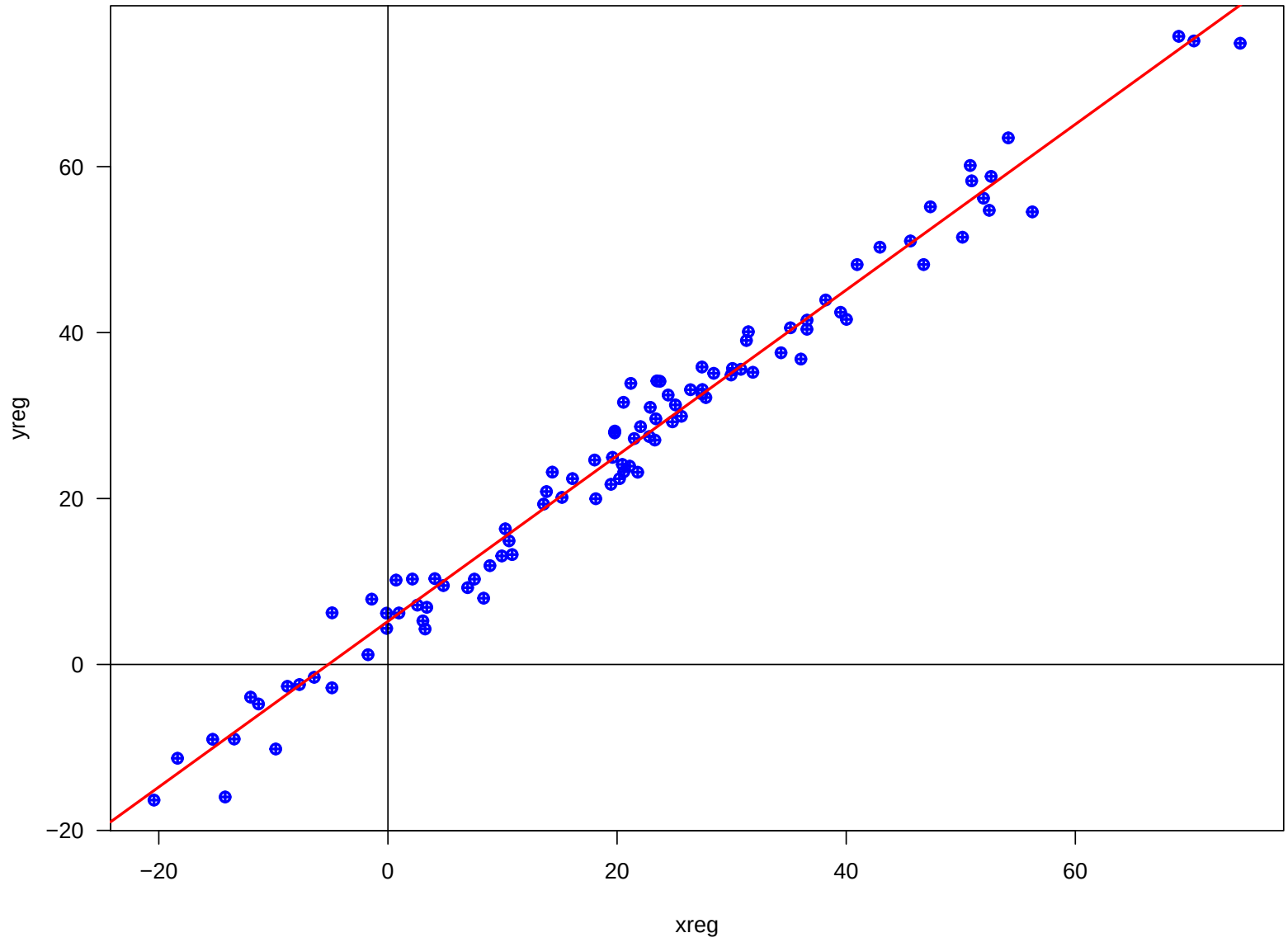
Artificial Intelligence - Data Science

Modellierung und Prädiktion mit Regression und verallgemeinerter Regression

Alexandra Posekany

WS 2020

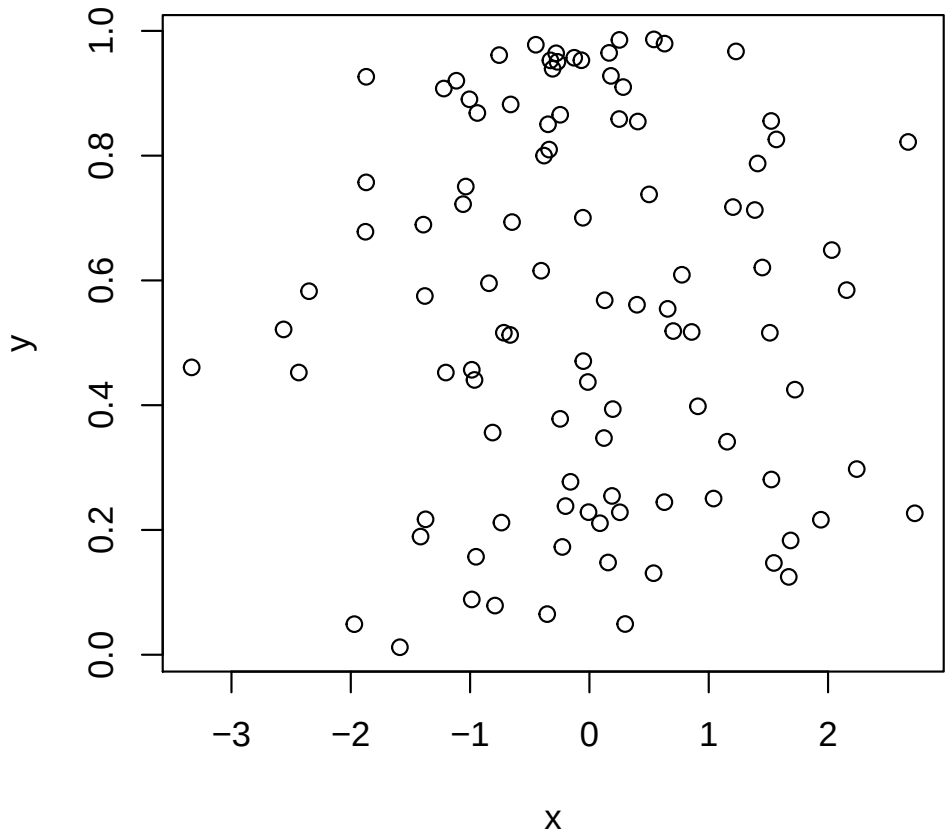
Modellierung und Prädiktion mit linearer Regression



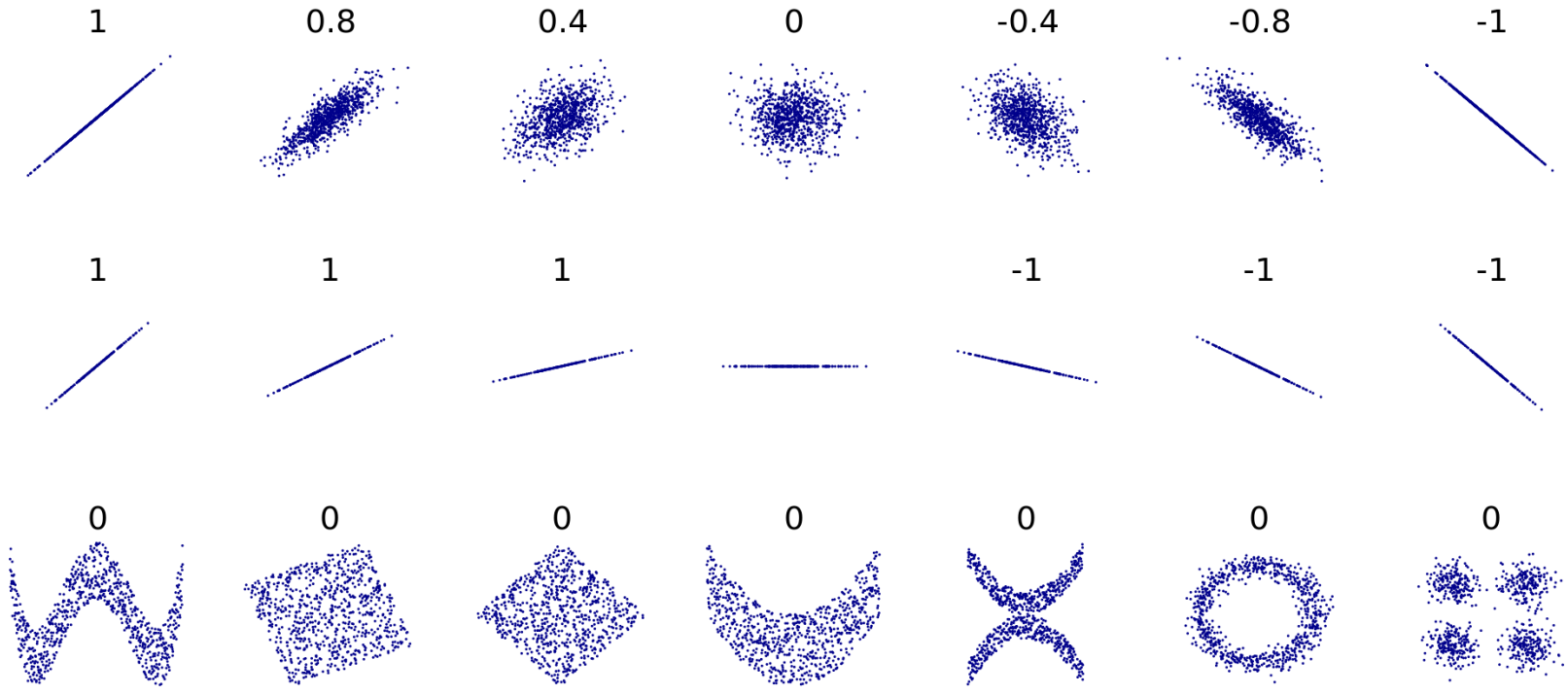
Zusammenhang zwischen zwei numerischen Variabeln

Der **Scatterplot** visualisiert den Zusammenhang zwischen 2 metrischen Variablen, von denen eine entlang der x-Achse, die andere entlang der y-Achse aufgetragen wird.

```
plot(dflm)
```



Scatterplot: Zusammenhänge erkennen (?)



Korrelation und Kovarianz

Kovarianz und Korrelation

- Der Pearson Korrelationskoeffizient

$$r = r(X, Y) = \frac{\widehat{\text{cov}}(X, Y)}{s(X) \cdot s(Y)},$$

misst den linearen Zusammenhang zwischen 2 metrischen Variablen X und Y, der sich auf Basis der Kovarianz

$$\widehat{\text{cov}}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}).$$

berechnet.

```
cor(dflm$x,dflm$y, method = c('pearson'))
```

Spearman Korrelation

Spearman Korrelation

Die **Spearman's rank correlation**

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)},$$

misst den linearen Zusammenhang zwischen 2 metrischen oder ordinalen Variablen X und Y, der sich auf Basis der Ränge von Beobachtungen.

