

Biometrie 2-Tutorium

**Skript
zu den Tutorien Biometrie 2
des Studienganges Public Health
der LMU München**

**Dozenten der Hauptvorlesung:
P. Dirschedl, S. Aydemir**

**WS 2004/05
WS 2005/06
WS 2006/07**

**Autoren:
A. Henneberger
H. Blankenfeld**

Stand: 22. Nov. 2006

Vorwort:

Dieses Skript entstand aus den Vorlesungsmitschriften der Hauptvorlesung Biometrie 2 sowie den Tutoriumsunterlagen von Alexandra Henneberger. Es kann naturgemäß keinen Anspruch auf Vollständigkeit und Fehlerfreiheit erheben. Zur Vertiefung des Stoffes verweisen wir daher auf die offiziellen Vorlesungsskripten der Dozenten sowie die einschlägigen Lehrbücher.

Tabellen können aufgrund des Copyrights der Verlage leider nicht mitgeliefert werden.

Wir hoffen jedoch, dass das Skript seinen Teil zur erfolgreichen Vorbereitung auf die Biometrie-2-Klausur beitragen kann.

München, im November 2006

A. Henneberger / H. Blankenfeld

Inhaltsverzeichnis

1	Einfache lineare Regression	4
1.1	allgemein	4
1.2	Vorraussetzungen	4
1.3	Varianzaufteilung.....	5
1.4	Korrelationskoeffizient r und Bestimmtheitsmaß B	5
1.5	Test auf Signifikanz des Korrelationskoeffizienten r	6
1.6	Minimierung der Residuenquadrate (kein Klausurstoff)	7
1.7	95%-Konfidenzintervall für die Regressionsgerade	7
1.8	95%-Voraussagebereich (prediction interval)	8
1.9	Wald-Test: Signifikanz von β_i	9
1.10	Konfidenzintervall für den Regressionskoeffizienten β_i	9
2	Multiple lineare Regression	10
2.1	Überblick: Was mache ich wann (und warum)?	10
2.2	Gleichung	11
2.3	Residuen und multiples Bestimmtheitsmaß	11
2.4	Anpassungstest (Goodness-of-fit-Test GOF / Overall-F-Test).....	11
2.5	Wald-Test: Signifikanz der β_i	12
2.6	Konfidenzintervall für β_i	12
2.7	Modellierung des linearen Prädiktors	12
3	Logistische Regression	15
3.1	Voraussetzungen	15
3.2	„Herleitung“	15
3.3	Wichtige Formeln: Odds; Odds Ratio; Wahrscheinlichkeit	18
3.4	Wald-Test zur Prüfung der einzelnen β_i	18
3.5	Likelihood-Ratio-Test (LR-Test)	19
3.6	Konfidenzintervall von OR_i	19
3.7	Dummy-Kodierung	20
3.8	Modellierung	21
3.9	Wechselwirkungen	24
4	Überlebenszeitanalyse	26
4.1	Vorbemerkungen.....	26
4.2	Cutler-Ederer-Methode.....	28
4.3	Kaplan-Meier-Verfahren	29
4.4	Mediane Überlebenszeit.....	30
4.5	Vergleich von Überlebenszeiten.....	31
4.6	Das Cox-Modell (Cox-Regression).....	33

1 Einfache lineare Regression

1.1 allgemein

Die drei Hauptgründe für eine Analyse des Zusammenhanges zwischen zwei kontinuierlichen Variablen sind:

1. sind beide Variablen voneinander abhängig? (**Korrelation**)
2. eine Vorhersage des Wertes der einen Variable aus der anderen zu ermöglichen (**Regression**) und ein geeignetes Modell zur Beschreibung der Realität zu suchen
3. den Grad der Übereinstimmung der Werte der beiden Variablen zu bestimmen (**Konkordanz**)

Korrelation und Regression:

Prädiktor-Variable x (unabhängig) \Rightarrow **Outcome-Variable y** (abhängig)

Eine lineare Beziehung zwischen Prädiktor und Outcome muss biologisch plausibel sein (Beispiel: nach der Pubertät ist ein linearer Zusammenhang zwischen Alter \rightarrow Größe nicht mehr plausibel).

1.2 Voraussetzungen

- Für jeden Wert von x sollte das y normalverteilt sein, zumindest aber symmetrisch verteilt
- Die Varianz von y sollte für jedes x gleich groß sein (**homoskedastisch**)
- **iid: identical** (y_i stammen aus einer Verteilung) **independent** (unabhängige Stichproben, nur je ein Wert der Variablen von einem Individuum) **distribution**

$$\bullet \quad y = \alpha + \beta x + \varepsilon$$

\uparrow \uparrow \uparrow
y-Achsenabschnitt Steigung Fehler

α und β sind die durch die Regression geschätzten Koeffizienten

- Der Fehler ε ist normalverteilt mit dem Erwartungswert 0 und der Varianz σ^2 :
 $\varepsilon \sim N(0, \sigma^2)$
- Die Fallzahl sollte möglichst > 100 sein (100 Paare x_i, y_i)
- **Überprüfung der Voraussetzungen:**
 - Scatterplot der Originalwerte
 - Residuenplot (Normalplot, Shapiro-Wilk)

1.3 Varianzaufteilung

Gesamte Varianz = Erklärte Varianz + Unerklärte Varianz

$$s^2 \left[= \frac{1}{n-1} \sum_i (y_i - \bar{y})^2 \right] = s_R^2 \left[= \frac{1}{n-1} \sum_i (\hat{y}_i - \bar{y})^2 \right] + s_F^2 \left[= \frac{1}{n-1} \sum_i (y_i - \hat{y}_i)^2 \right]$$

Abweichungsquadrate vom Mittelwert (Erwartungswert) = größste Schätzung von y_i

Schätzwerte von y_i (= Gerade)

Residuen → minimieren
"Least-square-Methode"

Die Residuen sollten für jedes x_i normalverteilt sein (um 0)

→ ist y_i normalverteilt ⇒ dann ist auch $(y_i - \hat{y}_i)$ normalverteilt

1.4 Korrelationskoeffizient r und Bestimmtheitsmaß B

Der **Korrelationskoeffizient r** ist ein Maß der Streuung der (Mess-)werte um einen zugrunde liegenden linearen Trend: je größer die Streuung desto geringer die Korrelation. r kann folgenderweise berechnet werden:

$$s^2 = s_R^2 + s_F^2 \quad | : s^2 \quad (\text{vgl. 1.3})$$

$$1 = \frac{s_R^2}{s^2} + \frac{s_F^2}{s^2} \quad r^2 = \frac{s_R^2}{s^2} = \text{Bestimmtheitsmaß } B$$

↑
 r^2

= relativer Anteil (Prozentsatz) der durch die Regression erklärten Varianz an der gesamten Varianz

$$r = \sqrt{\frac{s_R^2}{s^2}} \quad \text{Korrelationskoeffizient}$$

$$r \in [-1;1]$$

$|r| = 1$ ⇒ „perfekter“ Zusammenhang, alle Punkte auf einer Geraden

$r = 0$ ⇒ kein Zusammenhang

Beispiel:

$r = 0,7$ ⇒ $r^2 = B = 0,49$ ⇒ 49 % der Gesamtvarianz werden durch die lineare Regression erklärt.

Beachte:

- Die Gerade ist nur dort gültig, wo Messwerte vorliegen! (keine „Extrapolation“)
- Schichtungen können etwas vortäuschen:

a) falscher Zusammenhang durch Cluster = „Klumpung“ (Schichten, Strata)

b) fälschlich kein Zusammenhang, obwohl einer in den einzelnen Schichten vorhanden wäre.

Gegenmaßnahme: r in den Strata berechnen; multiple Regression (S. 10)

1.5 Test auf Signifikanz des Korrelationskoeffizienten r

Hypothese:

$H_0: r = 0$ (kein Zusammenhang zwischen x und y)

$H_1: r \neq 0$ (Zusammenhang) \rightarrow 2-seitiger Test!

(einseitig: $H_1: r > 0$ oder $H_1: r < 0$)

Testgröße: $\hat{t} = r \cdot \sqrt{\frac{n-2}{1-r^2}}$	$t_{n-k-1;\alpha}$ – verteilt	(für $n > 60$)
--	-------------------------------	-----------------

$t_{n-2;\alpha}$ bei der einfachen lineare Regression ($k = 1$)

k = Anzahl der Prädiktoren

n = Fallzahl

Entscheidung: Wenn $|\hat{t}| > t_{Tabelle} \Rightarrow r$ signifikant von 0 verschieden

Für große n kann die z -Verteilung genommen werden

Alternative:

Direkt in der **r-Tabelle** nachsehen:

$ r > r_{Tabelle}(n, \alpha) \Rightarrow r$ signifikant von 0 verschieden
--

n = Fallzahl

α = Irrtumswahrscheinlichkeit

Test von r gegen einen vorgegebenen Wert r_0 :

Hypothese:

$H_0: r = r_0$

$H_1: r \neq r_0$

(einseitig: $H_1: r > r_0$ oder $H_1: r < r_0$)

Testgröße:

$\hat{t} = (r - r_0) \cdot \sqrt{\frac{n-2}{(1-r^2) \cdot (1-r_0^2)}}$	$t_{n-k-1;\alpha}$ – verteilt	(für $n > 60$)
--	-------------------------------	-----------------

