

# Regressionsmodelle

- Einflussgrößen → Zielgröße (Alter, Geschlecht → Blutdruck)
- Zielgröße entscheidet über das Regressionsmodell
  - stetige Zielgröße → lineare Regression
  - binäre Zielgröße → logistische Regression
  - zensierte Zielgröße (Überlebenszeiten) → Cox-Regression
  - Anzahl: 1
- Einflussgrößen entscheiden über die Komplexität
  - 1 → einfache Regression
  - $> 1$  → multivariable / multiple (/ multivariate) Regression
  - quantitative oder qualitative Einflussgrößen möglich

# Einfache lineare Regression

- Beispiel NB 2004: hängt  $\log(\text{Ferritin})$  von  $\log(\text{LDH})$  ab
- Allgemein:
  - $Y$  = abhängige Variable = Zielgröße (z. B.  $\log(\text{Ferritin})$ )
  - $X$  = unabhängige Variable = Einflussgröße (z. B.  $\log(\text{LDH})$ )
- Modell:
  - deterministisch:  $Y = a + b \cdot X$
  - stochastisch:  $Y = a + b \cdot X + \varepsilon$   
 $\varepsilon$ : normalverteilt, Erwartungswert = 0, gleiche Varianz
- Begriffe: Regressionskoeffizienten  $a$  und  $b$ 
  - $a$  heißt Achsenabschnitt = Intercept
  - $b$  heißt Steigung = Slope

# Logistische Regression

- Wie hängt **eine binäre** Zielgröße von mehreren Einflussgrößen ab?
- Beispiel (B.W. Brown 1980 aus [1]):  
Population: Patienten mit Prostatakarzinom  
Zielgröße = **Befall regionaler Lymphknoten** (ja/nein)  
Einflussgrößen = Grading (niedrig/hoch),  
saure Phosphatase im Serum (King-Armstrong Einheiten),  
Röntgenbefund (negativ/positiv),  
Tumorgroße (klein/groß)

# Interpretation der Regressionsparameter

- $X$  eine Einflussgröße
- $p$  die Wahrscheinlichkeit für ein Ereignis
- $\text{logit}(p) = \log(p/(1-p)) = \log(\text{Odds}) = \beta_0 + \beta_1 X$
- $\text{Odds} = p/(1-p) = \exp(\beta_0 + \beta_1 X)$
- Das Odds verändert sich um Faktor  $\exp(\beta_1)$ ,

falls  $x$  um eine Einheit steigt – unabhängig von  $x$ :

$$\begin{aligned}\text{Odds}(x+1) &= \exp(\beta_0 + \beta_1(x+1)) = \exp(\beta_0 + \beta_1 x + \beta_1) \\ &= \exp(\beta_0 + \beta_1 x) \cdot \exp(\beta_1) = \text{Odds}(x) \cdot \exp(\beta_1)\end{aligned}$$

$$\Rightarrow \text{OR}(x+1, x) = \frac{\text{Odds}(x+1)}{\text{Odds}(x)} = \exp(\beta_1)$$

## Interpretation der Regressionsparameter

$$\text{OR}(x+1, x) = \frac{\text{Odds}(x+1)}{\text{Odds}(x)} = \frac{\exp(\beta_0 + \beta_1(x+1))}{\exp(\beta_0 + \beta_1 x)} = \exp(\beta_1).$$

- $\beta_1 = 0$ :  $\text{OR}(x+1, x) = \exp(\beta_1) = 1 \Rightarrow$   
 $\text{Odds}(x+1) = \text{Odds}(x) \Leftrightarrow$  **kein Einfluss von X**
- $\beta_1 > 0$ :  $\text{OR}(x+1, x) = \exp(\beta_1) > 1 \Rightarrow$   
 $\text{Odds}(x+1) > \text{Odds}(x) \Leftrightarrow$  **mit X steigende Chance**
- $\beta_1 < 0$ :  $\text{OR}(x+1, x) = \exp(\beta_1) < 1 \Rightarrow$   
 $\text{Odds}(x+1) < \text{Odds}(x) \Leftrightarrow$  **mit X fallende Chance**



## Binäre Einflussgrößen

- Die **Differenz** der beiden Werte einer binären Einflussgröße sollte **1** sein
- Beispiel einer **ungünstigen** Codierung
  - Geschlecht: **0** für **männlich**, **2** für **weiblich**
  - $\exp(\beta_1)$  ist das **Odds Ratio** zwischen **0** und **1**
  - aber **1** kommt **nicht** vor
- Beispiel für eine **gute** Codierung
  - Geschlecht: **1** für **weiblich**, **2** für **männlich**
  - Exposition: **0** für **nichtexponiert**, **1** für **exponiert**

# Kategoriale Einflussgrößen > 2 Ausprägungen

- Beispiel: Blutgruppe A, B, AB, 0
- Eine Kategorie wird Referenzkategorie
- Referenzkategorie =
  - Inhaltlichen Grund: Referenzpatient
  - häufigste Ausprägung → kleinste Varianz der Schätzer
  - In Europa: Blutgruppe A = Referenz
- Andere Kategorien (Blutgruppe B, AB, 0):  
umkodiert in eigene (Dummy-) Variable (SPSS: automatisiert)  
Blutgruppe B: ja vs. nein, AB: ja vs. nein, 0: ja vs. nein
- Odds Ratio: andere Kategorien vs. Referenz (B/A, AB/A, 0/A)

# A priori Bewertung von Einflussgrößen

1. Klinisch **etablierte** Einflüsse, **Confounder** →  
**Berücksichtigung** im Modell: **immer**
  2. Einflüsse mit **plausibel** vermutetem **kausalen** Zusammenhang  
→ **Berücksichtigung** im Modell: **möglichst**
  3. Identifizierung **neuer fraglich** wichtiger Einflüsse →  
**Berücksichtigung**: erst **zum Schluss**
- A priori **Bewertung**: **inhaltlich**, nicht statistisch
  - Falls **Einflüsse** inhaltlich **gleiches Gewicht** besitzen  
→ **statistische Modellwahl**



# Automatisierte Modellwahl

- Motivation: großer Aufwand bei der Modellwahl
- Automatisierte Regeln: Quick and dirty
- 3 Grundverfahren:
  1. Variablen gleichzeitig einschließen
  2. Forward Stepwise Selection = Bottom Up:  
Variablen Zug um Zug ins Modell aufnehmen
  3. Backward Stepwise Elimination = Top Down:  
Variablen Zug um Zug aus der Modell ausschließen

## Modellbildung - Bewertung

- Viele Auswahlverfahren für Modelle existieren
- Auswahlverfahren werden oft kontrovers diskutiert
- Automatisierte Verfahren:
  - wenig akzeptiert
  - in klinischen Studien oft benutzt
  - I. d. R. akzeptiert, falls: Forward = Backward
- Immer: aus inhaltlicher Sicht Konzept erforderlich
- Interaktionen bedenken
- Wichtig: in Publikationen Verfahren angeben und begründen!
- Mehrere Modelle: Alternativen diskutieren