Vorteil:

für **ordinale** und **metrische** Variablen anwendbar

robust gegenüber Ausreißern

```
cor(dflm$x,dflm$y, method = c('spearman'))
```

Interpretation der Korrelation

Werte für Korrelation liegen zwischen -1 und 1

negatives Vorzeichen = fallender Zusammenhang

positives Vorzeichen = steigender Zusammenhang

Faustregel zur Interpreation des Korrelationskoeffizient:

$r_{(s)} = 0$	keine Korrelation
$0 < \|r_{(s)}\| \leq 0.5$	schwache Korrelation
$0.5 < \|r_{(s)}\| \leq 0.75$	mittlere Korrelation
$0.75 < \|r_{(s)}\| < 1$	starke Korrelation
$\|r_{(s)}\| = 1$	vollständige Korrelation

Fehlgeleitete Korrelation

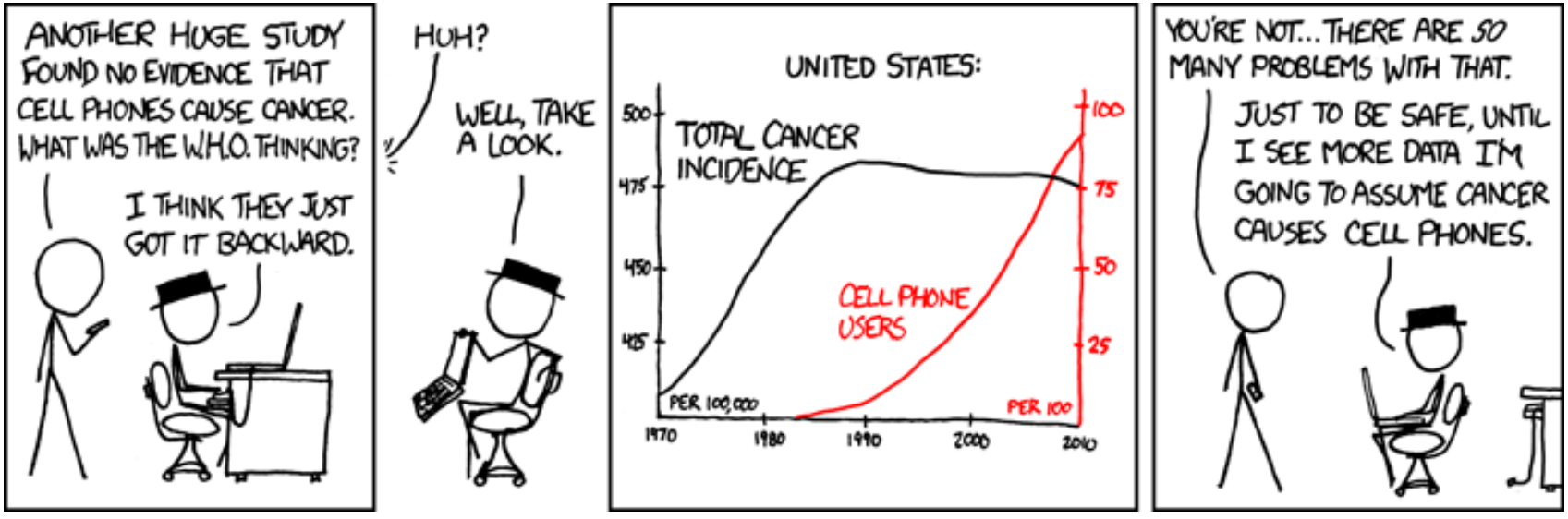
AIDS Infektionen und Handynutzer in der Schweiz

Jahr	1995	1996	1997	1998	1999
AIDS Infektionen	736	542	565	422	262
Handynutzer (in 1000)	447	663	1044	1699	3058

- Pearson Korrelation $r = -0.94$
  Spearman Korrelation $r_S = 0.9$
- Schützen Handies vor AIDS?
- Oder trügt die Zeit?

from Duller: Einführung in die Statistik mit EXCEL und SPSS (2007)

Fooled by correlation - Korrelation ist nicht Kausalität!



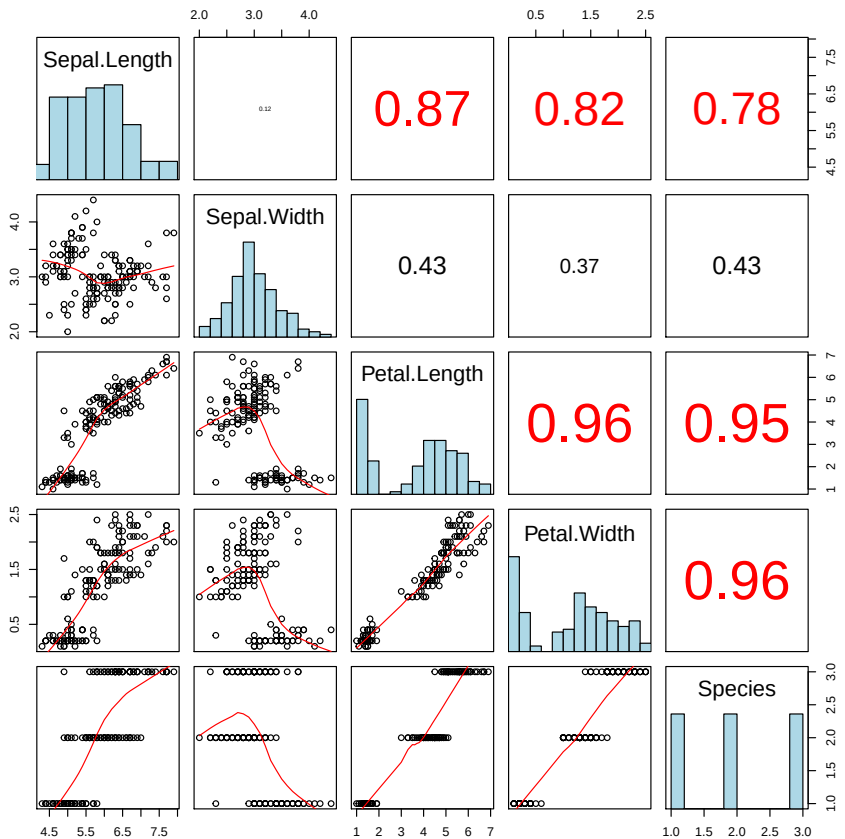
- Korrelation ist nur ein Maß für den **linearen** Zusammenhang!
- Korrelation **impliziert KEINE Kausalität** ("Babies und Störche")!

Pairwise Scatterplot Matrix

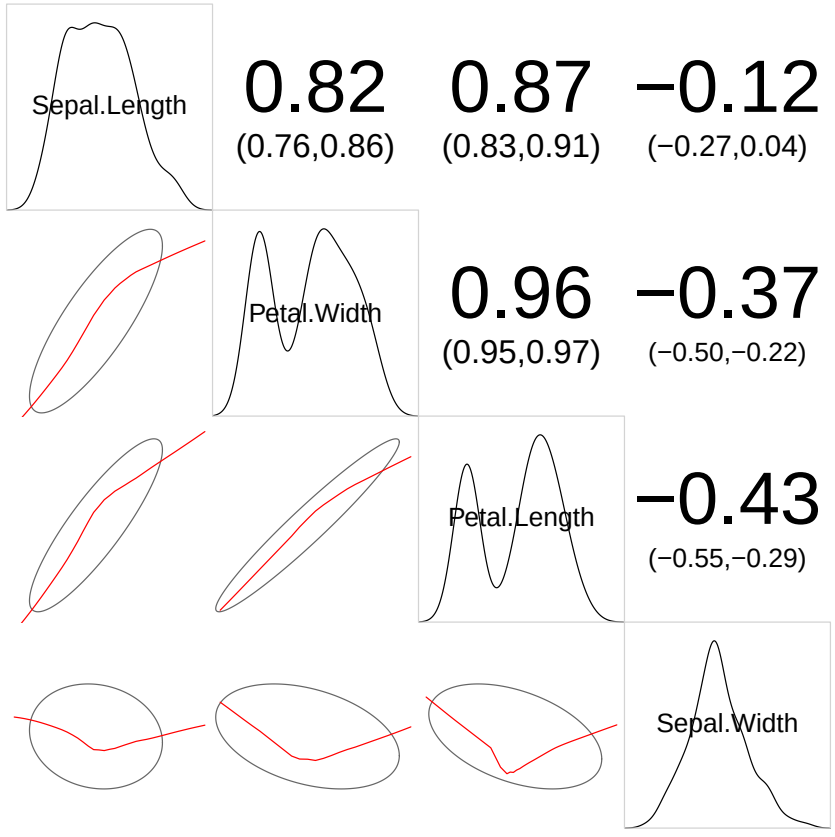
Pairwise Scatterplot Matrix

Die **Pairwise Scatterplot Matrix** dient der Darstellung der paarweisen Zusammenhänge zwischen je 2 Variablen.

```
pairs(iris)
```



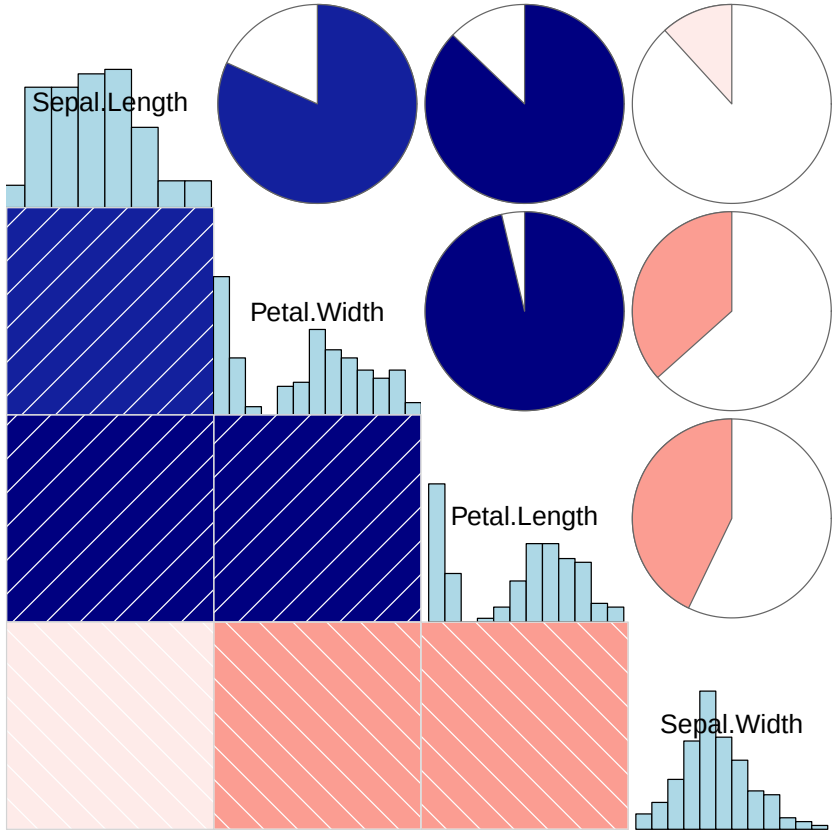
Variante der Pairwise Scatterplot Matrix



Korrelogramm

Ein **Korrelogramm** dient der kompakten Darstellung der Korrelationsmatrix zwischen je 2 Variablen.

```
library(corrgram); corrgram(dflm)
```



Lineare Regression - einfaches univariates Modell

(lineare) Regressionsmodelle modellieren den Zusammenhang zwischen

- einer **abhängigen** numerischen Variable, **Regressanden** Y, und
- einer (oder mehrerer) **unabhängiger** erklärender numerischer Variablen, **Regressoren** X, $\boldsymbol{X}$

Mathematisch wird das einfache lineare Regressionmodell geschrieben als

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

- Beobachtungen $(x_i|y_i)$, $i = 1, \ldots, n$
- Achsenabschnitt (Intercept) α und der Regressionkoeffizient (Steigung) β
- Residuen (Fehler) ε_i sind der Vertikalabstand der Punkte $(x_i|y_i)$ von der Gerade $y = \alpha + \beta \cdot x$

Regressionsmodell vs. Modellgleichung

Das Regressionsmodell

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

erklärt die beochachteten Werte y_i durch die beochachteten Werte x_i unter Berücksichtigung der einzelnen Fehler ε_i.

Durch numerischen Schätzung erhalten die tatsächliche Modellgleichung der Regressionsgerade

$$y = \hat{\alpha} + \hat{\beta} x$$

$\hat{\alpha}$ und $\hat{\beta}$ sind die konkret geschätzten Werte für den Achsenabschnitt und die Steigung der Geraden.

Interpretation der Regressionsparameter im Sachkontext

- Achsenabschnitt $\hat{\alpha}$
  Der mittlere Wert der abhängigen Variable y, wenn der Wert der Variable x=0 ist.
- Regressionskoeffizient $\hat{\beta}$, also die Steigung der Geraden
  Der mittlere Wert, um den die abhängige Variable y steigt/fällt, wenn der Wert der Variable x um 1 Einheit steigt.

Volumen eines Gases (in m^3, abhängig von der Temperatur in °C.

$$V = 0.0178 \cdot T + 3.65$$

3.65 m^3 ist das mittlere Volumen des Gases, wenn die Temperatur 0°C beträgt.