$t_{n-2;\alpha}$ bei der einfachen lineare Regression ($k = 1$)

k = Anzahl der Prädiktoren

n = Fallzahl

Entscheidung: Wenn $|\hat{t}| > t_{Tabelle} \Rightarrow r$ signifikant von r_0 verschieden

Für große n kann die z -Verteilung genommen werden

1.6 Minimierung der Residuenquadrate

= Minimierung der unerklärten Varianz \Rightarrow „beste“ Gerade durch die Punktwolke

$$S_{xx} = \sum_i x_i^2 - \frac{1}{n} \cdot (\sum_i x_i)^2 = \sum_i (x_i - \bar{x})^2 = (n-1) \cdot Var(x)$$

$$S_{yy} = \sum_i y_i^2 - \frac{1}{n} \cdot (\sum_i y_i)^2 = \sum_i (y_i - \bar{y})^2 = (n-1) \cdot Var(y)$$

$$S_{xy} = \sum_i x_i y_i - \frac{1}{n} \cdot \sum_i x_i \sum_i y_i = \sum_i (x_i - \bar{x})(y_i - \bar{y}) = (n-1) \cdot Cov(x, y) \quad (Cov(x, y) = Kovarianz)$$

Gerade: $\hat{y} = \alpha + \beta \cdot x$

$$\alpha: \text{ intercept (y-Achsenabschnitt) } = \bar{y} - \beta \cdot \bar{x}$$

$$\beta: \text{ slope (Steigung) } = \frac{S_{xy}}{S_{xx}}$$

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} = \sqrt{\frac{S_R^2}{s^2}} ; B = r^2$$

Anmerkung: $se(\beta) = \frac{s_{res}}{\sqrt{S_{xx}}}$ mit $s_{res} = \sqrt{\frac{S_{yy} - \beta \cdot S_{xy}}{n-2}}$ (Residual-/Restvarianz)

1.7 95%-Konfidenzintervall für die Regressionsgerade

Für bestimmten Wert der x_i : $x = x_0$ gilt

$$\hat{y}_i \pm t_{n-2, 1-\alpha/2} \cdot se(\hat{y}_i) \quad (df = n-2)$$

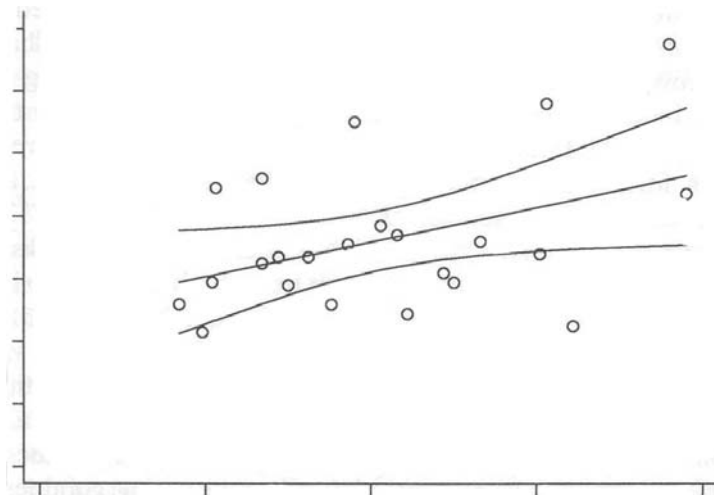
$$\text{mit } s_{res} = \sqrt{\frac{S_{yy} - \beta \cdot S_{xy}}{n-2}} \quad (\text{Residual-/Restvarianz}) \quad \text{und}$$

$$se(\hat{y}_i) = s_{res} \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

alle Geraden in diesem Konfidenzband sind theoretisch als Regressionsgeraden möglich (die wahre Gerade liegt zu 95% in diesem Konfidenzband):

Die Gerade repräsentiert die zu erwartenden Mittelwerte \bar{y}_i für jeden gegebenen Wert x_i

Je größer n , desto enger wird das Konfidenzintervall



95% Konfidenzintervall für die Regressionsgerade

1.8 95%-Voraussagebereich (prediction interval)

Für bestimmten Wert der x_i : $x = x_0$ gilt

$$\hat{y}_i \pm t_{n-2, 1-\alpha/2} \cdot s_{pred} \quad (df = n-2)$$

⇒ Prognoseband für y-Werte (breiter als das Konfidenzband und schließt oft alle y-Werte ein). Einzelne Punkte sind ablesbar (Für $x = \dots$ liegt der prognostizierte Wert zwischen ... und ...)

$$\text{mit } s_{res} = \sqrt{\frac{S_{yy} - \beta \cdot S_{xy}}{n-2}} \quad (\text{Residual-/Restvarianz}) \quad \text{und}$$

$$s_{pred} = s_{res} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

Das Prognoseband wird nur wenig enger mit steigendem n (siehe Altmann, S. 307)

2 Multiple lineare Regression

2.1 Überblick: Was mache ich wann (und warum)?

- | | | |
|------------------------------------|--|---|
| a) Outcome y metrisch | + ein <u>diskreter</u> Prädiktor x | Student's t-Test (x hat 2 Klassen)
bzw. ANOVA (x hat k Klassen) |
| b) Outcome y metrisch | + ein <u>metrischer</u> Prädiktor x | (einfache) lineare Regression |
| c) Outcome y metrisch | + <u>k metrische</u> Prädiktoren x_i
+ <u>k diskrete</u> Prädiktoren x_i
+ <u>k gemischte</u> Prädiktoren x_i | multiple lineare Regression
k-fache Varianzanalyse
gemischte multiple Regression / Kovarianzanalyse |
| d) Outcome y binär | + ein oder mehrere <u>diskrete</u>
<u>oder metrische</u> Prädiktoren x_i | Logistische Regression |
| e) Outcome y
= Überlebenszeit | + ein/mehrere <u>diskrete</u>
<u>oder metrische</u> Prädiktoren x_i | Cox-Modell |

2.1.1 Warum multipel und nicht einfach?

Man möchte...

- den Effekt eines Prädiktors unverzerrt** („unbiased“) schätzen = „Nettoeffekt“ (d.h. bereinigt von den Effekten der anderen Prädiktoren)
- den gemeinsamen Effekt vieler Prädiktoren gleichzeitig** schätzen \Rightarrow bessere Übertragbarkeit auf andere Populationen
- die Prognose des Outcome** für neue Werte der Prädiktoren verbessern

2.1.2 Probleme

Im Rahmen der multiplen linearen Regression sind folgende Punkte zu beachten (sind die Modellvoraussetzungen erfüllt? wie modelliere ich? ist das Modell gut?):

- ist der Zusammenhang zwischen x_i und y wirklich linear?
- Sind die Fehler ε normalverteilt?
- Art der Variablenselektion: forward oder backward?
- Schätzung der Effekte $\hat{\beta}_i$
- Signifikanz der $\hat{\beta}_i$ \rightarrow Wald-Test, Konfidenzintervall
- Prädiktoren unabhängig oder mit Wechselwirkungen untereinander?
- Erklärungswert des Modells \rightarrow multiples Bestimmtheitsmaß R^2
- Wie gut ist das Modell \rightarrow Residuenanalyse, Goodness-of-fit-Test, Kreuzvalidierungen
- Interpretation des Modells \rightarrow ...

2.2 Gleichung

„Realität“:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k + \varepsilon$$

↑
intercept
↑
Fehler ε : $N(0, \sigma^2)$ -verteilt

„Modellschätzung“:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \dots + \hat{\beta}_k x_k$$

= linearer Prädiktor l

2.3 Residuen und multiples Bestimmtheitsmaß (vgl. Seite 4)

Summe der Residuenquadrate: $Q_e = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ (n Fälle in der Regression)

(vgl. unerklärte Varianz) ↳ minimieren: „least-square-method“

Gesamtstreuung: $Q_o = \sum_{i=1}^n (y_i - \bar{y})^2$

(vgl. Gesamtvarianz)

Summe der Regressionsquadrate: $Q_o - Q_e = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$

(vgl. erklärte Varianz)

$$\Rightarrow \text{Güte der Regression: } R^2 \quad R^2 = \frac{Q_o - Q_e}{Q_o} = B \text{ (multiples Bestimmtheitsmaß)}$$

$R^2 \in [0;1]$ je näher R^2 bei 1, desto besser der Erklärungswert der Regression

$R^2_{x_1, x_2} - R^2_{x_1} = B_{x_1, x_2} - B_{x_1}$ macht eine Aussage darüber, wie viel Prozent durch die Aufnahme von Faktor x_2 zusätzlich erklärt werden.

„partieller Korrelationskoeffizient“: $\sqrt{\frac{Q_e(x_1) - Q_e(x_1, x_2)}{Q_e(x_1)}}$

Bedeutung: relativer Anteil der Varianz des ersten Modells (nur x_1), der zusätzlich durch Faktor x_2 erklärt wird. Der Erklärungswert des neuen Modells (x_1 und x_2) steigt insgesamt, jedoch sinkt der Erklärungswert von x_1 im neuen Modell im Vergleich zum alten Modell (nur x_1).

