

---

# **Regression**

---

**Dr. Dirk Hasenclever**

# Allgemeine Regressionsprobleme

---

Gegeben Daten :  $(y_i, z_i)$   $i = 1, \dots, n$

$z$  : Vektor von Kovariaten, die als gegeben angenommen werden.

$y$  : Realisierung einer zufälligen Zielgröße  $Y$ , deren Erwartungswert (und Varianz) von  $z$  abhängen kann.

Beispiel : Vorlesungsteilnehmer :

$y_i$  = Körpergröße Person  $i$

$z_i$  = (Körpergewicht, Alter, Geschlecht, Augenfarbe)

Fragen : - Welche der dokumentierten Einflußfaktoren sind relevant?  
- Wie wirken die relevanten ggfs. zusammen?  
- Wie gut läßt sich (bei weiteren Beobachtungen)  $Y$   
gegeben  $z$  vorhersagen?

# Allgemeine Regressionsprobleme II

---

Allgemeines Modell :  $Y(z) \sim \text{Verteilung}(\theta(z)),$

wobei die Parameter  $\theta(z)$  der Verteilung von  $z$  abhängen.

Ziel : Beschreibe der Einfluß von  $z$  auf  $Y$  durch  
Wahl einer geeigneten Verteilungsfamilie und  
einen Ansatz für die systematische Komponente  $\theta(z)$ .

# Multiple lineare Regression

---

Y metrisch.

Modell :  $Y(z) \sim \text{Normal}(\mu_0 + \beta^\top z; \sigma^2)$

Verteilung : Normalverteilung

Die Kovariaten haben einen linearen Einfluß  
auf den Erwartungswert :

$$\mu(z) = \mu_0 + \sum_j \beta_j \cdot z_j; \quad (\text{Linearer Prädiktor})$$

Die Varianz wird als unabhängig von z angenommen.

# Regression bei binären Daten I

---

$Y$  binär.

Modell :  $Y(z) \sim \text{Bernoulli}(g(\alpha + \beta^\top z))$

Verteilung : Bernoulli - Verteilung mit Wahrscheinlichkeit

$$p(z) = g(\alpha + \beta^\top z)$$

Der lineare Prädiktor  $\alpha + \beta^\top z$  wird i.A.  $\notin [0; 1]$  liegen.

Daher muss er durch eine monotone Linkfunktion

$g: [-\infty; +\infty] \rightarrow [0; 1]$  transformiert werden.

Die Varianz  $p(z) \cdot (1 - p(z))$  ist direkt durch  $p(z)$  gegeben.

# Regression bei binären Daten II

---

Mögliche Linkfunktionen :

a) Logistische Regression

$$g(x) = \frac{\exp(x)}{1 + \exp(x)}; \quad g^{-1}(p) = \ln\left(\frac{p}{1-p}\right) \text{ (Logodds)}$$

b) Probit Regression

$$g(x) = \Phi(x); \quad g^{-1}(p) = \Phi^{-1}(p)$$

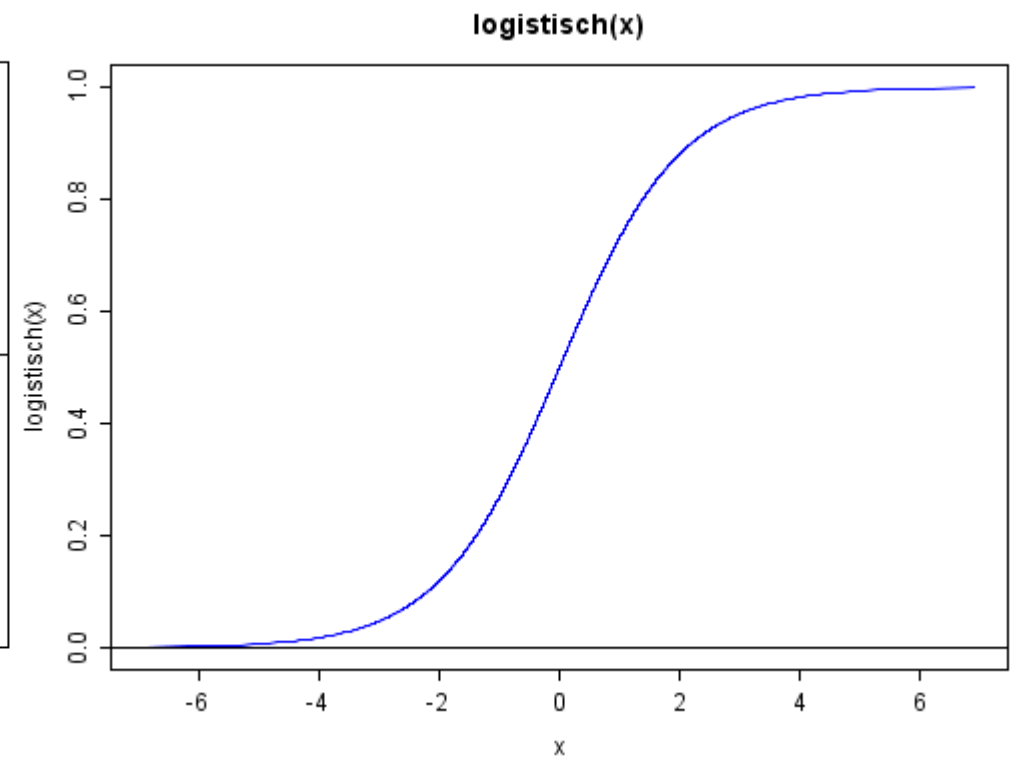
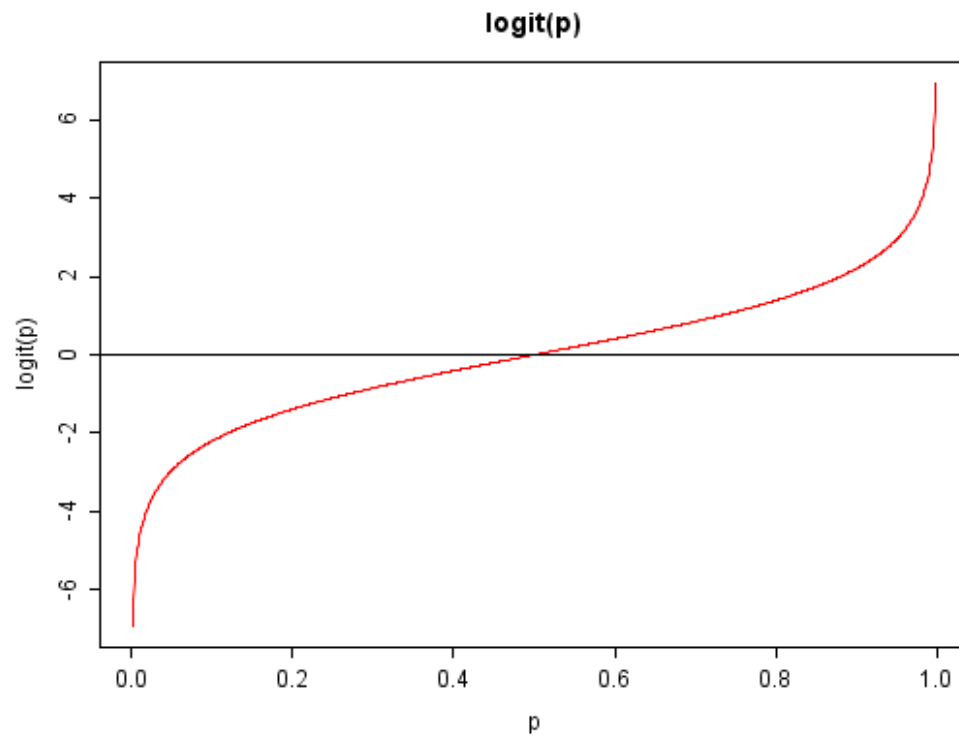
$\Phi$  Verteilungsfunktion der Standardnormalverteilung

