschreibt und

$$S(t) = \mathbb{P}(T > t) = 1 - F_T(t)$$

die zugehörige Survival–Funktion.

Die Hazard–Funktion wird definiert als

$$h(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

und die kumulative Hazard–Funktion durch

$$H(t) = \int_0^t h(s)ds = -\log S(t).$$

Ist $f = F'$ die stetige Dichte der Verteilung von T, so gilt $h(t) = f(t)/S(t)$.

Zensierte Daten

Eine der Hauptschwierigkeiten bei der Durchführung von Lebenszeitanalysen sind zensierte Daten.

Gründe:

- Zum Ende der Studie ist noch nicht bei allen Teilnehmenden das Zielereignis eingetreten
- Während der Durchführung der Studie scheiden Teilnehmer aus Gründen aus, die in keinem Bezug zum Zielereignis stehen

Dann ist die „naive“ Schätzung durch

$$\hat{S}(t) = \frac{\text{Überlebende zur Zeit } t}{n} \text{ verzerrt.}$$

Kaplan–Meier–Schätzer

Definition

Betrachten wir die geordneten Ereigniszeitpunkte $t_{(1)} < t_{(2)} < \ldots < t_{(n)}$.

Dann ist der Kaplan–Meier–Schätzer gegeben durch

$$\hat{S}_{KM}(t) = \prod_{j|t_{(j)} \leq t} \left(1 - \frac{d_j}{r_j}\right).$$

Dabei bezeichnet r_j die Anzahl der Einheiten, die unmittelbar vor $t_{(j)}$ noch unter Risiko stehen und d_j die Anzahl der Einheiten, die zum Zeitpunkt $t_{(j)}$ ausfallen.

Kaplan–Meier–Schätzer

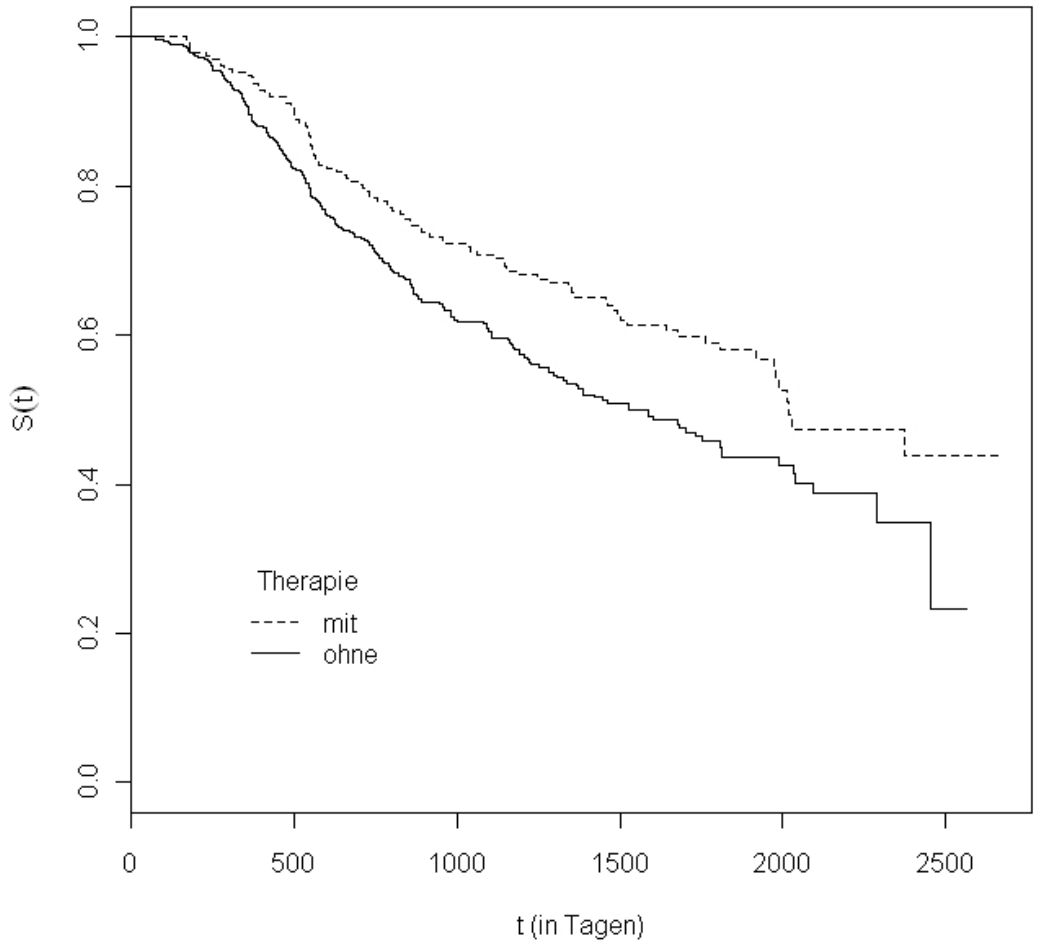
Der Kaplan–Meier–Schätzer ist der Maximum–Likelihood–Schätzer der Survival–Funktion.

Es gilt: $$\mathbb{V}\mathrm{ar}(\hat{S}_{KM}(t)) = \left(\hat{S}_{KM}(t)\right)^2 \sum_{j|t_{(j)}\leq t} \frac{d_j}{r_j(r_j - d_j)}.$$

Die (kumulative) Hazardfunktionen werden geschätzt durch

$$\hat{H}(t) = \sum_j \frac{d_j}{n_j} \quad \text{und} \quad \hat{h}(t) = \frac{d_j}{n_j(t_{(j+1)} - t_{(j)})}\ .$$

Beispiel: Überlebenschancen mit/ohne Therapie



Cox–Regression

Definition

Sei T eine Lebensdauer und $\vec{X} = (X_1, ..., X_p)$ ein p-dimensionaler Vektor von erklärenden Variablen.

Das Cox proportional hazard–Modell modelliert die Hazard–Funktion in Abhängigkeit von $\vec{X}$ durch

$$\log h(t) = \sum_{i=1}^{p} \beta_i X_i + \log h_0(t).$$

$h_0(t)$ heißt die Basis–Hazard–Funktion (baseline hazard) und ist Hazard–Rate zum Kovariablenvektor $\vec{X} = 0$.

$(\beta_j)_{j=1}^{p}$ ist ein Vektor von Regressionskoeffizienten.

Cox–Regression

Das Cox–Modell ist ein semiparametrisches Regressionsmodell.

Das Verhältnis von Risiken $h(t|\vec{x}^{(1)})/h(t|\vec{x}^{(2)})$ zweier Individuen ist konstant.

Es wird hierbei ein Regressionsansatz für die Hazard–Funktion gewählt, da diese nicht von den r_j abhängt.