2.4 Anpassungstest (Goodness-of-fit-Test GOF / Overall-F-Test)

Hypothesen:

$$H_0: \beta_1 = \beta_2 = \beta_3 = \dots = \beta_k = 0$$

$$H_1: \beta_i \neq 0$$

(mindestens ein $\beta_i \neq 0$!)

$$\text{Testgröße: } F = \frac{R^2}{1-R^2} \cdot \frac{n-k-1}{k}$$

$$\text{Testentscheidung: } F_{\text{Test}} > F_{\text{Tabelle}}$$

\Rightarrow signifikant (mind. ein β_i ist $\neq 0$),

$$\text{Tabelle: } F(k, n-k-1)$$

k = Prädiktoren, n = Fallzahl

2.5 Wald-Test: Signifikanz der $\hat{\beta}_i$

Hypothesen:

$$H_0: \beta_i = 0 \quad H_1: \beta_i \neq 0 \quad (\text{für jedes } \hat{\beta}_i \text{ einzeln})$$

$$\text{Testgröße: } \hat{t} = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)}$$

$t_{n-k-1; \alpha}$ -verteilt, k = Anzahl der Prädiktoren

$$se(\hat{\beta}_i) = \sqrt{Var(\hat{\beta}_i)}$$

$$\text{Entscheidung: } |\hat{t}| > t_{\text{Tabelle}}$$

$\Rightarrow \beta_i$ signifikant von 0 verschieden

$$\text{Tabelle: } t_{n-k-1}$$

n = Fallzahl, k = Prädiktoren

n groß $\Rightarrow t \rightarrow z$ (Normalverteilung)

2.6 Konfidenzintervall für $\hat{\beta}_i$

$$KI_{1-\alpha}(\hat{\beta}_i) = \hat{\beta}_i \pm se(\hat{\beta}_i) \cdot t_{n-k-1}$$

$$n > 100: KI_{95\%}: \hat{\beta}_i \pm 1,96 \cdot se(\hat{\beta}_i)$$

KI wird größer für kleine n , da die Schätzung dann ungenauer ist

Wie immer: bei $n > 100$ auch z -Verteilung statt t -Verteilung

2.7 Modellierung des linearen Prädiktors

2.7.1 Forward selection

a) Alle möglichen Prädiktoren einzeln testen:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2 \quad \text{etc.}$$

\rightarrow Wald-Test für β_1, β_2 usw.

\Rightarrow wähle das x_i mit dem kleinsten p-Wert (möglichst $< 0,05$) für β_i im Wald-Test

⇒ „ x_1 “ (erster Prädiktor im Modell)

b) Inklusion eines weiteren Parameters:

Alle Paare der restlichen x_i mit x_1 durchtesten:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_3 x_3 \quad \text{etc.}$$

→ Wald-Test für β_2, β_3 usw.

⇒ Auswahl des β_i mit dem kleinsten p-Wert
(nur wenn $p \leq 0,05$, sonst ist das Modell schon mit x_1 beendet)

⇒ „ x_2 “ (zweiter Prädiktor im Modell)

c) Gegebenenfalls weitere Inklusion von Parametern analog Schritt b)

d) Wenn keine p-Werte < 0.05 mehr auftreten ist das Modell fertig

2.7.2 Backward elimination

a) Beginn mit vollem Modell:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \dots + \hat{\beta}_k x_k$$

**b) Wald-Test für alle β_i ⇒ Elimination des x_i mit dem „schlechtesten“ β_i
(größter p-Wert [$> 0,05$] im Wald-Test)**

**c) ⇒ Kleineres Modell, wieder Wald-Test für die vorhandenen β_i
⇒ Elimination des β_i mit dem größten p-Wert**

**d) weitere Verkleinerung des Modells bis zum Stopp:
alle p-Werte der übrig gebliebenen β_i sind signifikant ($p \leq 0,05$)**

Problem: „Overfitting“ an der Stichprobe ⇒ schlechte Übertragbarkeit auf andere Populationen (Generalisierbarkeit / externe Validität)
⇒ lieber sparsam modellieren

Overfitting: Hiervon spricht man, wenn die Vorhersageleistung eines statistischen Modells, welches an einem abhängigen Datenkollektiv (Testkollektiv) entwickelt wurde, aufgrund zu vieler verwendeter Prädiktoren an einer unabhängigen Teilstichprobe (Examinationskollektiv, erneutes Testkollektiv) dramatisch einbricht.

CAVE: Forward selection und Backward elimination führen meistens nicht zum gleichen Modell!

2.7.3 Plausibilität der β_i prüfen

$\beta_i < 0$: präventiver Prädiktor x_i

$\beta_i > 0$: Risikofaktor x_i

Falls Unplausibilitäten auftreten:

- Modell überdenken
- Kodierung von Outcome / Prädiktor(en) überprüfen

Man braucht daher eine Funktion, die die „0/1-Realität“ besser modelliert.

Eine Möglichkeit (aber nicht die einzige) ist folgende:

$$p = y = \frac{e^x}{1+e^x} = \frac{\frac{e^x}{e^x}}{\frac{1+e^x}{e^x}} = \frac{1}{1+e^{-x}}$$

Lineare Regression: $y = \alpha + \beta x$

$$\text{Jetzt: } \ln\left(\frac{p}{1-p}\right) = \alpha + \beta x = \text{linearer Prädiktor } l$$

odds

log odds

= logit p

„Rechtfertigung“ warum $\text{logit } p = \alpha + \beta x$:

$$\left. \begin{array}{l} p=0 \Rightarrow \ln\left(\frac{p}{1-p}\right) = \ln\left(\frac{0}{1}\right) = -\infty \\ p=1 \Rightarrow \ln\left(\frac{p}{1-p}\right) = \ln\left(\frac{1}{0}\right) = +\infty \end{array} \right\} \text{passt zum Wertebereich von } y = \alpha + \beta x$$

$$\ln\left(\frac{p}{1-p}\right) = l = \alpha + \beta \cdot x$$

$$\frac{p}{1-p} = e^l = e^{\alpha + \beta \cdot x}$$

$$p = (1-p) \cdot e^{\alpha + \beta \cdot x}$$

$$p = e^{\alpha + \beta \cdot x} - p \cdot e^{\alpha + \beta \cdot x}$$

$$p + p \cdot e^{\alpha + \beta \cdot x} = e^{\alpha + \beta \cdot x}$$

$$p \cdot (1 + e^{\alpha + \beta \cdot x}) = e^{\alpha + \beta \cdot x}$$

$$p = \frac{e^{\alpha + \beta \cdot x}}{1 + e^{\alpha + \beta \cdot x}} = \text{„Eintrittswahrscheinlichkeit“}$$

- Minimum von $\alpha + \beta x$: $-\infty \Rightarrow p = 0$
- Maximum von $\alpha + \beta x$: $+\infty \Rightarrow p = 1$
- $x = 0 \Rightarrow \alpha + \beta x = \alpha \Rightarrow p = \frac{e^\alpha}{1 + e^\alpha} = \frac{1}{1 + e^{-\alpha}}$ „Baseline-Risiko“ (Grundrauschen)

Ziel: Schätzung für p hoch für Leute, die wirklich krank werden und niedrig für Leute, die gesund bleiben.

3.2.1 Beispiel:

$x = \text{exponiert ja/nein} = 1 \text{ oder } 0$

Proband 1: $x = 0$ (nicht exponiert)

Proband 2: $x = 1$ (exponiert)

Odds für Proband 1: $\frac{p_1}{1-p_1} = e^{\alpha+\beta \cdot 0} = e^\alpha$

Odds für Proband 2: $\frac{p_2}{1-p_2} = e^{\alpha+\beta \cdot 1} = e^{\alpha+\beta}$

Odds Ratio Proband 2 / Proband 1: $OR = \frac{\frac{p_2}{1-p_2}}{\frac{p_1}{1-p_1}} = \frac{e^{\alpha+\beta}}{e^\alpha} = \frac{e^\alpha \cdot e^\beta}{e^\alpha} = e^\beta$
 = OR (exp./nicht exp.)

3.2.2 Wie kommt man an α und β ? „Maximum likelihood-Verfahren“

Prinzip:

y -Abweichungen Realität – Modell sollen möglichst klein sein (vgl. kleinste Abstandsquadrate)

⇒ wenn „krank“ soll \hat{p} nahe bei 1 liegen, wenn „gesund“ soll \hat{p} nahe bei 0 liegen

⇒ Produkt: $L = \prod_{\text{Kranke}} \hat{p}_i \cdot \prod_{\text{Gesunde}} (1-\hat{p}_j)$

maximieren durch ausprobieren → PC

Likelihood
 sollte nahe 1 sein
 sollte auch nahe 1 sein (da \hat{p}_j nahe 0)

$\hat{p}_i \in [0;1] \Rightarrow L \in [0;1]$

Dieses so genannte „**Maximum likelihood-Verfahren**“ liefert dann $\alpha + \beta_i$

Warum Produkt? Beobachtungen sind unabhängig ⇒ $P(A \cap B) = P(A) \cdot P(B)$

3.2.3 Devianz

$D = -2 \cdot \ln L$
 $= -2 \cdot LL$ → minimieren (entspricht L maximieren)

Log-Likelihood-Funktion

Minuszeichen bei D macht die Werte positiv

$L \in [0;1] \Rightarrow LL \in]-\infty; 0] \Rightarrow D \in [0; +\infty[$

Wichtig für die Devianz: $L^2 \Rightarrow D = -\ln L^2 = -2 \cdot \ln L = -2 \cdot LL$

vgl. least-square-method

3.3 Wichtige Formeln: Odds; Odds Ratio; Wahrscheinlichkeit

$$\ln\left(\frac{p}{1-p}\right) = l = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_i x_i = \text{linearer Prädiktor } l$$

$$\text{Odds} = e^l = e^{\alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_i x_i}$$

$$OR = e^{l-l_0} = e^{\Delta l} \Rightarrow OR_i = e^{\beta_i} \text{ (Anstieg um 1 Einheit)}$$

$$OR_i = e^{c \cdot \beta_i} \text{ (Anstieg um } c \text{ Einheiten)}$$

$$p = \frac{1}{1+e^{-l}} = \frac{e^l}{1+e^l} = \frac{\text{Odds}}{1+\text{Odds}}$$