c) Komplementäre log - log Regression

$$g(x) = 1 - \exp(-\exp(x)); \quad g^{-1}(p) = \ln(-\ln(1-p))$$

# Logit Scala I

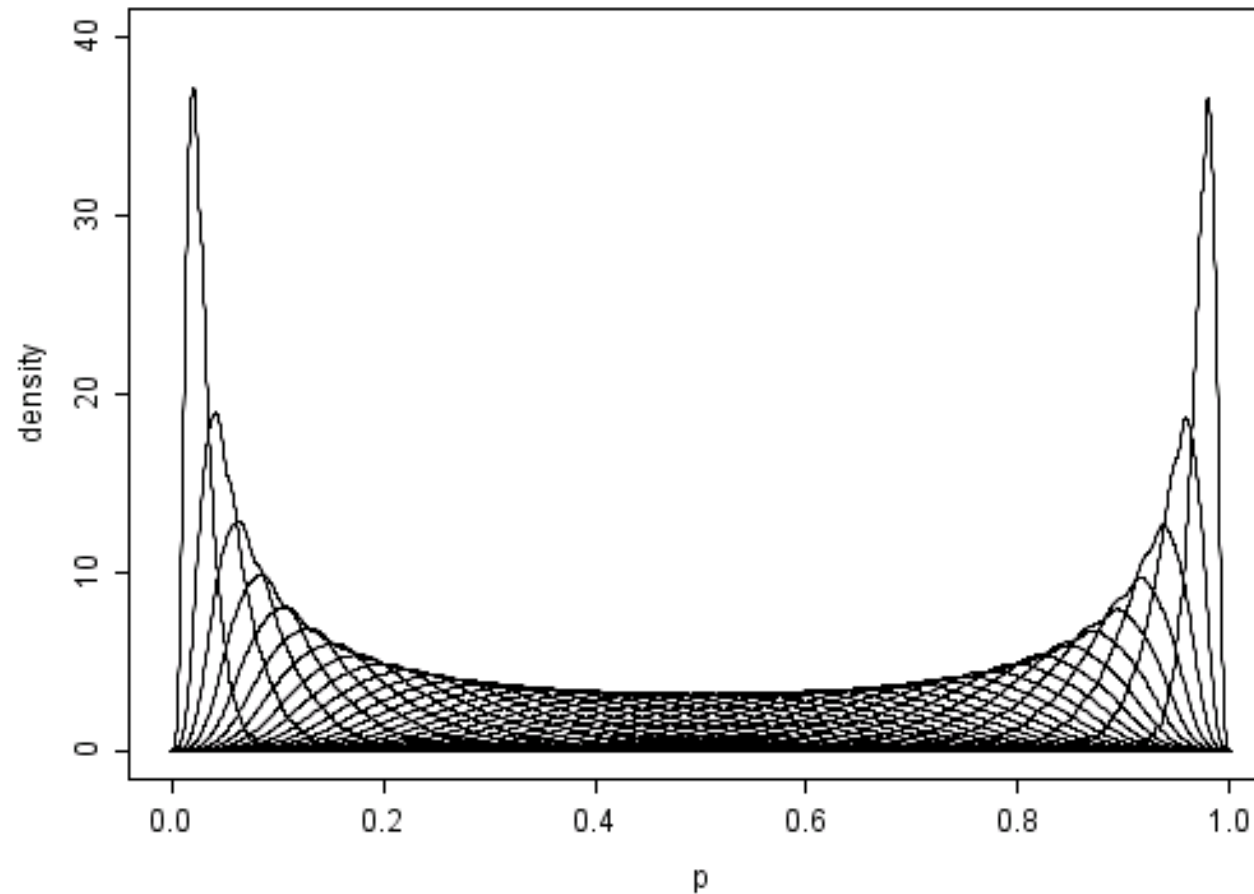
---



# Logit Scale II

---

Illustration logit scale



Random noise  
on logit scale

# Beispiel logistische Regression I

---

## ■ Klinisches Problem

- Hodgkin's Lymphom
- Klinisch nur eine Lymphknotenregion oberhalb des Zwerchfells befallen
- Diagnostische Bauchöffnung (Laparotomie) zur Abklärung eines möglichen abdominalen Befalls
- Daten: befallen= $j/n$  + diverse Kovariaten: Alter, Geschlecht, Mediastinum befallen= $j/n$ , Histologie, extranodaler Befall= $j/n$