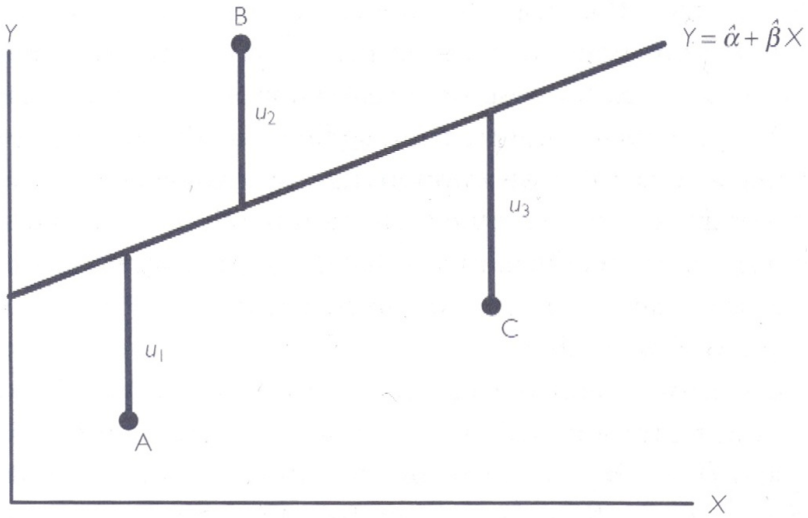
Wenn die Temperatur um 1°C steigt, dann steigt das Volumen des Gases um 0.0178 m^3.

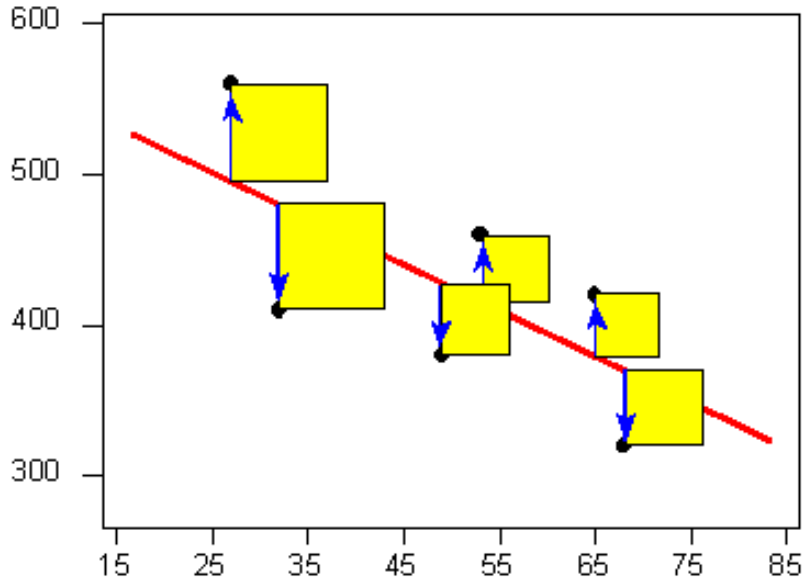
Kleinste Quadrate Schätzung

Die Kleinste Quadrate Schätzung (Least Squares Estimation) der Regression bedeutet, dass die Summe der quadratischen Residuen (residual sum of squares RSS)

$$e_i = (y_i - \hat{\alpha} - \hat{\beta} y_i)^2, i = 1, 2, \ldots, N.$$

minimiert wird, um $\hat{\alpha}$ und $\hat{\beta}$ zu schätzen.





Linear Regression - OLS estimate

Das univariate einfache Regressionsmodell hat die koeffizienten

$$\begin{aligned} \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} \\ \hat{\beta} &= r_{xy}\frac{s_y}{s_x} = \frac{\frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}}{\frac{1}{n}\sum x_i^2 - \bar{x}^2} \end{aligned}$$

Positive Werte von β bedeuten eine positive Korrelation zwischen X und Y. Negative Werte eine negative Korrelation. $\beta \approx 0$ bedeutet, entweder dass X und Y praktisch unkorreliert sind oder dass die Varianz der x-Werte weitaus größer ist als die der y-Werte.

Regressionskoeffizient β

Der Korrelationskoeffizient r_{xy} und der Regressionskoeffizient \ (= die Steigung) β stehen in engem Zusammenhang:

- Bei der einfachen linearen Regression hat $\hat{\beta}$ dasselbe Vorzeichen wie der Korrelationskoefficient r_{xy}!
- $\beta_{XX} = 1$
- $-\infty < \beta < \infty$
- Größere Werte von β bedeuten NICHT einen stärkeren linearen Zusammenhang (Korrelation).
- $\beta_{XY} \neq \beta_{YX}$ (im Allgemeinen)

- r_{xy} misst nur lineare Abhängigkeit.
- $\beta \neq 0$ bedeutet nicht automatisch einen kausalen Zusammenhang. (Babies und Störche!)

Annahmen und Voraussetzungen für Regression

Die Schätzung kann nur funktionieren, wenn die Daten bestimmte Annahmen erfüllen. Diese Annahmen betreffen insbesonders den Modellfehler ε

(A1) Das Modell hat keinen systematischen Fehler.

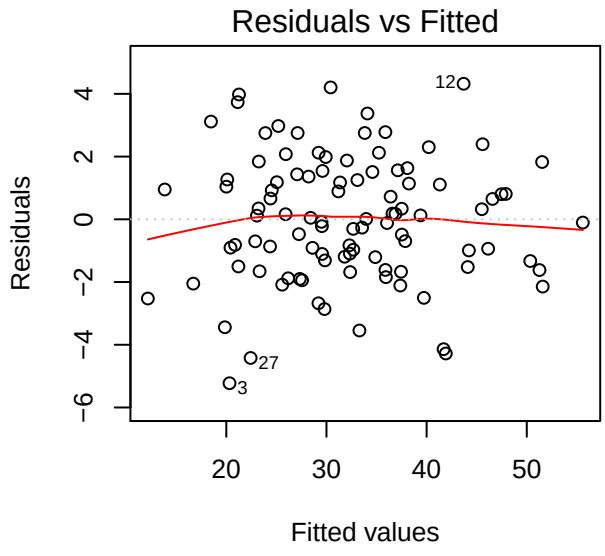
(A2) Die Fehlervarianz ist für alle Beobachtungen gleich groß (homoskedastisch).

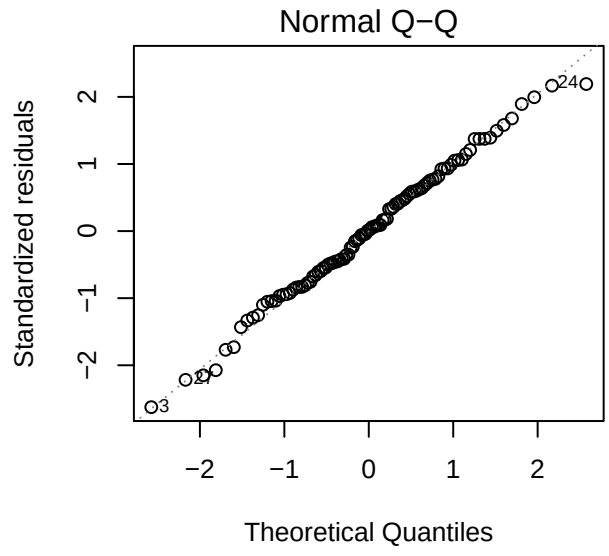
(A3) Die Komponenten des Fehlerterms sind nicht korreliert.

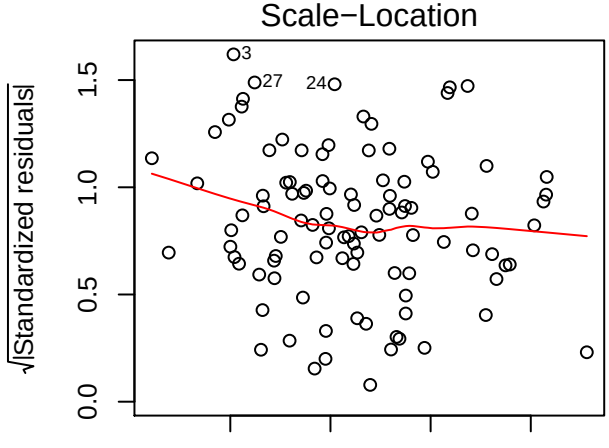
(A4) Der Modellfehler sei normalverteilt.

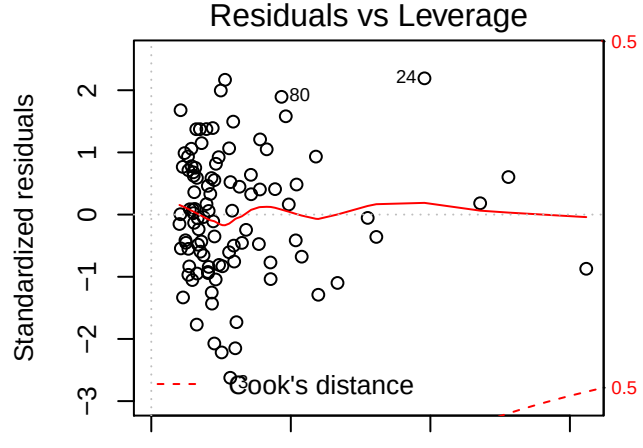
Überprüfung der Annahmen A1 - A4

```
linearesModell <- lm(y ~ x, data=Daten)
plot(linearesModell)
```

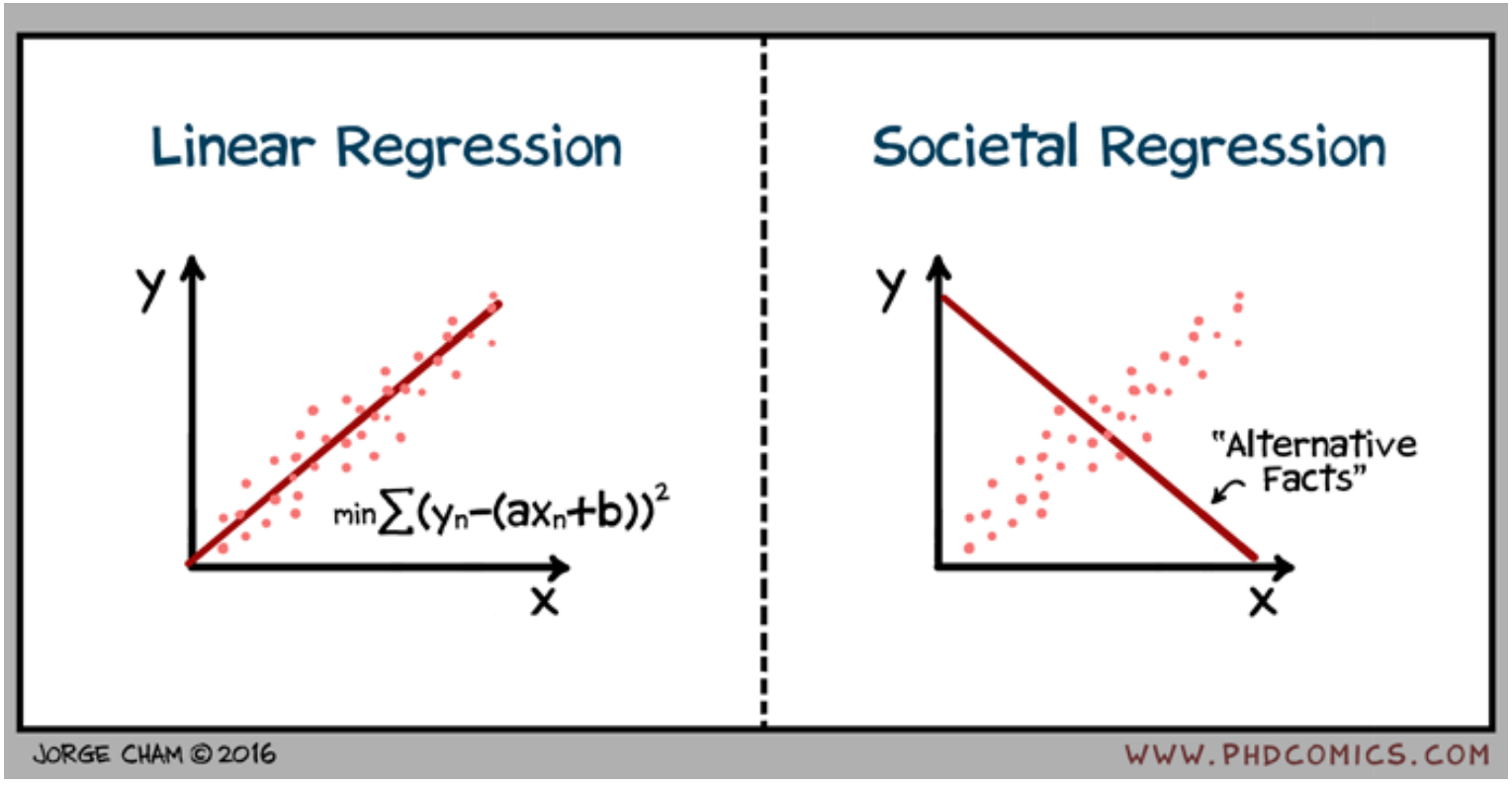








Leverage Points



Multiple lineare Regression

$$y_i = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_k x_{k,i} + \varepsilon_i$$

in der Notation mittels Matrizen und Vektoren

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$\boldsymbol{y}$	...	n-dimensionaler Vektor der Beobachtungen
$\boldsymbol{X}$	...	n $\times$ p Matrix der p Beobachtungsspalten
$\boldsymbol{\beta}$	...	p-dimensionaler Vektor der Regressionskoeffizienten der p verschiedenen Variablen
$\boldsymbol{\varepsilon}$	...	n-dimensionaler Vektor der Fehler (Residuen)

Annahmen und Voraussetzungen für multiple Regression

Die Schätzung kann nur funktionieren, wenn die Daten bestimmte Annahmen erfüllen. Diese Annahmen betreffen insbesonders den Modellfehler ε, aber auch die Regressoren X_j.