3.4 Wald-Test zur Prüfung der einzelnen β_i

Hypothesen:

$$H_0: \beta_i = 0$$

$$H_1: \beta_i \neq 0$$

$$\text{Testgröße: } \hat{t} = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)}$$

$t_{n-k-1; \alpha}$ -verteilt, k = Anzahl der Prädiktoren

$$se(\hat{\beta}_i) = \sqrt{\text{Var}(\hat{\beta}_i)}$$

$$\text{Entscheidung: } |\hat{t}| > t_{\text{Tabelle}}$$

$\Rightarrow \beta_i$ signifikant von 0 verschieden

$$\text{Tabelle: } t_{n-k-1}$$

n = Fallzahl, k = Anzahl Prädiktoren

n groß (> 100) $\Rightarrow t \rightarrow z$ (Normalverteilung)

Hinweis:

$$\text{Statt } \hat{t} = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)} \approx z \text{ geht auch: } \chi_{df=1}^2 = \left(\frac{\hat{\beta}_i}{se(\hat{\beta}_i)}\right)^2$$

Klausur: mit t_{n-1} rechnen (laut E-Mail Dirsched!!)

Standardisiertes β_i : $K_i = \beta_i \cdot sd(\text{Prädiktor})$

3.5 Likelihood-Ratio-Test (LR-Test)

Hypothesen:

$$H_0: \quad \vec{\beta}_+ = 0 \quad \text{„}\vec{\beta}_+ \text{“ steht für zusätzliche } \beta_i \text{ im Modell}$$

$$H_1: \quad \vec{\beta}_+ \neq 0$$

Testgröße:

$$LR = 2 \cdot (LL(\vec{\beta}, \vec{\beta}_+) - LL(\vec{\beta})) = 2 \cdot LL(\vec{\beta}, \vec{\beta}_+) - 2 \cdot LL(\vec{\beta})$$

$$LR = D_{\text{Untermmodell}} - D_{\text{Obermodell}}$$

χ_{df}^2 - verteilt; df = Unterschied der Variablenanzahl in den beiden Modellen

Entscheidung: $LR > \chi_{1-\alpha, df}^2 \Rightarrow$ Test signifikant, zusätzliches β_i signifikant

Tabelle: χ_{df}^2 df = Unterschied der Variablenanzahl in den beiden Modellen

Der LR-Test wird für die Modellierung (forward oder backward) benötigt.

3.6 Konfidenzintervall von OR_i

$$KI_{(1-\alpha)} : e^{\hat{\beta}_i \pm z_\alpha \cdot \underbrace{\sqrt{Var(\hat{\beta}_i)}}} \Rightarrow OR_i : [e^{\hat{\beta}_i - z_\alpha \cdot \sqrt{Var(\hat{\beta}_i)}}; e^{\hat{\beta}_i + z_\alpha \cdot \sqrt{Var(\hat{\beta}_i)}}]$$

oder

$$KI_{(1-\alpha)} : e^{\hat{\beta}_i \pm z_\alpha \cdot se(\hat{\beta}_i)} \Rightarrow OR_i : [e^{\hat{\beta}_i - z_\alpha \cdot se(\hat{\beta}_i)}; e^{\hat{\beta}_i + z_\alpha \cdot se(\hat{\beta}_i)}]$$

(Bei $n < 100$ lieber t -Verteilung nehmen, $t_{n-1, \alpha}$)

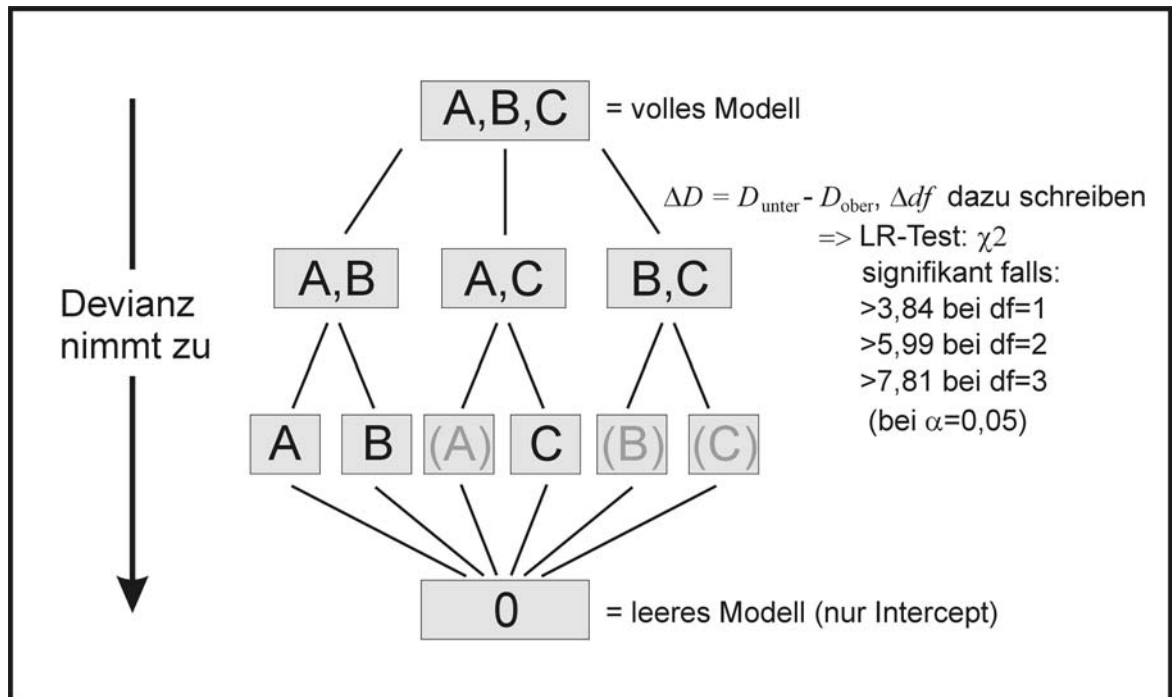
Testbasiert (Miettinen):

$$KI_{(1-\alpha)} : \left[OR^{1 - \frac{z_\alpha}{\sqrt{\chi^2}}}; OR^{1 + \frac{z_\alpha}{\sqrt{\chi^2}}} \right] \quad \chi^2 \text{ aus dem Wald-Test bzw. LR-Test}$$

3.8 Modellierung

3.8.1 Backward elimination

→ Hierarchisch arbeiten, Beispiel mit 3 Einflussfaktoren A,B,C:



- **Verschiedene Pfade mit gleichem Start und Ziel haben gleiches ΔD :**

z.B. $A,B,C \rightarrow A,B \rightarrow A \rightarrow 0$
 $A,B,C \rightarrow B,C \rightarrow B \rightarrow 0$

ΔD gleich

- **bei nominalen Variablen mit c Klassen: $df = c - 1$**
- **LR-Test signifikant** (Seite 19)
⇒ Obermodell ist überlegen (ausgesonderte Variable war signifikant), **Stopp**
- **LR-Test nicht signifikant**
⇒ Untermodell ist überlegen ⇒ **weitere Elimination**
- **Falls mehrere Pfade nicht signifikant sind:**
kleinster χ^2 -Wert entscheidet (falls df bei den Pfaden gleich)
bzw. größter p-Wert (egal welche df)
- **Ausgesonderte Faktoren = keine Confounder**
- **Vorteil „backward elimination“:** alle Variablen ohne Vor-Screening im Modell
- **Nachteil:**
zu viele Variablen ⇒ ungenaues Modell (valide aber nicht präzise) ⇒ Konfidenzintervalle werden riesig

- **Faustregeln für die Fallzahl gilt:**

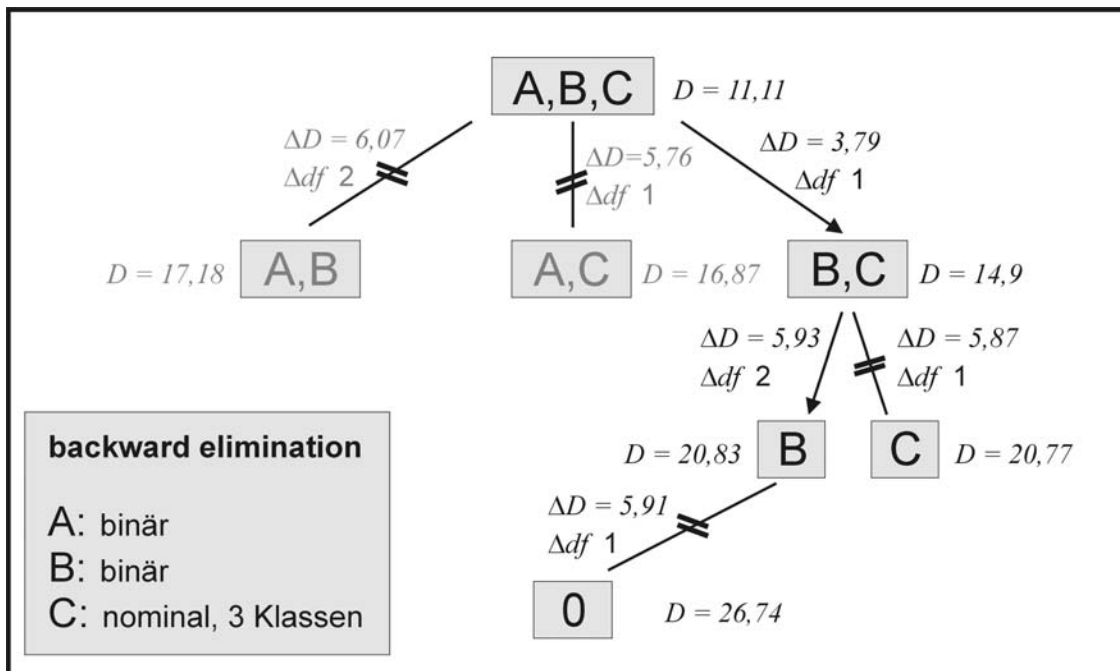
Ein gutes Verhältnis zwischen Fallzahlen und Anzahl der Variablen ist:

Zahl der Variablen * 10 = Fallzahl n

Variablen/Termanzahl $\approx \sqrt{n}$ n = Probandenzahl

Variablen/Termanzahl $\approx \frac{n}{10}$

3.8.1.1 Beispiel:



Beispiel: Backward elimination bei einem Modell mit 3 Prädiktoren (A,B binär, C nominal mit 3 Klassen)

1.) Start mit vollem Modell A,B,C: Devianz 11,11

2.) Elimination eines Prädiktors aus dem Modell

⇒ aus ΔD und df (C nominal mit 3 Klassen ⇒ 2 df) LR-Test berechnen:

A,B,C → A,B: $\Delta D = 6,07$, $df = 2$ > 5,99 ⇒ LR-Test signifikant!

A,B,C → A,C: $\Delta D = 5,76$, $df = 1$ > 3,84 ⇒ LR-Test signifikant!

A,B,C → B,C: $\Delta D = 3,79$, $df = 1$ < 3,84 ⇒ LR-Test nicht signifikant!

⇒ A aus dem Modell eliminieren ⇒ mit B,C weiter

3.) Elimination eines weiteren Prädiktors aus dem Modell

B,C → B: $\Delta D = 5,93$, $df = 2$ < 5,99 ⇒ LR-Test nicht signifikant!

B,C → C: $\Delta D = 5,87$, $df = 1$ > 3,84 ⇒ LR-Test signifikant!

⇒ C aus dem Modell eliminieren ⇒ mit B weiter

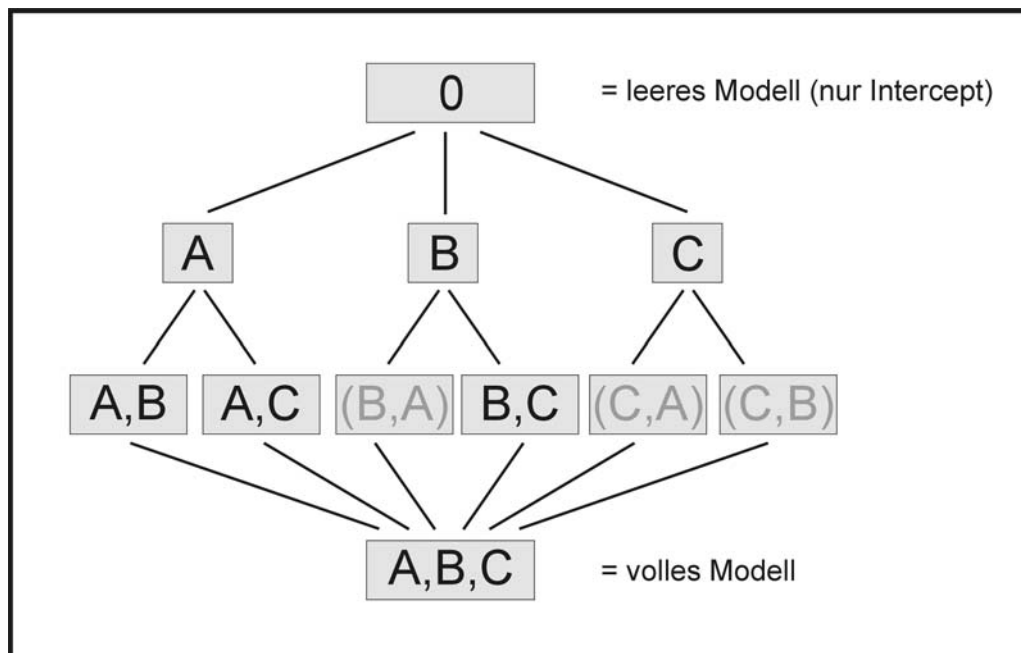
4.) B → 0: $\Delta D = 5,91$, $df = 1$ > 3,84 ⇒ LR-Test signifikant!

⇒ B kann nicht aus dem Modell eliminiert werden

⇒ fertiges Modell: **B**

3.8.2 Forward selection

Start mit dem Minimalmodell (leeres Modell):



Vorgehen:

- 1.) Start mit dem Null-Modell
- 2.) LR-Test für Obermodell mit einem Prädiktor
- 3.) Hinzufügen des Prädiktors mit dem „signifikantestem Ergebnis“ im LR-Test
(=größter χ^2 -Wert (bei gleichen df) bzw. kleinster p-Wert (unabhängig von df))
- 4.) LR-Test für Hinzufügen eines weiteren Prädiktors \Rightarrow evtl. weiteres Hinzufügen eines Prädiktors (analog 3.)
- 5.) Stopp der Forward selection, wenn kein LR-Test signifikant ist \Rightarrow Endmodell

BEACHTE:

In der **Epidemiologie** ist man nur an Modellen interessiert, die die untersuchte **Exposition** enthalten \Rightarrow Start mit der Exposition (nicht Nullmodell).

Bei der backward elimination wird die Exposition nie eliminiert!

3.9 Wechselwirkungen: OR und Konfidenzintervall

Bei Wechselwirkungen können Haupteffekte nicht mehr global angegeben werden: Es können nur noch bedingte OR berichtet werden, da jetzt neben A und B auch die **Wechselwirkung A*B** im Modell als Prädiktor steckt:

$$\text{Modell: } \hat{y} = \hat{\alpha} + \hat{\beta}_A \cdot A + \hat{\beta}_B \cdot B + \hat{\beta}_{AB} \cdot (A * B)$$

$$\text{z.B. OR } (A=1 \cap B=0 \text{ versus } A=0 \text{ und } B=1) = \frac{\text{Odds}(A=1 \cap B=0)}{\text{Odds}(A=0 \cap B=1)}$$

$$\text{z.B. OR}(A|B=1) = \frac{\text{Odds}(A=1)}{\text{Odds}(A=0)} \text{ im Stratum } B=1$$

Vorgehen:

OR (A=1 ∩ B=1 vs. A=0 ∩ B=1) berechnet sich wie folgt:

$$\begin{aligned} \Delta l &= l_{A=1;B=1} - l_{A=0;B=1} \\ \Delta l &= (\hat{\alpha} + \hat{\beta}_A \cdot 1 + \hat{\beta}_B \cdot 1 + \hat{\beta}_{AB} \cdot (1*1)) - (\hat{\alpha} + \hat{\beta}_A \cdot 0 + \hat{\beta}_B \cdot 1 + \hat{\beta}_{AB} \cdot (0*1)) \\ \Delta l &= (\hat{\alpha} + \hat{\beta}_A + \hat{\beta}_B + \hat{\beta}_{AB}) - (\hat{\alpha} + \hat{\beta}_B) \\ \Delta l &= \hat{\beta}_A + \hat{\beta}_{AB} \end{aligned}$$

$$\Rightarrow OR = e^{\beta_A + \beta_{A*B}}$$

Wechselwirkungen treten nur „in Aktion“ wenn beide Variablen „in Aktion“ sind!

Konfidenzintervall:

$$KI_{1-\alpha}: e^{(\beta_A + \beta_{A*B})} \pm z_\alpha \cdot se(\beta_A + \beta_{A*B})$$

$$se(\beta_A + \beta_{A*B}) = \sqrt{\text{Var}(\beta_A + \beta_{A*B})}$$

$$\text{Var}(\beta_A \pm \beta_{A*B}) = \underbrace{\text{Var}(\beta_A)}_{=(se(\beta_A))^2} + \underbrace{\text{Var}(\beta_{A*B})}_{=(se(\beta_{A*B}))^2} \pm 2 \text{cov}(\beta_A, \beta_{A*B})$$

Kovarianzmatrix:

	β_A	β_B	β_{A*B}
β_A	var (β_A)	cov (β_B, β_A)	cov (β_A, β_{A*B})
β_B	cov (β_B, β_A)	var (β_B)	cov (β_B, β_{A*B})
β_{A*B}	cov (β_{A*B}, β_A)	cov (β_{A*B}, β_B)	var (β_{A*B})

Anmerkungen zum logistischen Modell / Klausur:

Beispiel:

- A : ordinales Merkmal mit 3 Ausprägungen = 0;1;2
 B : metrisches Merkmal
 C : nominales Merkmal mit 3 Ausprägungen \Rightarrow **Dummy-Kodierung** mit D_1 und D_2 !
 E : dichotomes Merkmal = 0;1
 F : dichotomes Merkmal = 0;1
 $E*F$: Wechselwirkung zwischen E und F

$$\Rightarrow \hat{y} = \hat{\alpha} + \hat{\beta}_A \cdot A + \hat{\beta}_B \cdot B + \hat{\beta}_{D_1} \cdot D_1 + \hat{\beta}_{D_2} \cdot D_2 + \hat{\beta}_E \cdot E + \hat{\beta}_F \cdot F + \hat{\beta}_{EF} \cdot (E * F)$$

Variablen im Modell: 5 (A, B, C, E, F)

Terme im Prädiktor: 8 ($\alpha, \beta_A, \beta_B, \beta_{D_1}, \beta_{D_2}, \beta_E, \beta_F, \beta_{E*F}$)

„Risiko“-Klassen: $3 \cdot 3 \cdot 2 \cdot 2 = 24$ ($\#A \cdot \#C \cdot \#E \cdot \#F$) [metrische werden nicht berücksichtigt]
(Kategorien)

Beobachtungs-Null: Kombination von Merkmalen (=Risikoklasse), die in der beobachteten Kohorte zufällig nicht beobachtet wurden, aber prinzipiell vorkommen kann

Struktur-Null: Kombination von Merkmalen (=Risikoklasse), die in der Realität nicht vorkommen kann, aber der Vollständigkeit wegen mit aufgeführt wird

$e^{\beta_{Dummy\ i}}$ ist immer nur ein Vergleich zur Referenzkategorie (alle Dummies = 0)

Nur $e^{\beta_{D_1} - \beta_{D_2}}$ liefert Vergleiche von 2 Dummy-Kategorien!