# Beispiel logistische Regression II

Table V-2. Prognostic factors of positive laparotomy findings in CS IA supradiaphragmatic patients (N=1,169)

Patients with positive laparotomy (PS III-IV): n = 287 (24.6 %)

## STEPWISE LOGISTIC REGRESSION

Parameters included in the model (number of dummy variables) = sex, age (4), mediastinal involvement (2), histological subtype (2), extra nodal localization.

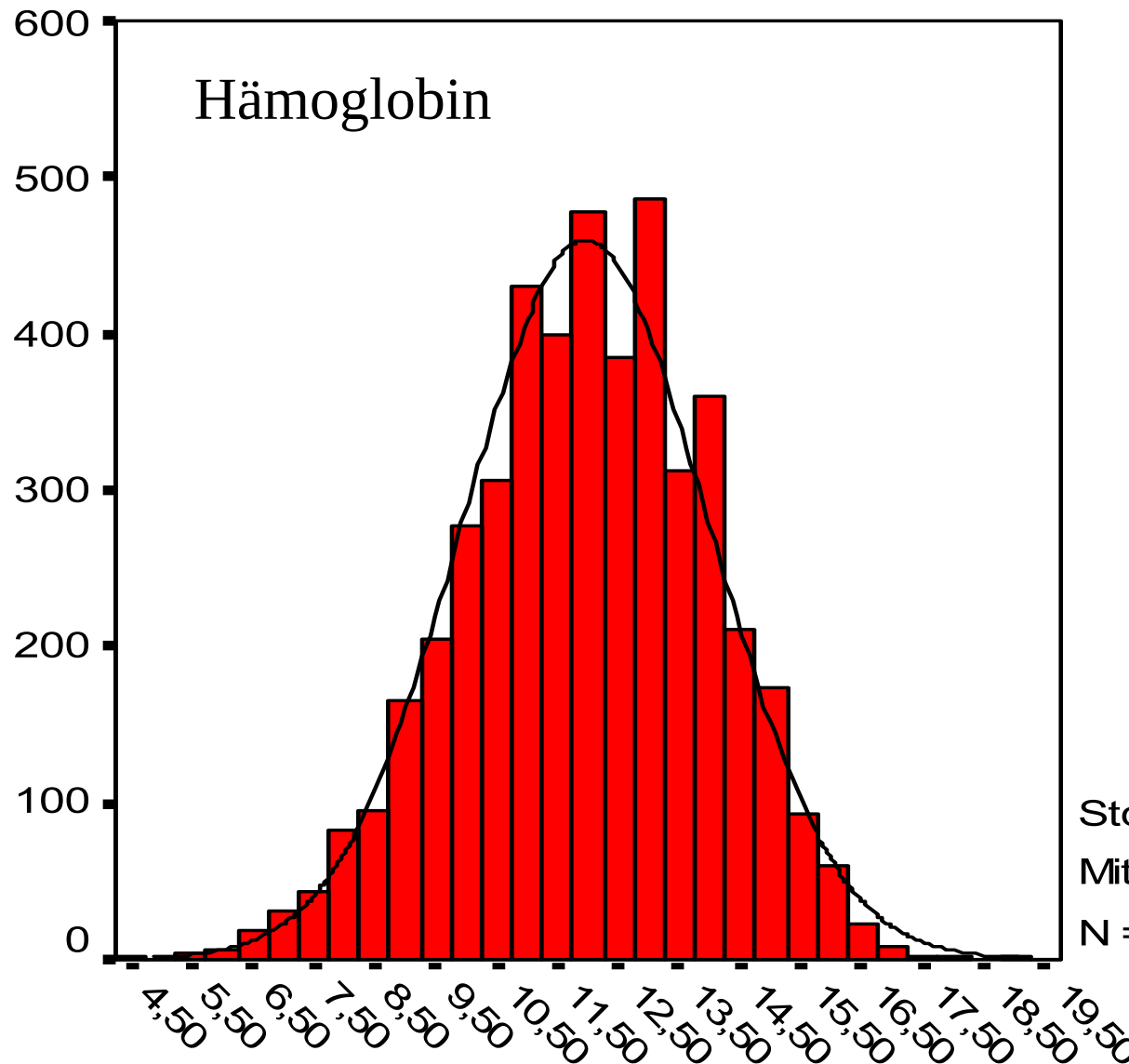
| Step | Variable (codes)                    | log.<br>likelihood | Coeff/s.e.    | p value |
|------|-------------------------------------|--------------------|---------------|---------|
| 0    |                                     | - 649.84           |               |         |
| 1    | SEX (M=-1, F=1)                     | - 635.03           | - 0.396/0.084 | < 0.001 |
| 2    | HISTOLOGY (LP-NS=-1, MC-LD-uncl.=1) | - 625.90           | + 0.295/0.070 | < 0.001 |
| 3    | AGE (<50=-1, 50+=1)                 | - 622.50           | + 0.243/0.091 | 0.009   |
|      | CONSTANT                            |                    | - 1.079/0.102 |         |