(A1) Das Modell hat keinen systematischen Fehler.

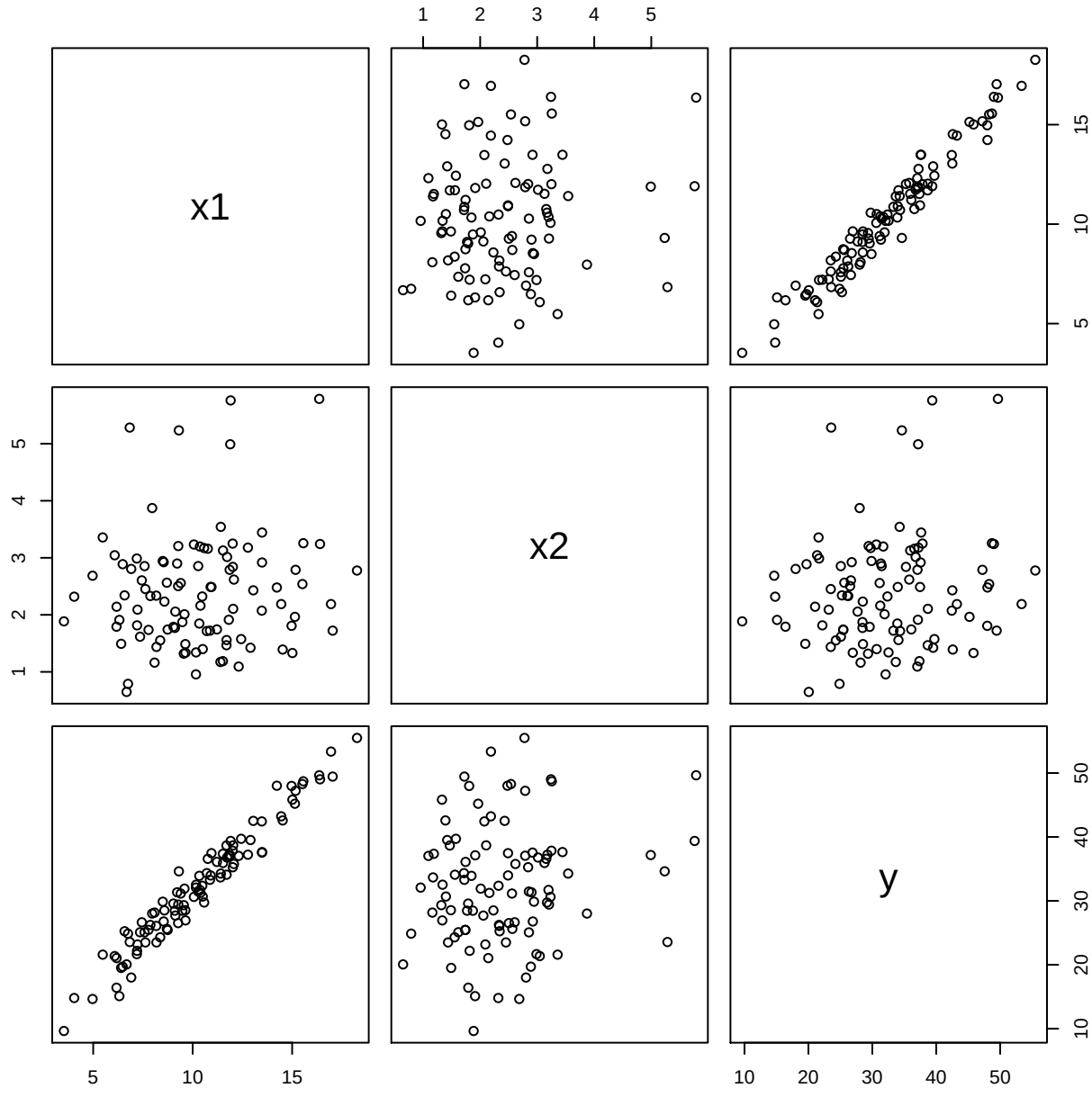
(A2) Die Fehlervarianz ist für alle Beobachtungen gleich groß (homoskedastisch).

(A3) Die Komponenten des Fehlerterms sind nicht korreliert.

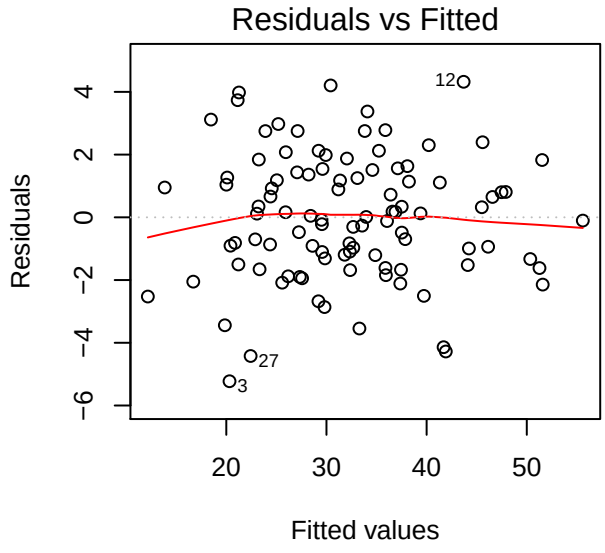
(A4) Der Modellfehler sei normalverteilt.

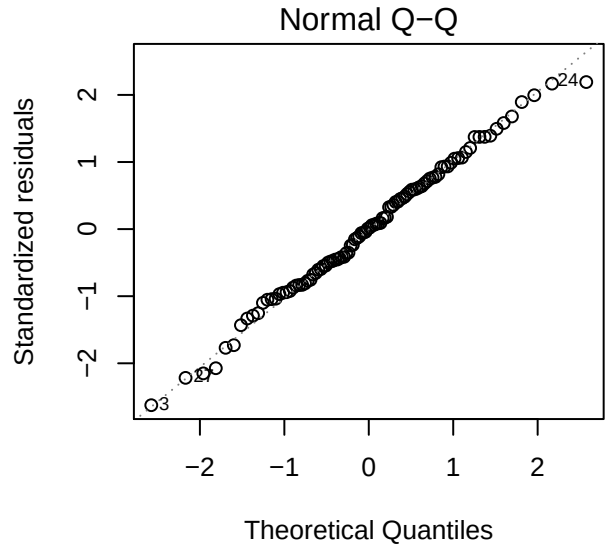
(A5) Es gibt keine lin. Abhängigkeiten zwischen den Regressoren.

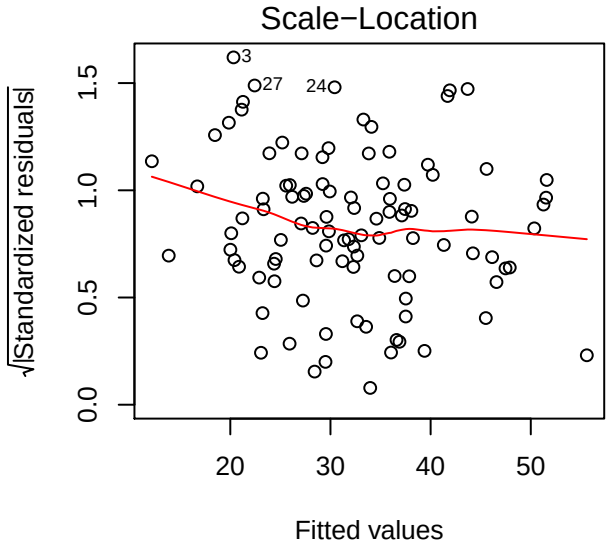
Überprüfung der Annahme A5

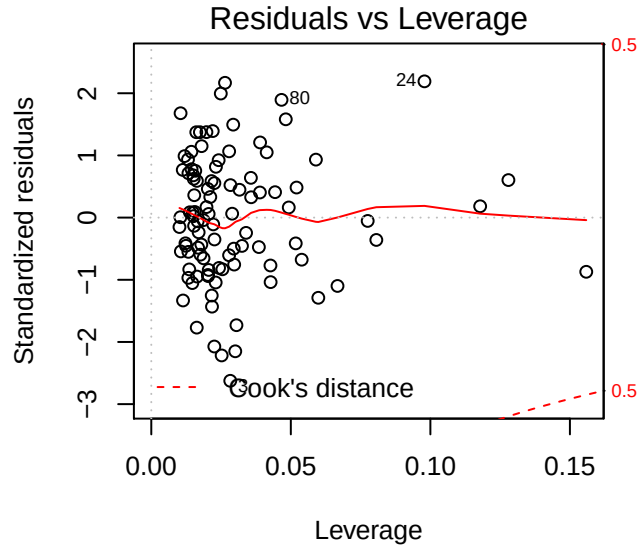


Überprüfung der Annahmen A1 - A4









Modellselektion

Wenn mehrere Variablen als erklärend erwogen werden, müssen noch lange nicht alle erklärend sein. Man kann daher auf unterschiedliche Weise selektieren, welche relevant sind.

Es gilt: "so klein wie möglich, so groß wie notwendig soll ein Model sein"