Beispiel:		D_1	D_2	
Alter in drei nominalen Klassen	1	0	0	\rightarrow Referenz
	2	1	0	$\rightarrow \beta_{D_1}$
	3	0	1	$\rightarrow \beta_{D_2}$

$$OR\left(\frac{\text{Klasse2}}{\text{Klasse1}}\right) = e^{\beta_{D_1}} = \frac{\text{Odds}(\text{Klasse2})}{\text{Odds}(\text{Klasse1})} \quad OR\left(\frac{\text{Klasse3}}{\text{Klasse1}}\right) = e^{\beta_{D_2}} = \frac{\text{Odds}(\text{Klasse3})}{\text{Odds}(\text{Klasse1})}$$

$$OR\left(\frac{\text{Klasse3}}{\text{Klasse2}}\right) = e^{\beta_{D_2} - \beta_{D_1}} = \frac{e^{\beta_{D_2}}}{e^{\beta_{D_1}}} = \frac{\text{Odds}(\text{Klasse3}) / \text{Odds}(\text{Klasse1})}{\text{Odds}(\text{Klasse2}) / \text{Odds}(\text{Klasse1})}$$

4 Überlebenszeitanalyse

4.1 Vorbemerkungen

Warum ist bei der Analyse von Überlebenszeiten ein neues Verfahren notwendig?

Vergleich von Therapie A und B:

Fall a) Status: verstorben (1), lebt (0), Überlebenszeit unberücksichtigt
⇒ **Analyse mit 4-Felder-Tafel, Häufigkeiten, χ^2 -Test**

Fall b) Studie läuft bis alle Patienten aus A und B verstorben sind
⇒ Überlebenszeiten
⇒ **Analyse mit Mittelwertsvergleich der Überlebenszeit, t-Test**

Fall c) Studienende festgelegt ⇒ Lebende und Verstorbene in A und B
⇒ Status (0/1) und Überlebenszeiten von Verstorbenen und Lebenden
⇒ **neues Verfahren notwendig**

Für jeden Patienten liegen zwei Informationen vor:

Status (dichotom) und beobachtete **Überlebenszeit** (kontinuierlich)

Dabei unterscheiden sich die Überlebenszeiten bei Verstorbenen (tatsächliche ÜLZ) und Lebenden (zensierte ÜLZ, tatsächliche ÜLZ ist nicht bestimmbar)

Survivorfunktion (s): Wahrscheinlichkeit p, einen bestimmten Zeitpunkt t zu überleben (Überlebenswahrscheinlichkeit)

Abnahme mit der Zeit: $t = 0 \Rightarrow p = 1 \rightarrow t = \infty \Rightarrow p = 0$

Hazardfunktion (h): Wahrscheinlichkeit p, dass unmittelbar nach einem Zeitpunkt t ein Ereignis eintritt unter der Voraussetzung, dass es bisher noch nicht eingetreten ist

Zunahme mit der Zeit: $t = 0 \Rightarrow p = \text{klein} \rightarrow t = \infty \Rightarrow p = 1$

Zusammenhang zwischen Survivorfunktion s und Hazardfunktion h:

$$s_i(t) = e^{-\int_0^t h_i(u) du}$$

Events und Zensierungen:

Event: Ziel-Ereignis (z.B. Tod, Rezidiv etc.) tritt ein ⇒ Status = 1 und beobachtete ÜLZ

Zensierung: Kein Ziel-Ereignis beobachtet ⇒ Status = 0 und zensierte ÜLZ

im Beobachtungszeitraum

Zensierungsarten:

Zensierung 1: Patient lebt bei Studienende

Zensierung 2: Patient scheidet vorzeitig aus der Studie aus:

- Patient zieht weg
 - Patient erscheint nicht mehr
 - Patient verweigert weitere Teilnahme
- } Lost to Follow-up
- Event aus anderer Ursache
Problem: wirklich unabhängig von Therapie?
Nebenwirkungen \Rightarrow echtes Event!

Eventrate:

$$R_E: \frac{\# \text{ Events}}{\text{beobachtete Population}}$$

Zensierungsrate:

$$R_C: \frac{\# \text{ Zensierungen}}{\text{beobachtete Population}}$$

Es stehen 2 Verfahren zur Verfügung: Cutler-Ederer und Kaplan-Meier

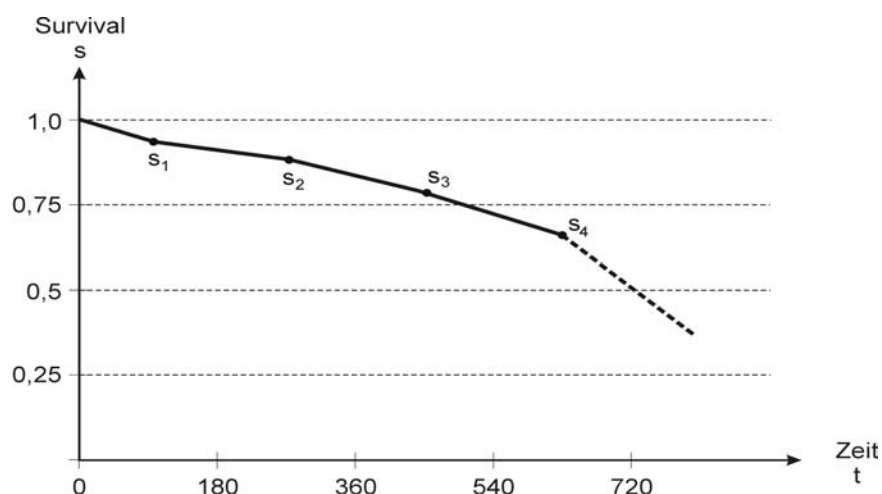
4.2 Cutler-Ederer-Methode

- **Synonyme:** Sterbetafelmethode, Aktuarsmethode
- **Konstante Intervalleinteilung**
- **Annahmen:**
Zensierungen und Events im Intervall gleichmäßig verteilt
⇒ Schätzungen in der Intervallmitte (daher $a_i/2$ in der Formel)
- Eventzeitpunkte oder Zensierungszeitpunkte müssen nicht ganz genau bekannt sein (nur das Intervall in dem das Event bzw. die Zensierung eingetreten ist)
⇒ **Plot: Polygonzug mit Punkten in den Intervallmitten**

Rechenschema - Beispiel: Life-Table

Zeit-intervall	# Patienten unter Risiko zu Intervallbeginn	# Ereignisse im Intervall	# Zensierungen im Intervall	Event-wahrscheinlichkeit q	Überlebens-wahrscheinlichkeit p	Survivorfunktion S
i	n_i	d_i	a_i	$q_i = d_i / (n_i - a_i/2)$	$p_i = 1 - q_i$	$S_i = s_{i-1} * p_i = \prod p_i$
$i = 0$ 0	$n_0 = 60$	0	0	$q_0 = 0$	$p_0 = 1$	$S_0 = 1$
$i = 1$ 1 - 180	$n_1 = 60$	2	1	$2 / (60 - 1/2)$ $q_1 = 0,0336$	$1 - 0,0336$ $p_1 = 0,9664$	$1 * 0,9664$ $S_1 = 0,9664$
$i = 2$ 181 - 360	$60 - 2 - 1$ $n_2 = 57$	4	5	$4 / (57 - 5/2)$ $q_2 = 0,0734$	$1 - 0,0734$ $p_2 = 0,9266$	$0,9664 * 0,9266$ $S_2 = 0,8955$
$i = 3$ 361 - 540	$57 - 4 - 5$ $n_3 = 48$	5	2	$5 / (48 - 2/2)$ $q_3 = 0,1064$	$1 - 0,1064$ $p_3 = 0,8936$	$0,8955 * 0,8936$ $S_3 = 0,8002$
$i = 4$ 541 - ...	$48 - 5 - 2$ $n_4 = 39$	$q_4 = ...$	$p_4 = ...$	$S_4 = ...$

Graph:



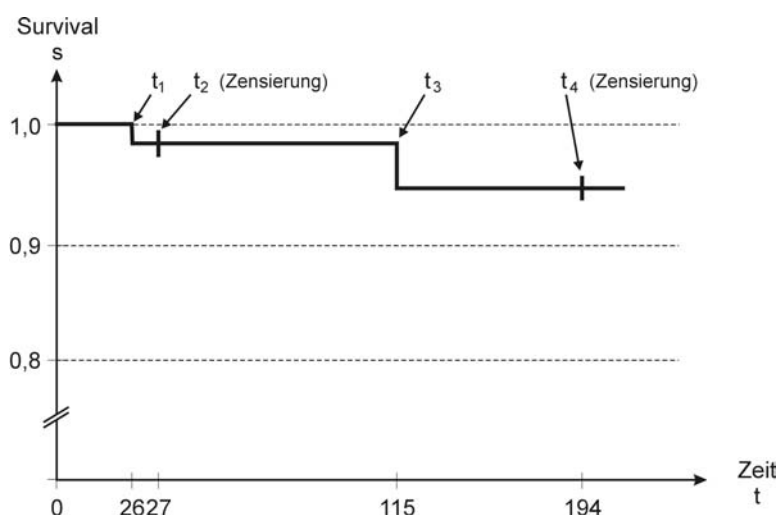
4.3 Kaplan-Meier-Verfahren

- ungleiche Intervalle
- in jedem Intervall ein bzw. mehrere gleichzeitige(!) Ereignisse (event oder Zensierung)
- Ereigniszeitpunkt = Intervallbeginn
- genaue Angabe von Event-/Zensierungszeitpunkten (nötig)
 - ⇒ genauer als Cutler-Ederer
 - ⇒ Plot: „Treppenfunktion“

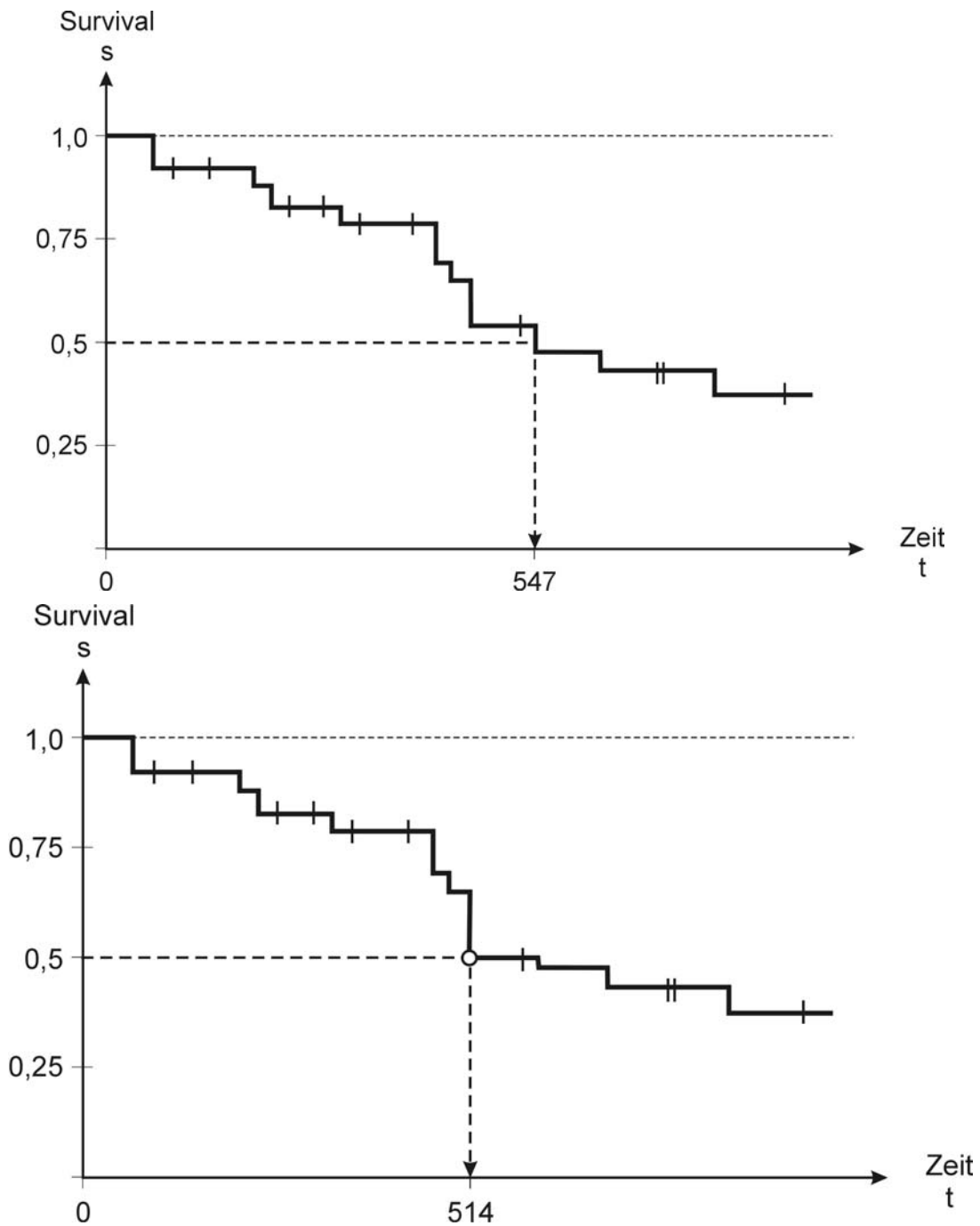
Rechenschema - Beispiel: Life-Table

Zeitintervall	# Patienten unter Risiko zu Intervallbeginn	# Ereignisse zum Zeitpunkt t_i	# Zensierungen zum Zeitpunkt t_i	Eventwahrscheinlichkeit q	Überlebenswahrscheinlichkeit p	Survivorfunktion S
i	n_i	d_i	a_i	$q_i = d_i/n_i$	$p_i = 1 - q_i$	$S_i = S_{i-1} * p_i = \prod p_i$
$i = 0$ $t_0 = 0$	$n_0 = 60$	0	0	0	$p_1 = 1$	$S_0 = 1$
$i = 1$ $t_1 = 26$	$n_1 = 60$	1	0	$1 / 60$ $q_1 = 0,0167$	$1 - 0,0167$ $p_1 = 0,9833$	$1 * 0,9833$ $S_1 = 0,9833$
$i = 2$ $t_2 = 27$	$60 - 1$ $n_2 = 59$	0	1	0 $q_2 = 0$	$1 - 0$ $p_2 = 1$	$0,9833 * 1$ $S_2 = 0,9833$
$i = 3$ $t_3 = 115$	$59 - 1$ $n_3 = 58$	2	0	$2 / 58$ $q_3 = 0,0345$	$1 - 0,0345$ $p_3 = 0,9655$	$0,9833 * 0,9655$ $S_3 = 0,9494$
$i = 4$ $t_4 = 194$	$58 - 2$ $n_4 = 56$	0	1	$0 / 56$ $q_4 = 0$	$1 - 0$ $p_4 = 1$	$0,9664 * 1$ $S_4 = 0,9494$

Graph:



4.3.1 Mediane Überlebenszeit



$s = 0,5 = 50\% \Rightarrow$ Zeitpunkt t ablesen (vom linken Punkt)

Falls mehr als 50 % das Studienende „überleben“:

\Rightarrow keine mediane Überlebenszeit anzugeben

Falls die Treppenkurve auf $p = 0$ endet:

\Rightarrow Event beim Patienten mit längster Beobachtungsdauer

Falls die Treppenkurve auf $p > 0$ endet:

\Rightarrow Zensierung beim Patienten mit längster Beobachtungsdauer

4.4 Vergleich von Überlebenszeiten

Log-Rank-Test:

- **Mantel-Haenszel-Verfahren:** 2 Survivorkurven
- **Peto-Pike-Verfahren:** 2 oder mehr (r) Survivorkurven
 - Vereinfachung vom M.-H.-Verfahren
 - konservativer als M.-H. (hält länger an H_0 fest), d.h. Testgröße stets kleiner als die vom M.-H.
 - keine Gewichtung der Ereignisse

4.4.1 Log-Rank-Test: Peto-Pike-Verfahren

Vorraussetzung: die Survivorkurven kreuzen sich nicht

Hypothesen:

H_0 : Gleichheit der Survivorkurven

H_1 : Unterschied der Survivorkurven (mindestens 2 von r)

Testgröße:

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

O_i = gesamte Zahl der beobachteten Ereignisse

E_i = gesamte Zahl der erwarteten Ereignisse
(erwartet = gleichwahrscheinlich bei allen Kurven)

$$E_i = \sum_{\text{alle Kurven}} O_i \cdot \frac{n_i}{\sum_{\text{alle Kurven}} n_i}$$

n_i = Anzahl der Probanden der jeweiligen Kurve

Entscheidung: $\chi^2 \geq \chi_{df=r-1}^2$

\Rightarrow mindestens 2 Survivorkurven unterscheiden sich

Tabelle: $\chi_{df=r-1}^2$

r = Anzahl der Survivorkurven

4.4.2 Welcher Test?

Log-Rank-Test:

- gewichtet alle Zeitpunkte gleich
- entdeckt eher spätere Unterschiede

Gehan-Test:

- spezieller gewichteter Mantel-Haenszel-Test
- gewichtet frühere Zeitpunkte stärker als spätere (sinnvoll, da n_i im Verlauf kleiner)
- weniger sensitiv für spätere Unterschiede
- weniger konservativ als Mantel-Haenszel-Test

Survivorkurven kreuzen sich nicht \Rightarrow Log-Rank- oder Gehan-Test

Survivorkurven kreuzen sich:

(Hinweis auf Interaktion der Variablen mit der Zeit: Die Risiken ändern sich offensichtlich mit der Zeit ⇒ Verletzung der proportional-hazards-Annahme)

- sie kreuzen sich „hinten“ ⇒ Gehan-Test
- sie kreuzen sich „vorne“ ⇒ Modell von Aalen (vgl. „proportional hazards“)
- sie kreuzen sich „mittig“ ⇒ kein Test, da offensichtlich kein Unterschied; (Studiendauer ändern?)