**Henry-Amar et al. (IDHD) 1990**

# Beispiel logistische Regression III

| Pattern                | number of patients | observed proportion of LAP + | predicted probability of LAP + |
|------------------------|--------------------|------------------------------|--------------------------------|
| Male, 50+ yrs, MC-LD   | 50                 | 0.38                         | 0.46                           |
| Female, 50+ yrs, MC-LD | 22                 | 0.36                         | 0.28                           |
| Male, <50 yrs, MC-LD   | 276                | 0.36                         | 0.35                           |
| Male, 50+ yrs, LP-NS   | 66                 | 0.32                         | 0.32                           |
| Female, 50+ yrs, LP-NS | 35                 | 0.26                         | 0.18                           |
| Male, <50 yrs, LP-NS   | 402                | 0.23                         | 0.23                           |
| Female, <50 yrs, MC-LD | 76                 | 0.17                         | 0.19                           |
| Female, <50 yrs, LP-NS | 236                | 0.11                         | 0.12                           |

Henry-Amar et al. (IDHD) 1990



Beispiel multiple  
lineare Regression

Std.abw . = 2,02

Mittel = 11,96

N = 4663,00

RHB

Daten des Internationalen prognostischen Faktorprojekts Hasenclever 19



# HB nach Geschlecht (T-Test)

Gruppenstatistiken

|     | MALE | N    | Mittelwert | Standardabweichung | Standardfehler des Mittelwertes |
|-----|------|------|------------|--------------------|---------------------------------|
| RHB | ,00  | 1821 | 11,3444    | 1,71457            | ,04018                          |
|     | 1,00 | 2842 | 12,3540    | 2,10009            | ,03939                          |

Test bei unabhängigen Stichproben

|     |                             | Levene-Test der Varianzgleichheit |             | T-Test für die Mittelwertgleichheit |          |                 |                    |                              |                                      |         |
|-----|-----------------------------|-----------------------------------|-------------|-------------------------------------|----------|-----------------|--------------------|------------------------------|--------------------------------------|---------|
|     |                             | F                                 | Signifikanz | T                                   | df       | Sig. (2-seitig) | Mittlere Differenz | Standardfehler der Differenz | 95% Konfidenzintervall der Differenz |         |
|     |                             |                                   |             |                                     |          |                 |                    |                              | Untere                               | Obere   |
| RHB | Varianzen sind gleich       | 87,765                            | ,000        | -17,173                             | 4661     | ,000            | -1,0096            | ,05879                       | -1,12486                             | -,89435 |
|     | Varianzen sind nicht gleich |                                   |             | -17,942                             | 4397,593 | ,000            | -1,0096            | ,05627                       | -1,11992                             | -,89929 |

Daten des Internationalen prognostischen Faktorprojekts Hasenclever 1998

# HB nach Stadium IV (T-Test)

Gruppenstatistiken

|     | IV   | N    | Mittelwert | Standardabweichung | Standardfehler des Mittelwertes |
|-----|------|------|------------|--------------------|---------------------------------|
| RHB | ,00  | 2680 | 12,2398    | 1,90526            | ,03680                          |
|     | 1,00 | 1983 | 11,5812    | 2,10620            | ,04730                          |

Test bei unabhängigen Stichproben

|     |                             | Levene-Test der Varianzgleichheit |             | T-Test für die Mittelwertgleichheit |          |                 |                    |                              |                                      |
|-----|-----------------------------|-----------------------------------|-------------|-------------------------------------|----------|-----------------|--------------------|------------------------------|--------------------------------------|
|     |                             | F                                 | Signifikanz | T                                   | df       | Sig. (2-seitig) | Mittlere Differenz | Standardfehler der Differenz | 95% Konfidenzintervall der Differenz |
|     |                             |                                   |             |                                     |          |                 |                    |                              | Untere      Obere                    |
| RHB | Varianzen sind gleich       | 24,185                            | ,000        | 11,155                              | 4661     | ,000            | ,6586              | ,05904                       | ,54284      ,77433                   |
|     | Varianzen sind nicht gleich |                                   |             | 10,989                              | 4018,736 | ,000            | ,6586              | ,05993                       | ,54109      ,77608                   |

Daten des Internationalen prognostischen Faktorprojekts Hasenclever 1998

# Lineare Regression

$$HB = \text{Konstante} + B1 \cdot \text{MALE} + B2 \cdot \text{IV} + B3 \cdot \text{IV} \cdot \text{MALE} + \text{Fehler}$$

Koeffizienten<sup>a</sup>

| Modell |             | Nicht standardisierte Koeffizienten |                | Standardisierte Koeffizienten | T       | Signifikanz |
|--------|-------------|-------------------------------------|----------------|-------------------------------|---------|-------------|
|        |             | B                                   | Standardfehler | Beta                          |         |             |
| 1      | (Konstante) | 11,512                              | ,059           |                               | 194,123 | ,000        |
|        | MALE        | 1,201                               | ,076           | ,290                          | 15,759  | ,000        |
|        | IV          | -,401                               | ,092           | -,098                         | -4,377  | ,000        |
|        | IVMALE      | -,437                               | ,117           | -,095                         | -3,733  | ,000        |

a. Abhängige Variable: rHb (+estimates from HCT)

Modellzusammenfassung

| Modell | R                 | R-Quadrat | Korrigiertes R-Quadrat | Standardfehler des Schätzers |
|--------|-------------------|-----------|------------------------|------------------------------|
| 1      | ,298 <sup>a</sup> | ,089      | ,088                   | 1,92810                      |

Fehler ~ Normal(0,  $\sigma^2$ )

$\sigma^{\wedge}$

a. Einflußvariablen : (Konstante), IVMALE, MALE, IV

Daten des Internationalen prognostischen Faktorprojekts Hasenclever 1998

# Modelle im Vergleich

| Modell        | Schätzer $\pm$ se |
|---------------|-------------------|
| const         | 11,35 $\pm$ 0,05  |
| male (61%)    | 1,01 $\pm$ 0,06   |
|               |                   |
| const         | 11,63 $\pm$ 0,05  |
| male          | 1,02 $\pm$ 0,06   |
| iv (41%)      | -0,67 $\pm$ 0,06  |
|               |                   |
| Const         | 11,52 $\pm$ 0,06  |
| Male          | 1,20 $\pm$ 0,08   |
| Iv            | -0,41 $\pm$ 0,09  |
| iv*male (25%) | -0,43 $\pm$ 0,12  |