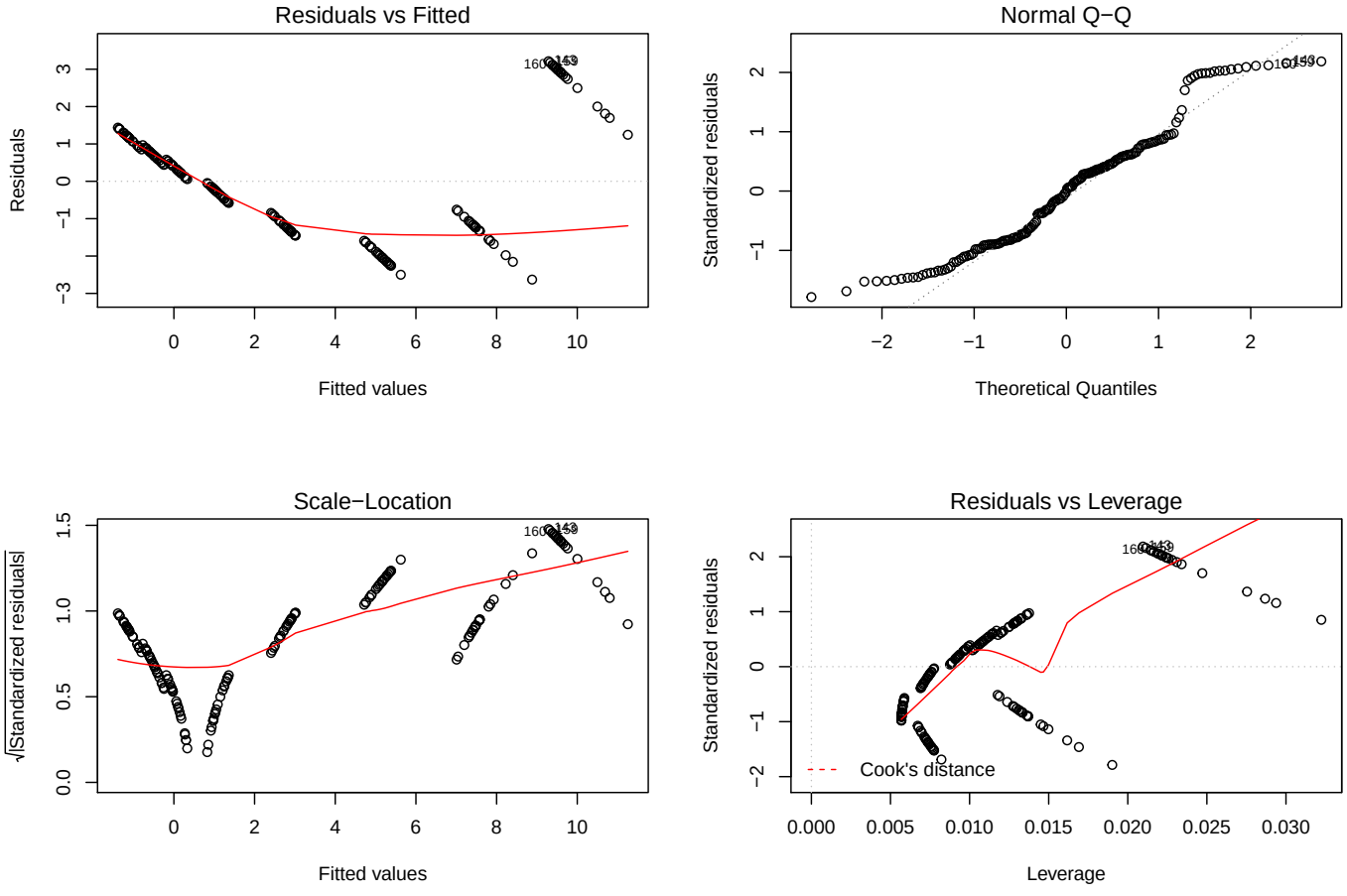
- **t-Tests** für die Koeffizienten β_i
  einfacher und direkter Weg, auf linearen Zusammenhang zu testen. (inkludiert in summary(linearesModell))
- allgemeine schrittweise Modellselection mittels **"goodness-of-fit"** Maßen
  nutzt Maßzahlen wie **Akaike Information Criterion (AIC)** oder **Bayesian Information Criterion (BIC)** zum Vergleich von Modellen und wählt das am besten Passende aus.

Regressionsmodell - Beispiel

```r
linearesModell<-lm(conc~density,data = DNase)
summary(linearesModell)
```

```
##
## Call:
## lm(formula = conc ~ density, data = DNase)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6307 -1.2371  0.0060  0.9051  3.2132
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.4572     0.1760  -8.281 3.13e-14 ***
## density       6.3462     0.1887  33.635  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.487 on 174 degrees of freedom
## Multiple R-squared:  0.8667, Adjusted R-squared:  0.8659
## F-statistic:  1131 on 1 and 174 DF,  p-value: < 2.2e-16
```

Regressionsmodell - Beispiel



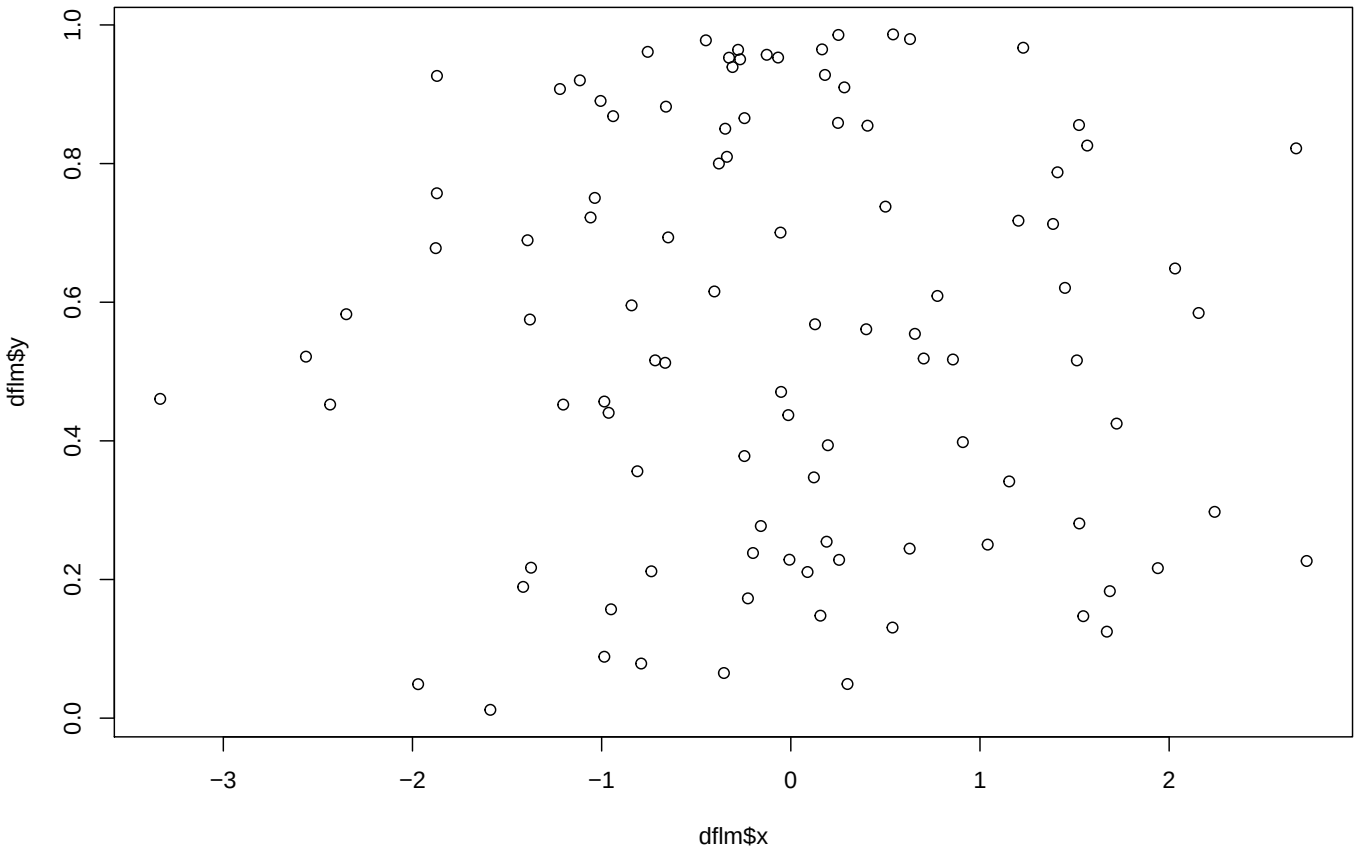
Probleme mit Modell

- Heteroskedastische Fehler - Schwankungen wird größer (A2 nicht erfüllt)
- "Batch-Effekt" Punkt zu Gruppen zusammenclustern (A3 nicht erfüllt)
- 2. Modus in den Fehlern - großen Fehlerwerten (A4 nicht erfüllt)
- keine Punkt ist Hebelpunkt

Modell passt gut (R-squared $= 0.8667$), aber nicht gültig, weil Voraussetzungen nicht erfüllt sind!

Beispiel 2

```
plot(dflm$x,dflm$y)
```

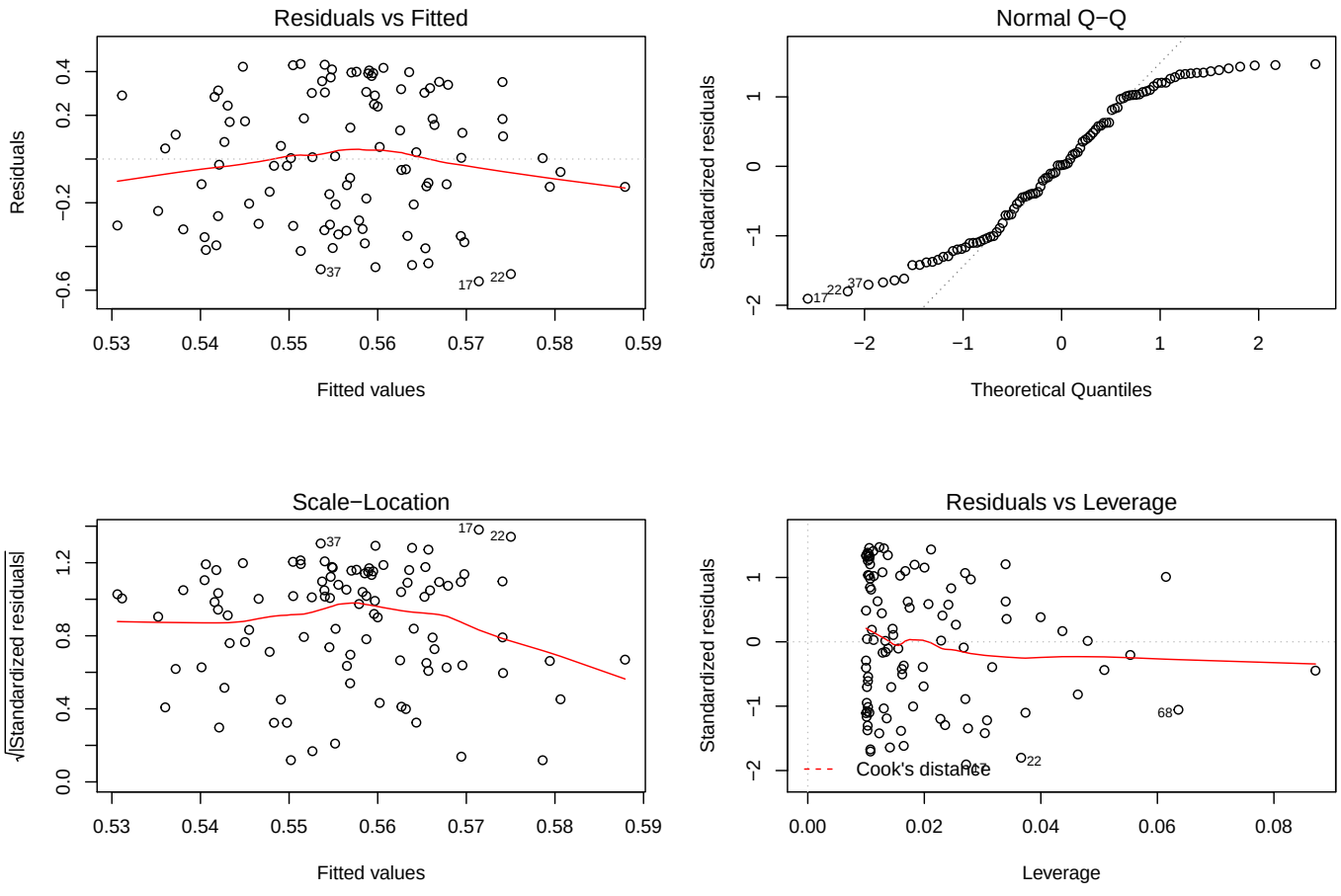


Regressionsmodell - Beispiel 2

```r
linearesModell2<-lm(y~x,data = dflm)
summary(linearesModell2)
```

```
##
## Call:
## lm(formula = y ~ x, data = dflm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5594 -0.2846  0.0049  0.2936  0.4351
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.556402   0.029762  18.695   <2e-16 ***
## x           -0.009452   0.025000  -0.378    0.706
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2975 on 98 degrees of freedom
## Multiple R-squared:  0.001457,   Adjusted R-squared:  -0.008733
## F-statistic: 0.143 on 1 and 98 DF,  p-value: 0.7062
```