Vorgehen Log-Rank-Test:

1.)			# Patienten unter Risiko zu Intervallbeginn			# beobachtete Events (keine Zensurierungen)			# erwartete Events	
			$n_i = n_{i-1} - (d_{i-1} + a_{i-1})$			O_i			$E_i = \sum_{A+B} O_i \cdot \frac{n_i}{\sum_{A+B} n_i}$	
<i>I</i>	<i>t</i>	was?	Therapie A	Therapie B	$\sum_{A+B} n_i$	A	B	$\sum_{A+B} O_i$	A	B
1	26	Event A	30	30	60	1	0	1	$1 \cdot \frac{30}{60} = \frac{1}{2}$	$1 \cdot \frac{30}{60} = \frac{1}{2}$
2	27	Event B	30 - 1 = 29	30	59	0	1	1	$1 \cdot \frac{29}{59}$	$1 \cdot \frac{30}{59}$
3	115	Zens. B	29	30 - 1 = 29	58	0	0	0	0	0
4	194	Zens. A	29	28	57	0	0	0	0	0
5	211	2 Events A	28	28	56	2	0	2	$2 \cdot \frac{28}{56}$	$2 \cdot \frac{28}{56}$
6	215	Zens. B	26	28	54	0	0	0	0	0

usw.

2.)

$$\begin{matrix} \sum O_{i \text{ Therapie A}} & \sum E_{i \text{ Therapie A}} \\ \sum O_{i \text{ Therapie B}} & \sum E_{i \text{ Therapie B}} \end{matrix}$$

3.)
$$\chi^2_{df=1} = \frac{(\sum O_{iA} - \sum E_{iA})^2}{\sum E_{iA}} + \frac{(\sum O_{iB} - \sum E_{iB})^2}{\sum E_{iB}}; \quad r = 2 \Rightarrow df = 1$$

4.) Falls $\chi^2 \geq 3,84 \Rightarrow$ signifikanter Unterschied ($\alpha = 0.05$) zwischen den beiden Survivorkurven

Check:
$$\sum O_{i \text{ A+B}} = \sum E_{i \text{ A+B}}$$

4.5 Das Cox-Modell (Cox Regression)

Simultane Untersuchung mehrerer Einflussgrößen bei zensierten Überlebenszeiten

• Hazardfunktion (vgl. S. 26):

$$\begin{array}{ccc} \text{Bestimmte Kovariablen-Konstellation} & & \text{Linearkombination} \\ \underbrace{h_i(t | x_1, \dots, x_p)}_{\substack{\uparrow \\ \text{i-tes Individuum}}} = \underbrace{h_0(t)}_{\text{Baseline-Hazard:}} \cdot \underbrace{e^{\beta_1 x_1 + \dots + \beta_p x_p}}_{\text{individueller Teil: parametrisch}} \end{array}$$

- geschätzt aus geschätzten β_i
- nur Zeitkomponente
- nicht-parametrisch
- allen Individuen gemeinsam
- „enthält Konstante β_0 “ (intercept)
- ist für Aussagen (relative Risiken) nicht relevant (kürzt sich raus)

β_i = Kovariableneffekte (Richtung und Stärke der Kovariablen x_i)

- **Schätzung der β_i aus dem Maximum der partial Likelihood** (partial, da die Zensierungen zur Likelihood keinen Beitrag liefern, sie bleiben aber in der Risikomenge).

• Voraussetzungen:

- **proportional hazards:** Risikoverhältnisse sind über die Zeit gleich
 $\Rightarrow \beta_i = \text{const.}$ über die Zeit t , denn jedes x_i wird nur einmal pro Proband gemessen
- **iid (identical independent distribution)**
- Zensierungen sind unabhängig voneinander und auch vom Outcome unabhängig
- notwendige Daten: Beobachtungszeit, Status/Zensierung, unabhängige Variablen x_i

• Prüfung der „proportional hazards“-Annahme

Überprüft wird eine Interaktion von x mit der Zeit:

1.) grafisch:

Survivorfunktionen (bzgl. Der untersuchten Kovariablen) sollten sich nicht schneiden und keine „Bäuche“ haben.

Die LLS-Kurven [$\log(-\log(\text{survival}))$ -Kurve] verlaufen parallel.

2.) Statistische Tests: Wechselwirkung der Variablen mit der Zeit ($x^* \log(t)$):

Überprüfung ob die Wechselwirkung mit der Zeit signifikant ist
falls signifikant: Annahme verletzt! Beispiel: im Epi2-Skript

• Sich ändernde (zeitabhängige) Risiken:

- Cox-Modell kann nicht verwendet werden \Rightarrow **Aalen-Modell; time-dependent Cox:**
(x_i mehrfach gemessen $\Rightarrow \beta_i = \beta_i(t)$)

- Relatives Risiko:**

$$RR = e^{\hat{\beta}_i}$$

= **Hazardrisiko** (-ratio) beim Übergang von einer Merkmalsstufe in die nächste

$$\hat{\beta}_i = \ln RR$$

Übergänge immer von „niedrig“ nach „hoch“: $RR = \frac{\text{höhere Stufe}}{\text{niedrigere Stufe}}$
 $\beta = 0 \Rightarrow RR = 1$

Beispiel:

RR = 1,27 Risiko für Event steigt beim Übergang von einer Stufe zur nächsten
 ($\beta > 0$) Ausprägungsstufe um 27% (\Rightarrow kürzere Überlebenszeit)

RR = 0,87 Risiko für Event nimmt beim Übergang von einer Stufe zur nächsten
 ($\beta < 0$) Ausprägungsstufe um 13 % ab (\Rightarrow längere Überlebenszeit); protekti-
 ver Faktor

- 95%-Konfidenz-Intervall des Relativen Risiko:**

$$e^{c \cdot \beta \pm 1,96 \cdot |c| \cdot se(\beta)}$$

c = Anzahl der Risiko-Stufen

Begründung: $\text{var}(c \cdot \beta_i) = c^2 \cdot \text{var}(\beta_i) \Rightarrow se(c \cdot \beta_i) = c \cdot se(\beta_i)$

- Test auf Signifikanz von β_i**

1.) **Wald-Test** (mit z oder χ^2 bei großen n oder t_{n-1} für kleine n)

$$z = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)} \quad \text{oder} \quad \chi^2 = \left(\frac{\hat{\beta}_i}{se(\hat{\beta}_i)} \right)^2 \quad \text{oder} \quad t_{n-1} = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)}$$

2.) **KI für RR** enthält die 1, dann ist RR nicht signifikant.

- Prognosescore**

= linearer Prädiktor = $\sum_{i=1}^n \beta_i x_i$ (kein Intercept!)

Je größer der Prognose-Score, desto schlechter die Überlebenszeit (high risk).

Wertebereich: Minimum und Maximum angeben („aus l austüfteln“) \rightarrow dient zur Abschätzung des Risikoprofils einer Einzelperson.

Anhand des Prognose-Scores können Prognose-Gruppen gebildet werden, deren Überlebenszeiten anhand von K.-M.-Kurven geschätzt werden können.

• **Verhältnis Fallzahl – Variablen im Modell:**

Faustregeln:

Termanzahl $\approx \sqrt{n}$ $n =$ Probandenzahl

Termanzahl $\approx \sqrt{n_{\text{effektiv}}}$ $n_{\text{effektiv}} =$ # events in n
(Zensierungen \rightarrow Powerverlust)

Termanzahl $\approx \frac{n}{10}$

• **Zusammenhang bei Änderung der Kodierung der Prognosevariable x_i :**

1.) männlich = 1, weiblich = 0 $\Rightarrow \beta \Rightarrow RR = e^\beta$ bei w \rightarrow m

2.) männlich = 0, weiblich = 1 $\Rightarrow \beta^* \Rightarrow RR^* = e^{\beta^*}$ bei m \rightarrow w

Zusammenhang: $\beta = -\beta^*$ und $RR^* = e^{-\beta} = \frac{1}{RR}$

• **Probleme bei der Cox-Regression**

- Falsches Modell für gegebene Datenstruktur
- iid verletzt
- selection bias bei den Probanden (vgl. Epidemiologie-Skript 2)
- zeitabhängige Kovariablen \Rightarrow besser: time-dependent-Cox / Aalen-Modell
- Beobachtungsdauer unpassend für die Erkrankung
- unterschiedliche Beobachtungszeiten der Gruppen
- Klumpungen (Ausreißer)
- Fallzahl zu niedrig (zu wenige Events)
- $B = r^2$ sehr klein \Rightarrow wenig Aussagekraft
- zu viele Kovariablen

• **Weitere Modelle**

Poisson-Regression: Personenzeit + poisson-verteilte Ereignisse

Polytome logistische Regression: Ausprägungen von $y =$ polytom

Konditionale logistische Regression: für gematchte Daten

Modell von Aalen

Time-dependent Cox-Modell