Interaktions-  
term 

# Instabilität des Interaktionsterms

4 x Zufallsauswahl N=200

| Modell  | Schätzer ± se | Schätzer ± se | Schätzer ± se | Schätzer ± se |
|---------|---------------|---------------|---------------|---------------|
| const   | 11,37 ± 0,19  | 11,09 ± 0,22  | 11,21 ± 0,22  | 11,61 ± 0,22  |
| male    | 0,91 ± 0,24   | 1,24 ± 0,27   | 1,15 ± 0,28   | 0,59 ± 0,28   |
|         |               |               |               |               |
| const   | 11,67 ± 0,21  | 11,33 ± 0,25  | 11,46 ± 0,24  | 11,80 ± 0,23  |
| male    | 0,90 ± 0,24   | 1,22 ± 0,27   | 1,20 ± 0,26   | 0,69 ± 0,28   |
| iv      | -0,63 ± 0,23  | -0,47 ± 0,25  | -0,68 ± 0,26  | -0,58 ± 0,28  |
|         |               |               |               |               |
| Const   | 11,49 ± 0,26  | 11,24 ± 0,31  | 11,29 ± 0,28  | 11,62 ± 0,26  |
| Male    | 1,18 ± 0,32   | 1,36 ± 0,38   | 1,46 ± 0,35   | 1,02 ± 0,36   |
| Iv      | -0,56 ± 0,37  | -0,28 ± 0,43  | -0,22 ± 0,46  | -0,02 ± 0,46  |
| iv*male | -0,60 ± 0,47  | -0,28 ± 0,53  | -0,68 ± 0,56  | -0,86 ± 0,57  |