Regressionsmodell - Beispiel 2



Zusammenfassung

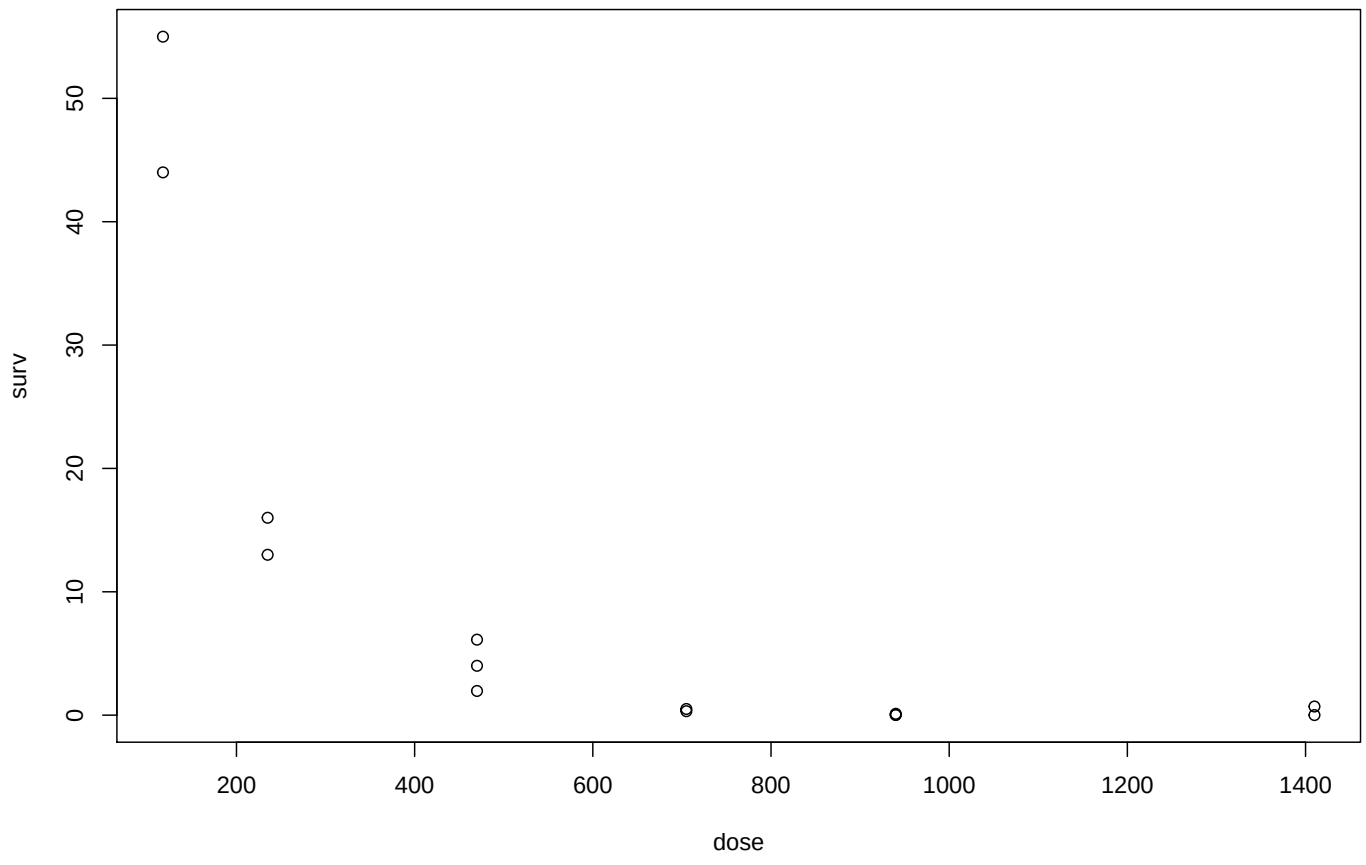
Steigung irrelevant

Güte des Modells R-squared=0.02, Modell erklärt nur 2% der Varianz der Daten.

Aber notwendigen Voraussetzungen A1-A3 sind erfüllt. Modell ist gültig.

Beispiel 3 mit Datentransformation

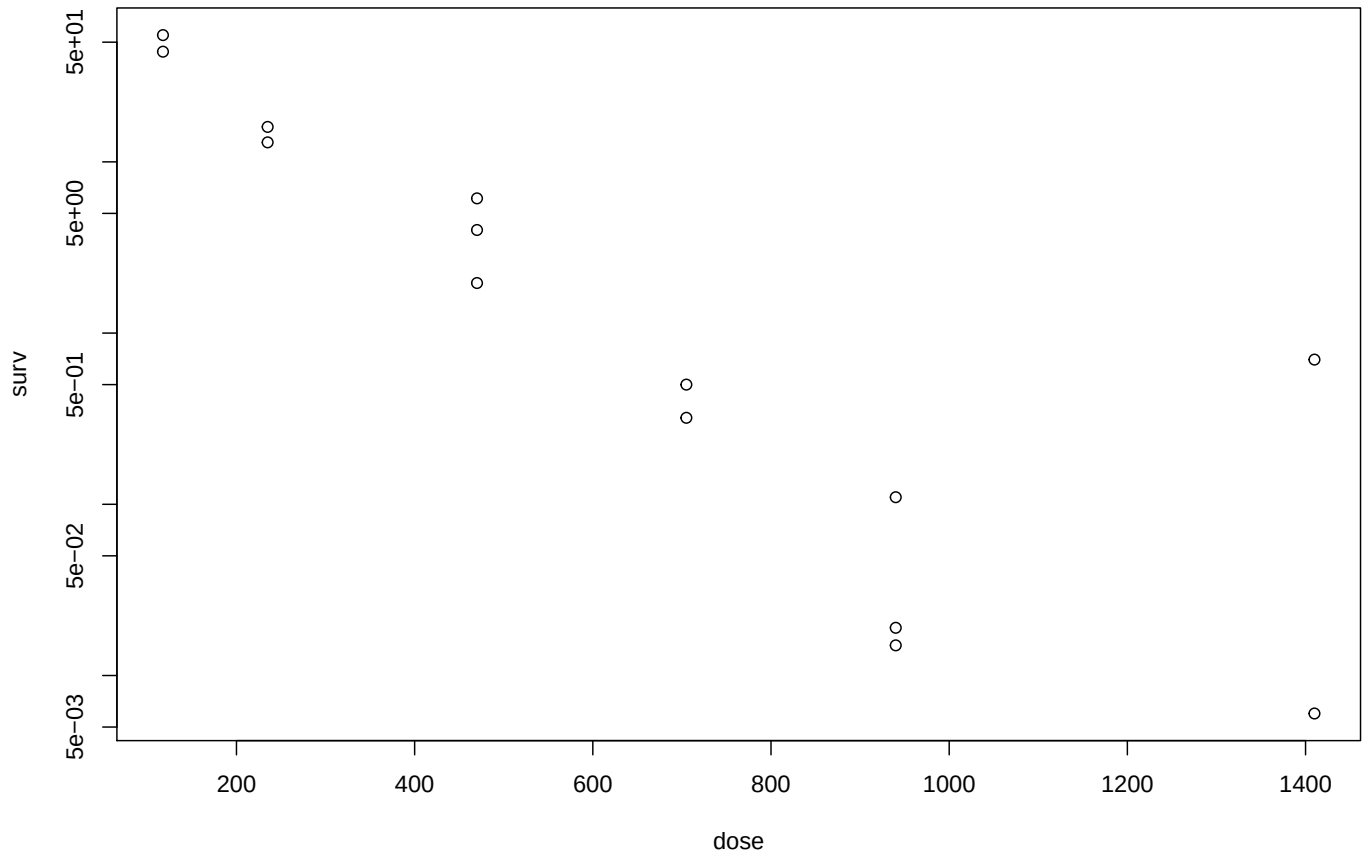
```
library(boot); data(survival)
with(survival, plot(dose, surv))
```



Beispiel 3 Log-transformation

Wir werden hier die y-Werte log-transformieren.

```
with(survival, plot(dose, surv, log="y"))
```



Beispiel 3 Linear Model Summary

```r
fit <- lm(log(surv)~dose, data=survival)
summary(fit)
```

```
##
## Call:
## lm(formula = log(surv) ~ dose, data = survival)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4637 -0.5679 -0.1079  0.5772  4.1592
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.823648   0.811788   4.710 0.000505 ***
## dose        -0.005915   0.001047  -5.651 0.000107 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.629 on 12 degrees of freedom
## Multiple R-squared:  0.7269, Adjusted R-squared:  0.7041
## F-statistic: 31.93 on 1 and 12 DF,  p-value: 0.0001071
```

Modelle

Lineare Modell für Logarithmen

$$log(surv) = \underbrace{3.82}_{\text{mittlere logarithmierte Überlebensdauer surv ist, wenn Dosis=0 ist}} - 0.006*$$

Wenn die Dosis um 1 Einheit steigt, fällt logarithmierte Überlebensdauer um 0.006 Einheiten.

Exponentielles Modell für Originaldaten

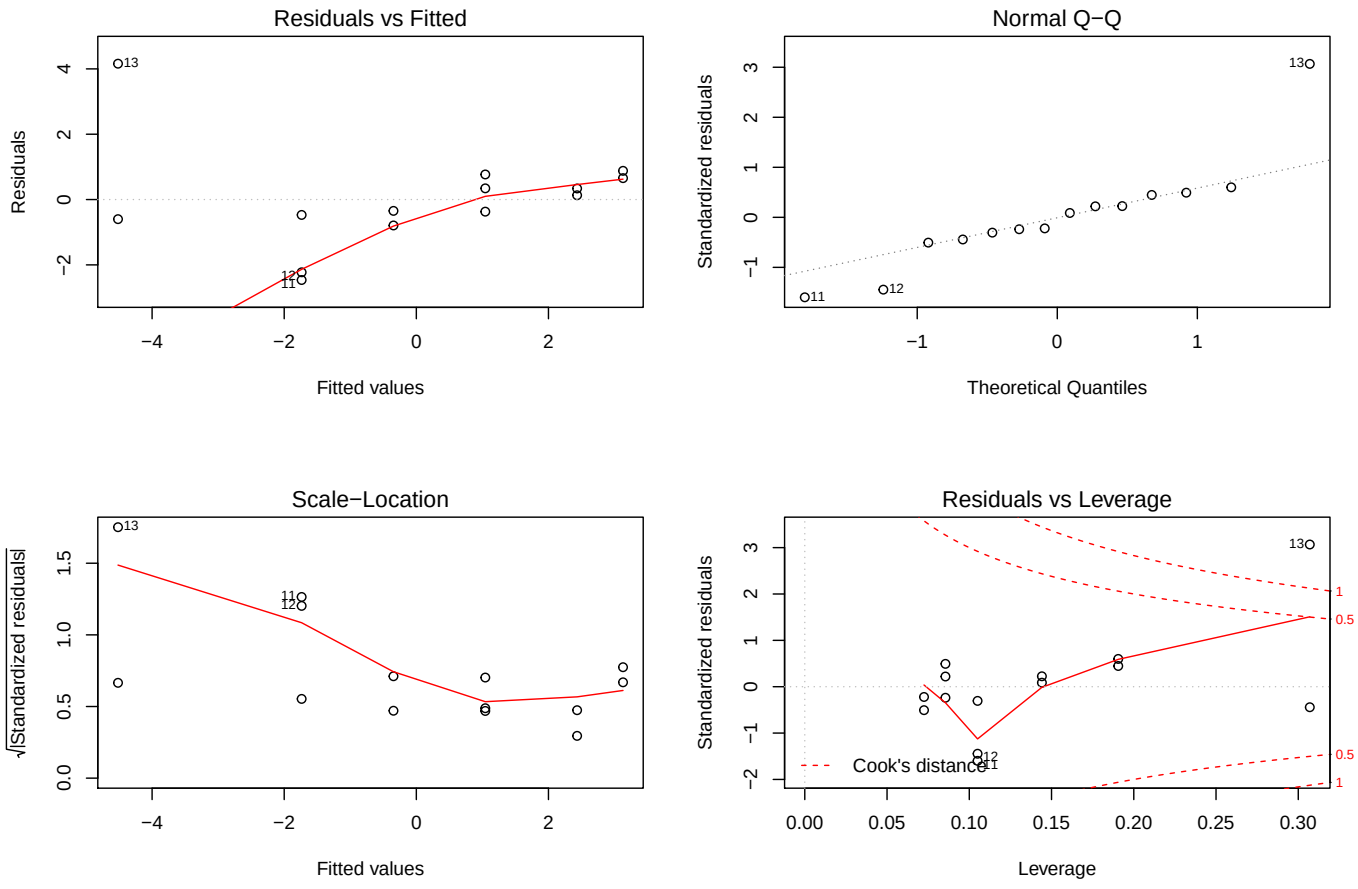
$$surv = exp(3.82 - 0.006*dose) = exp(3.82)*exp(\underbrace{-0.006}_{Wachstumskonstante}*dose)$$

$$surv = exp(3.82) * exp(-0.006)^{dose} = 45.60 * \underbrace{0.994018}_{Wachstumsfaktor}^{dose}$$