---

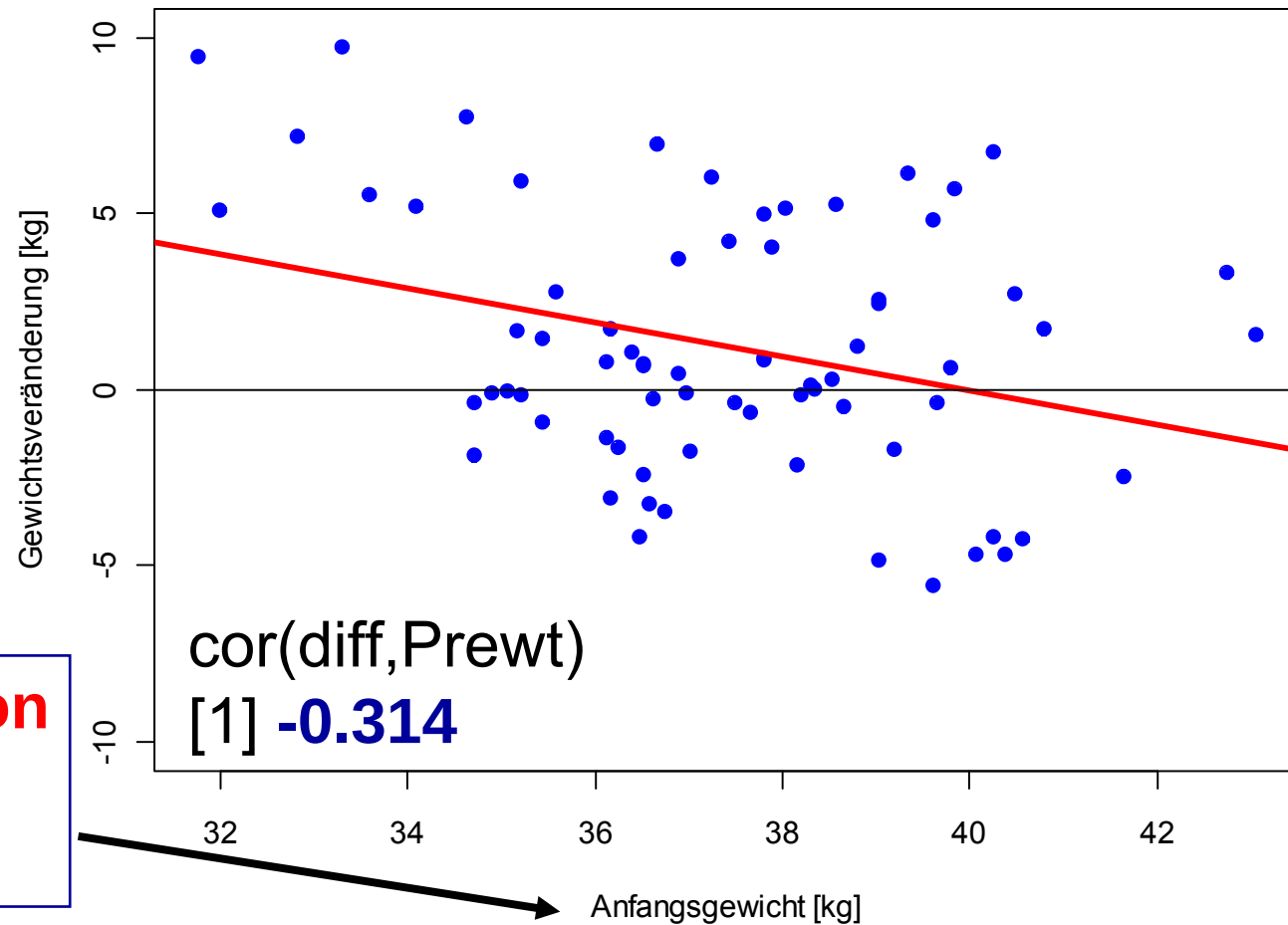
# **Vorher-Nachher Daten**

---

# Regression to the mean

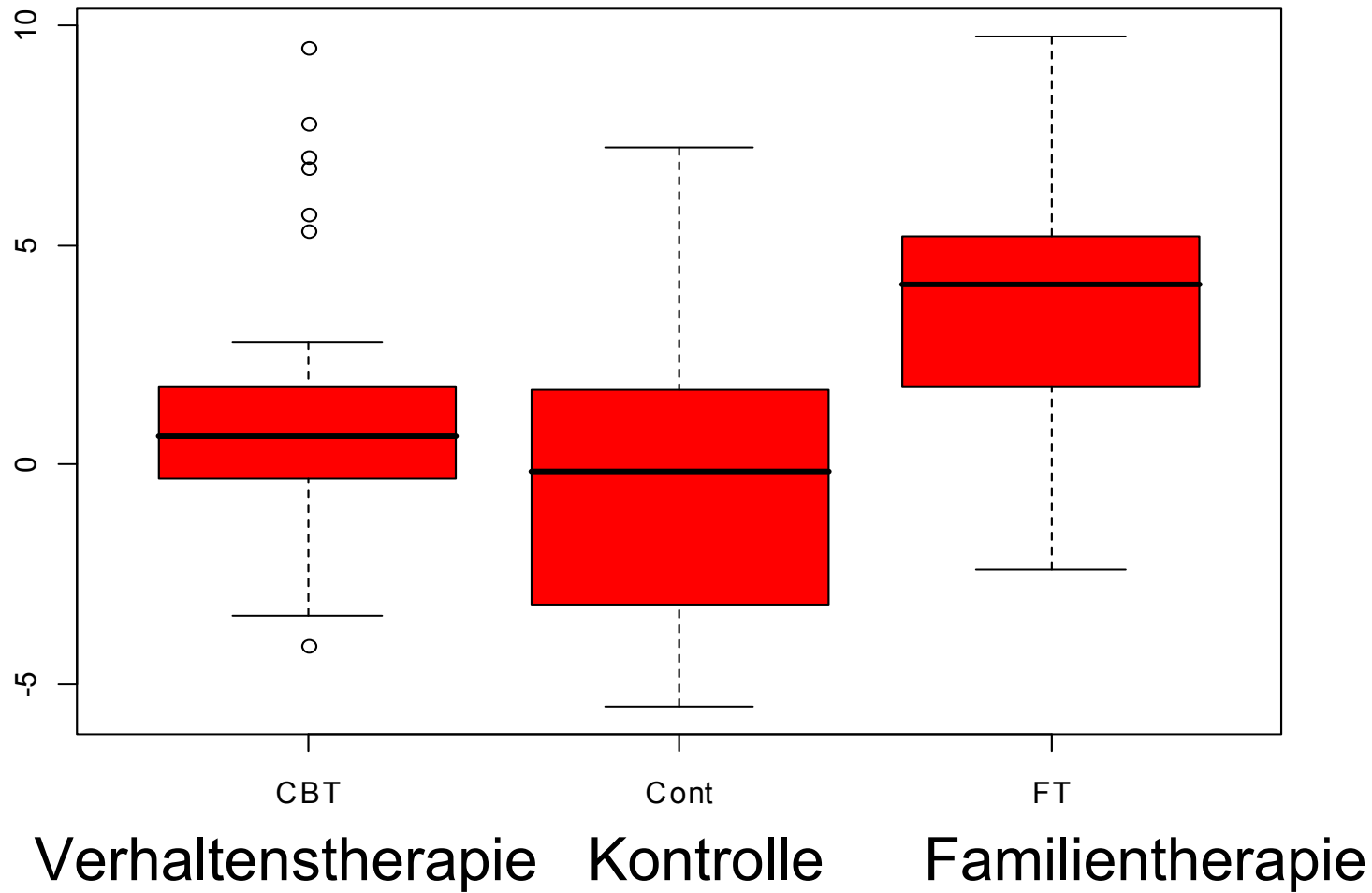
Vorher -> Nachher

Patientinnen  
mit Magersucht



**Qua Konstruktion  
abhängig  
VORSICHT**

# Gewichtsveränderung nach Therapieform



# Modellansatz

---

*(Nachher – Vorher) =*

*const +  $\beta \cdot \text{Vorher}$  +  $\gamma \cdot \text{Therapieindikator}$  + Fehlerterm*

## Coefficients:

|             | Estimate | Std. Error | t value | Pr(> t ) |     |
|-------------|----------|------------|---------|----------|-----|
| (Intercept) | 22.5682  | 6.0736     | 3.716   | 0.000411 | *** |
| Prewt       | -0.5653  | 0.1612     | -3.507  | 0.000807 | *** |
| TreatCont   | -1.8601  | 0.8588     | -2.166  | 0.033820 | *   |
| TreatFT     | 2.0699   | 0.9676     | 2.139   | 0.036012 | *   |

# Vergleich mit T-Tests auf Differenzen

---

Welch Two Sample t-test

```
data: diff[Treat == "Cont"] and diff[Treat == "CBT"]
t = -1.6708, df = 50.976, p-value = 0.1009
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 -3.4577094  0.3165953
sample estimates:
 mean of x  mean of y
-0.2053846  1.3651724
```

---

```
data: diff[Treat == "FT"] and diff[Treat == "CBT"]
t = 1.9323, df = 34.225, p-value = 0.06164
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 -0.09944705  3.96204340
sample estimates:
 mean of x  mean of y
```

---

3.296471 1.365172

Bessere Datenbeschreibung und  
höhere statistische Power  
mittels Regressionsmodell

# COX-Regression I

---

Regression auf Überlebenszeiten

$$Y(z) \sim S[z](t)$$

(Theoretische) Überlebenskurve abhängig von Kovariatenkonstellation

Wie kann man Einfluß der Kovariaten auf Überlebenskurve spezifizieren?

- a) vollparametrisch : z.B. modelliere Einfluß der Kovariablen auf  
Parameter der Weibullverteilung
- b) semi - parametrisch unter Annahme, dass Unterschiede  
dem Proportional Hazard Modell genügen :
  - Schätze gemeinsame Basis - Hazardfunktion
  - beschreibe, wie Kovariaten die Basis - Hazardfunktion  
proportional beeinflussen.

# COX-Regression II

---

Notation : Daten  $(t_j, \delta_j, z_j)$   $j := 1, \dots, n$

$t_j$  Zeitdauer,  $\delta_j$  Zensierungsstatus,  $z_j$  Vektor von Kovariaten

Ansatz :  $h(t, z) = h_0(t) \cdot \exp(\beta^T z)$   
Hazardfunktion = Baseline Hazard  $\cdot$  Linkfunktion( Linearer Prädiktor )

$$S(t|z) = S_0(t)^{\exp(\beta^T z)} \quad \text{mit} \quad S_0(t) = \exp\left(-\int_0^t h_0(\tau) d\tau\right)$$

# COX-Regression III

---

Hazardratios werde als konstant über die Zeit angenommen.

Patienten mit Kovariaten  $z_0$  und  $z_1$

$$\frac{h(t|z_1)}{h(t|z_0)} = \frac{h_0(t) \cdot \exp(\beta^T z_1)}{h_0(t) \cdot \exp(\beta^T z_0)} = \exp(\beta^T (z_1 - z_0))$$

Bei einem Therapievergleich :

Therapieindikator :  $z = \begin{cases} 0 & \text{Standardtherapie} \\ 1 & \text{experimentelle Therapie} \end{cases}$

$$\exp(\beta) = \frac{h(t|\text{experimentelle Therapie})}{h(t|\text{Standardtherapie})} \quad \text{"relatives risiko"}$$

$$\beta : \quad \text{"Loghazard Ratio"} \quad S_1(t) = S_0(t)^{\exp(\beta)}$$

# COX-Regression IV

---

Maximum Likelihood Schätzung zur Anpassung eines COX - Modells an gegebene Daten :

Zur numerischen Vereinfachung separiert man die Schätzung der Koeffizienten  $\beta$  von der Schätzung der Baseline Hazardfunktion  $h_0(t)$  mittels Einführung einer partiellen Likelihoodfunktion :

Der Patient  $i$  mit Kovariate  $z_i$  sterbe zur Zeit  $t$ .

Betrachte die Wahrscheinlichkeit, dass genau dieser Patient  $i$  zu  $t$  stirbt, gegeben, dass genau ein Patient zu  $t$  stirbt :

$$pr(T_i = t \mid d(t) = 1) = \frac{h(t|z_i)}{\sum_{j \in R(i)} h(t|z_j)} = \frac{h_0(t) \cdot \exp(\beta^T z_i)}{\sum_{j \in R(i)} h_0(t) \cdot \exp(\beta^T z_j)} = \frac{\exp(\beta^T z_i)}{\sum_{j \in R(i)} \exp(\beta^T z_j)}$$

# COX-Regression V

---

Definiere die partielle Likelihoodfunktion :

$$L_{\text{partiell}}(\beta) = \prod_{i=1}^D \frac{\exp(\beta^T z_i)^{\delta_i}}{\sum_{j \in R(i)} \exp(\beta^T z_j)}$$

wobei  $i$  über alle Ereigniszeitpunkte läuft.

Diese partielle Likelihoodfunktion beschreibt gerade, wie gut die  $\beta$  erklären, welche Patienten unter denen "at risk" das Ereignis bekommen.

Die Schätzung der Regressionskoeffizienten  $\beta$  erfolgt durch Maximierung der partiellen Likelihood.

# COX-Regression VI

---

Alternativ bekommt man diesen Ansatz auch als sogenannte Profile - Likelihood :

Volle Likelihood - Funktion :  $L(\beta, h_0(t)) \propto \prod_{i=1}^n h(t_i|z_i)^{\delta_i} \cdot S(t_i|z_i)^{1-\delta_i}$

$$L_{profile}(\beta) = \sup_{h_0(t)|\beta} (L(\beta, h_0(t)))$$

(Supremum bzgl. Funktionen  $h_0(t)$  aus einer Funktionenmenge bei jeweils festem  $\beta$ .)

Setzt man  $S_0$  als monoton fallende Treppenfunktion an mit Sprungstellen genau in den Ereigniszeitpunkten, so erhält man :

$$L_{profile}(\beta) \propto L_{partiell}(\beta)$$

(Übergang zur Profile - Likelihood ist ein Standardtrick zur Elimination von Störgrößen.)

# Allgemeine Sätze über Likelihood (informell)

---

- Maximum Likelihood Schätzer (MLE) sind i.A. asymptotisch multivariat normalverteilt:

$$\hat{\beta} \sim \text{Normal}[\beta, I^{-1}(\beta)]$$

$$I(\beta) = E(J(\beta)) \quad \text{mit} \quad J(\beta) = -\left(\frac{\partial^2 l(\beta)}{\partial \beta_i \partial \beta_j}\right)_{i,j} \quad l = \text{Loglikelihood}$$

$$I(\beta) \approx I_{\text{obs}} = -\left(\frac{\partial^2 l(\beta)}{\partial \beta_i \partial \beta_j}\right)_{i,j} \Big|_{\hat{\beta}} \quad \text{Observed information matrix}$$

# Heuristik dazu

---

- Taylor-Entwicklung im Maximum  $\beta$ :
- $\text{Loglikelihood}(\beta) \approx$   
 $\text{Matrix\_zweite\_Ableitungen}(\beta) * (\hat{\beta} - \beta)^2$   
(in kleiner Umgebung)
- Je gekrümmter die Loglikelihood desto genauer ist  $\hat{\beta}$  durch die Daten bestimmt.
- Also Standardfehler<sup>2</sup>  $\sim$   
 $1 / \text{Matrix\_zweite\_Ableitungen}(\beta)$

# COX-Regression VII

---

Der Vektor von Regressionskoeffizienten  $\beta$   
wird durch Maximierung  
der partiellen Likelihood geschätzt :  $\hat{\beta}_{ML}$

Die allgemeinen Sätze über Maximum - Likelihood - Schätzer  
übertragen sich  
auf die partielle Likelihood der COX - Regression.