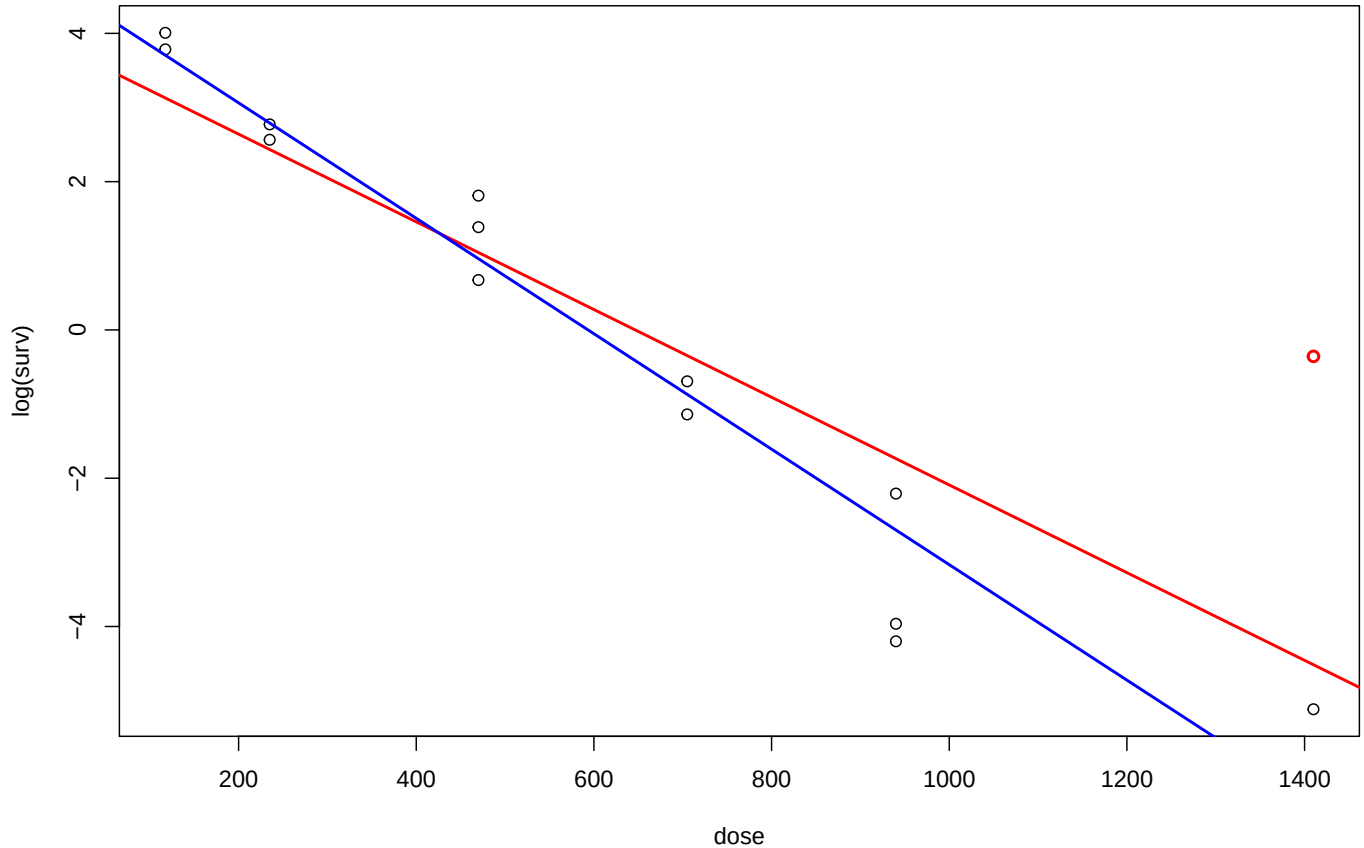
Überlebensdauer von 45.6 Einheiten, wenn Dosis = 0.

Example 3 Residual Diagnostic Plots

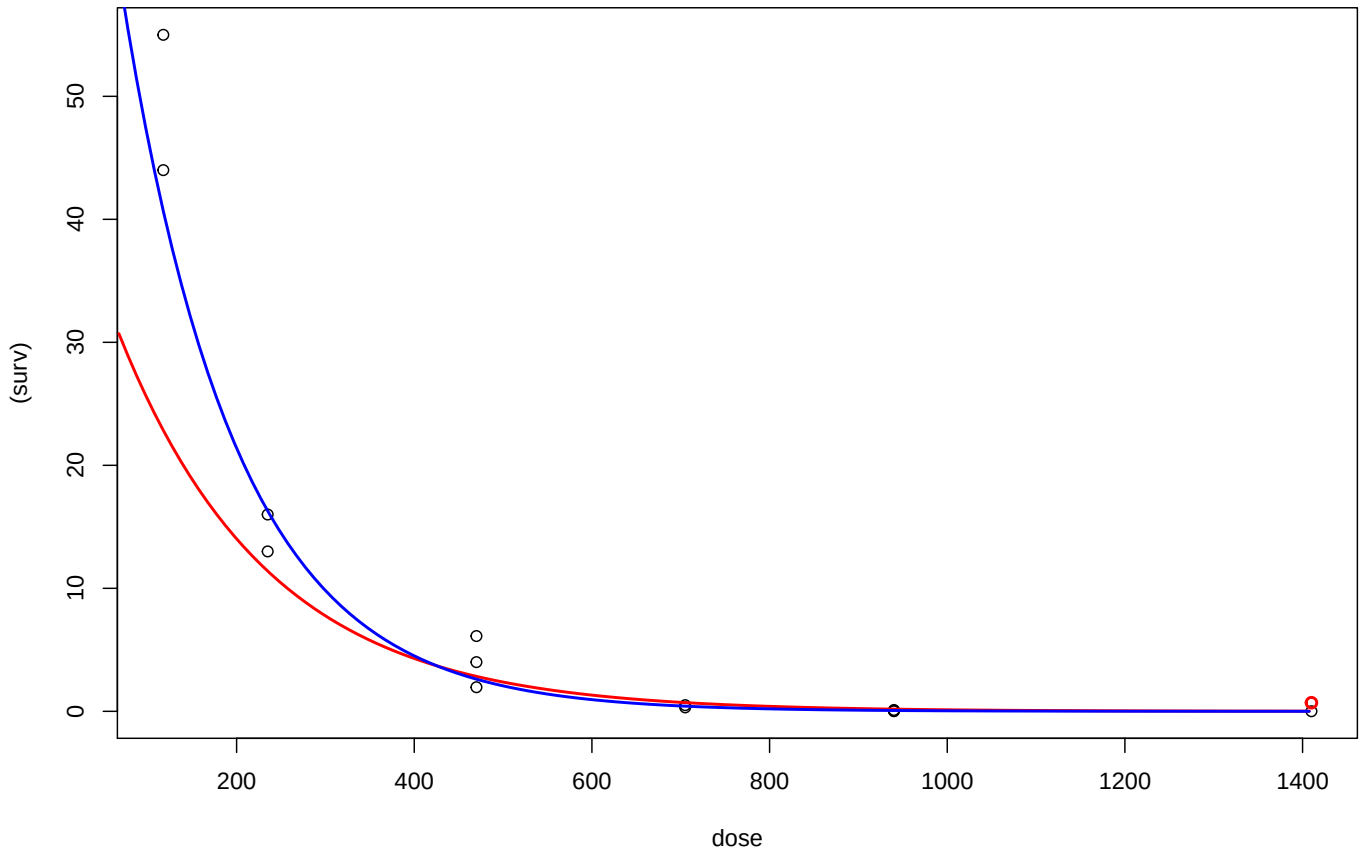
```
par(mfrow=c(2,2)); plot(fit); par(mfrow=c(1,1))
```



Example 3 - Hebelpunkt (Leverage Point)



Example 3 - Hebelpunkt (Leverage Point)

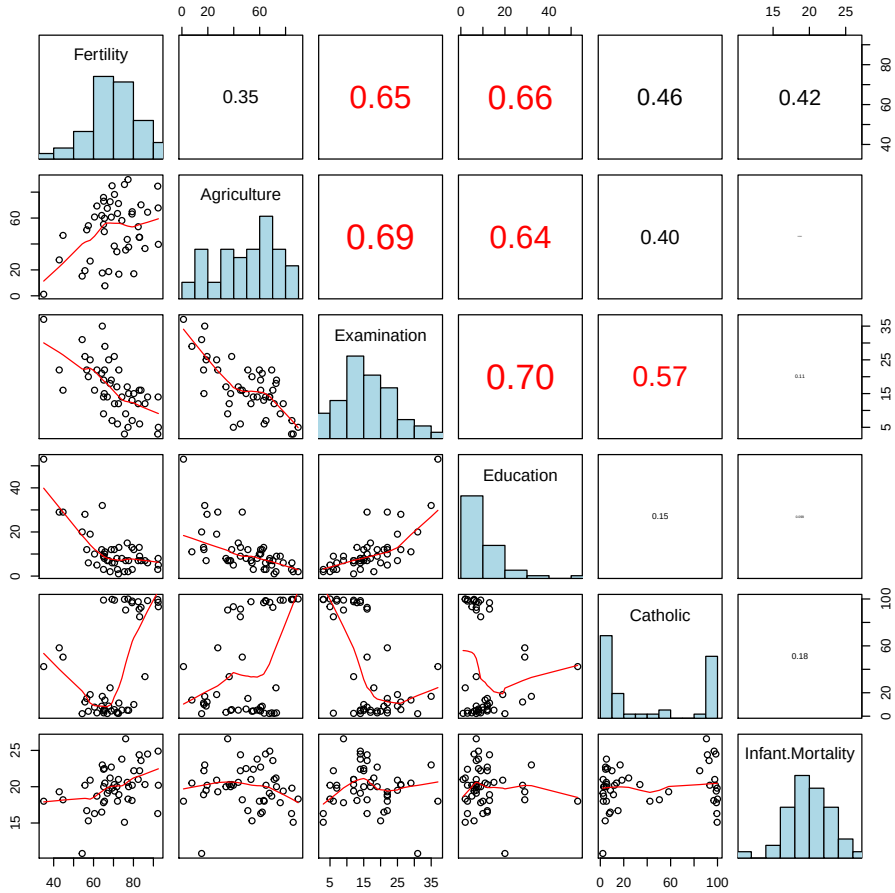


Example 3 - Hebelpunkt (Leverage Point)

```
library(boot); data(survival)
with(survival, plot(dose, (surv)))
dosenew<-seq(min(survival$dose)-50, max(survival$dose),by=1
lines(dosenew,exp(predict.lm(lm(log(surv)~dose,data=surviva
lines(dosenew,exp(predict.lm(lm(log(surv)~dose,data=surviva
points(survival[13,"dose"],(survival[13,"surv"]),col="red",
```

Hebelpunkt muss raus!