# COX-Regression VII

---

Gegeben  $\hat{\beta}$  kann man in einem zweiten Schritt die Baseline Überlebensfunktion (für  $z = 0$ ) schätzen :

$$\hat{S}(t|0) = \prod_{t_i < t} \left( 1 - \frac{\delta_i \cdot \exp(\hat{\beta}^T z_i)}{\sum_{j \in R(t_i)} \exp(\hat{\beta}^T z_j)} \right)$$

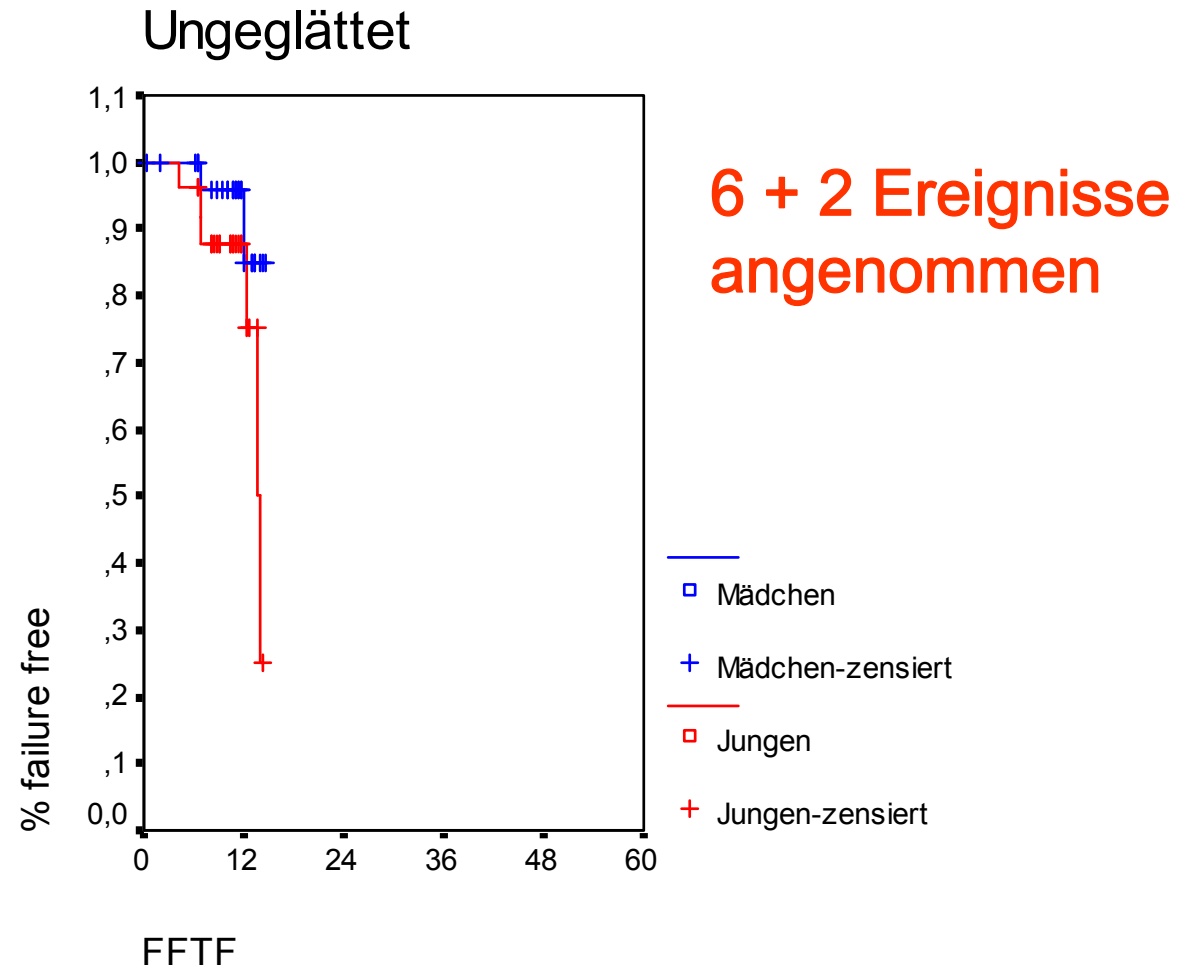
wobei  $R(t_i)$  die Menge der Patienten unter Risiko zum Zeitpunkt  $t_i$  bezeichnet.

$$\hat{S}(t|z) = \hat{S}(t|0)^{\exp(\hat{\beta}^T z)} \text{ liefert dann Modell - Schätzer}$$

für die Untergruppe von Patienten mit Kovariatenvektor  $z$ .

(Diese sind im Vergleich zum Kaplan - Meier Schätzer für die Untergruppe durch das Modell geglättet, da in die Schätzung der Baseline - Überlebenskurve alle Daten mittels der Modellannahme "Proportional Hazard" eingehen.)

# Unreife Kaplan-Meier Kurven (ziemlich instabil)



# Glätten und Vergleich mit Referenzkurve im Cox-Modell

---

- Nehme Referenzdaten in Cox-Modell hinzu.
- Schätzungen werden durch angenommenem bekanten Baseline-Hazard geglättet.
- Aus den Koeffizienten Schätzern kann man Test gegen Referenzkurve mit vorgegebenem Plateau ableiten.

# Geglättete Kurven im Vergleich zur Referenz