Multiples Regressionsmodell - Beispiel

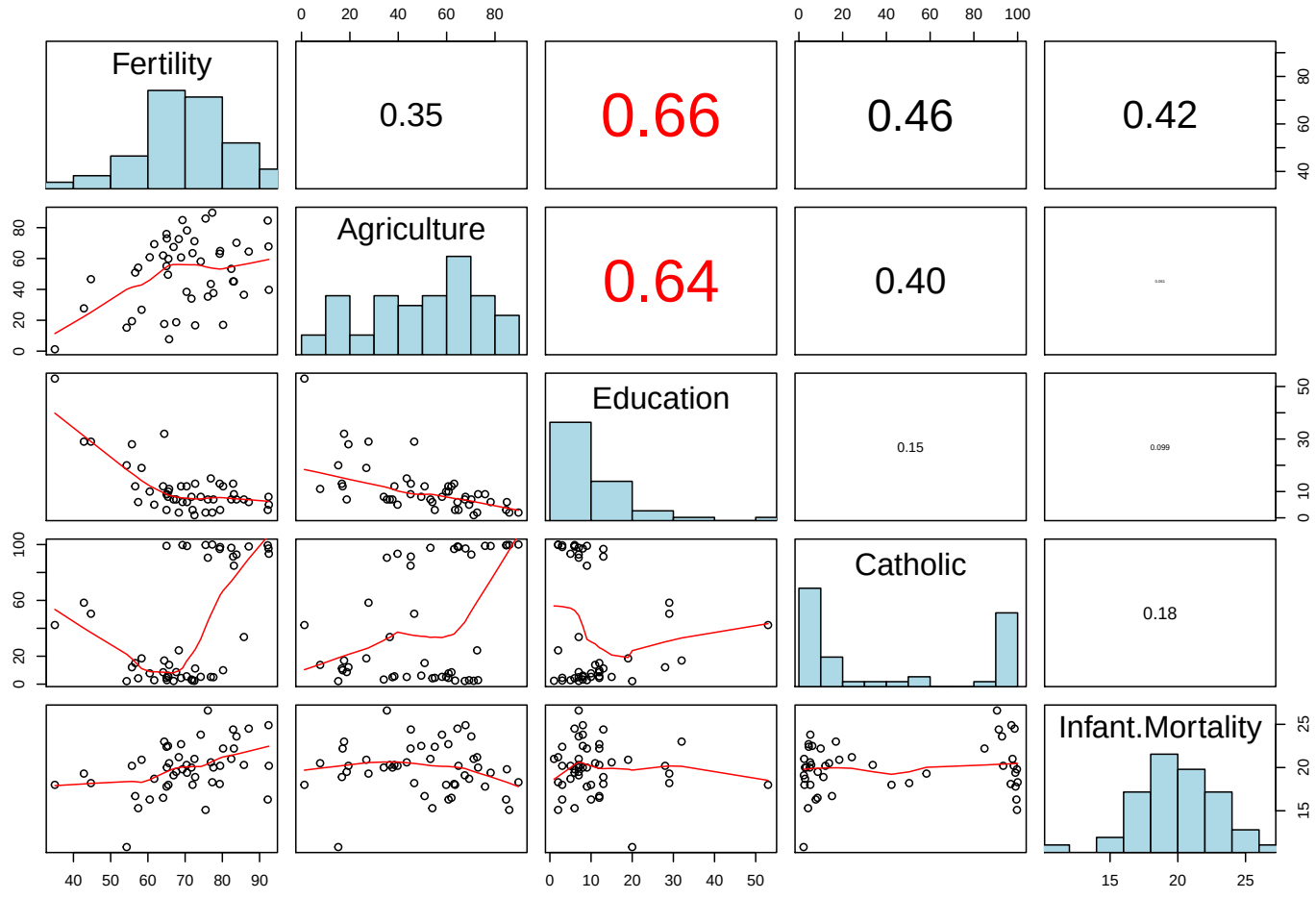


Multiples Regressionsmodell - Beispiel

```
panel.hist <- function(x, ...)
{
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(usr[1:2], 0, 1.5) )
    h <- hist(x, plot = FALSE)
    breaks <- h$breaks; nB <- length(breaks)
    y <- h$counts; y <- y/max(y)
    rect(breaks[-nB], 0, breaks[-1], y, col = "lightblue", ...)
}
panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...)
{
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- abs(cor(x, y))
    txt <- format(c(r, 0.123456789), digits = digits)[1]
    txt <- paste0(prefix, txt)
    if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
    text(0.5, 0.5, txt, cex = cex.cor * r,col=ifelse(r>=0.5,"red","black"))
}
pairs(swiss, lower.panel = panel.smooth,diag.panel = panel.hist, upper.panel = panel.cor)
```

Examiniation Filtern

```
pairs(swiss[,-3], lower.panel = panel.smooth,diag.panel = p
```



Multiples Regressionsmodell - Modellanpassung

```
##
## Call:
## lm(formula = Education ~ . - Examination, data = swiss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.4029  -2.7803  -0.7571   2.4934  12.8590
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      49.99303    6.18641   8.081 4.31e-10 ***
## Fertility        -0.52070    0.07869  -6.617 5.14e-08 ***
## Agriculture      -0.22880    0.03906  -5.857 6.37e-07 ***
## Catholic          0.08333    0.02179   3.825 0.000428 ***
## Infant.Mortality  0.28437    0.30040   0.947 0.349243
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.224 on 42 degrees of freedom
## Multiple R-squared:  0.7305, Adjusted R-squared:  0.7048
## F-statistic: 28.46 on 4 and 42 DF,  p-value: 1.804e-11
```

Multiples Regressionsmodell - Modellanpassung

```
##
## Call:
## lm(formula = Education ~ . - Examination - Infant.Mortality,
##     data = swiss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.0852  -2.9521  -0.6678   3.2519  12.9706
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 53.85051    4.64907  11.583 8.30e-15 ***
## Fertility   -0.48883    0.07104  -6.881 1.91e-08 ***
## Agriculture -0.23799    0.03779  -6.298 1.35e-07 ***
## Catholic     0.08440    0.02173   3.884  0.00035 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.218 on 43 degrees of freedom
## Multiple R-squared:  0.7247, Adjusted R-squared:  0.7055
## F-statistic: 37.73 on 3 and 43 DF,  p-value: 4.123e-12
```

Multiples Regressionsmodell - Modellgleichung

Variante 1

Education = 49.99 + (-0.52) · Fertility + (-0.23) · Agriculture+ 0.08 · Catholic+ 0.28 · Infant.Mortality

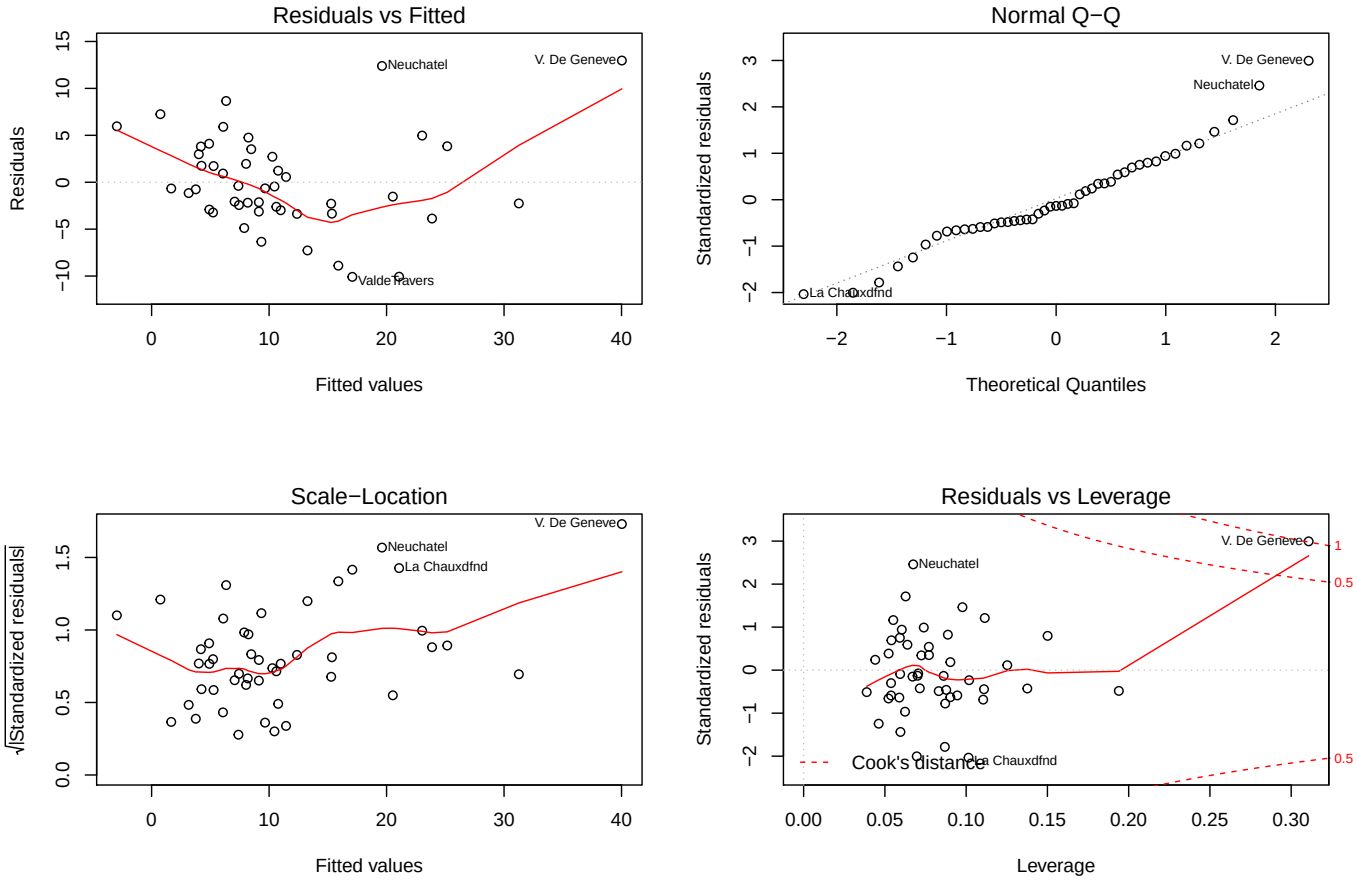
Variante 2 ohne Infant Mortality

Education = 53.85 + (-0.49) · Fertility + (-0.24) · Agriculture+ 0.08 · Catholic

gute Modellanpassung bei R^2 72.47%

Multiples Regressionsmodell - Quality Plots

```
par(mfrow=c(2,2))
plot(multilinearesModell)
```



Hebelpunkt entfernen

```
##
## Call:
## lm(formula = Education ~ . - Examination - Infant.Mortality,
##     data = swiss[-45, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0144 -2.4407 -0.7688  2.6409 14.0202
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 45.26863    4.91660   9.207 1.24e-11 ***
## Fertility   -0.38468    0.07120  -5.403 2.86e-06 ***
## Agriculture -0.20240    0.03566  -5.676 1.16e-06 ***
## Catholic     0.06188    0.02070   2.989  0.00466 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.697 on 42 degrees of freedom
## Multiple R-squared:  0.6216, Adjusted R-squared:  0.5945
## F-statistic: 22.99 on 3 and 42 DF,  p-value: 5.776e-09
```

Multiples Regressionsmodell - Modellgleichung

Variante 1

Education = 49.99 + (-0.52) · Fertility + (-0.23) · Agriculture+ 0.08 · Catholic + 0.28* Infant.Mortality

Variante 2 ohne Infant Mortality

Education = 53.85 + (-0.49) · Fertility + (-0.24) · Agriculture+ 0.08 · Catholic

Variante 3 ohne Infant Mortality und ohne Val de Genenve

Education = 45.27 + (-0.38) · Fertility + (-0.20) · Agriculture+ 0.06 · Catholic

Einfluss des Hebelpunkts!!!

gute Modellanpassung bei R^2 62.16%

Genf ein “guter Hebelpunkt” -> Modellanpassung besser

Einfache Varianzanalyse (ANOVA)

Bisher haben wir im linearen Regressionsmodell

$$y_i = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_k x_{k,i} + \varepsilon_i$$

als Erklärungsvariablen immer nur quant. Variablen x_{ij} verwendet. Aber natürlich können auch qual. Erklärungsvariablen einen Einfluß auf y_i haben.

In der explorativen Datenanalyse haben wir dies Problem auch schon durch parallele Boxplots visualisiert.

Das Modell wird also einfacher, nämlich im einfachsten Fall mit einer erklärenden Variable mit Kategorien j= 1, . . ., K

$$y_{i,j} = \mu + \alpha_j + \epsilon_{i,j}$$

wobei wir hier die verschiedenen Kategorieausprägungen mit dem Index j versehen. α_j misst also die mittlere Abweichung des Mittelwertes der Kategorie j vom Mittelwert aller Daten y.

Einfache Varianzanalyse (ANOVA)

Das Modell für die einfache Varianzanalyse mit einer erklärenden Variable mit Kategorien j= 1, . . ., K

$$Y_{ij} = \underbrace{\mu}_{\text{Gesamtmittelwert}} + \underbrace{\alpha_i}_{\text{Abstand der Gruppenittelwerte von } \mu} + \underbrace{\epsilon_{ij}}_{\text{Fehler=Residuen}}$$

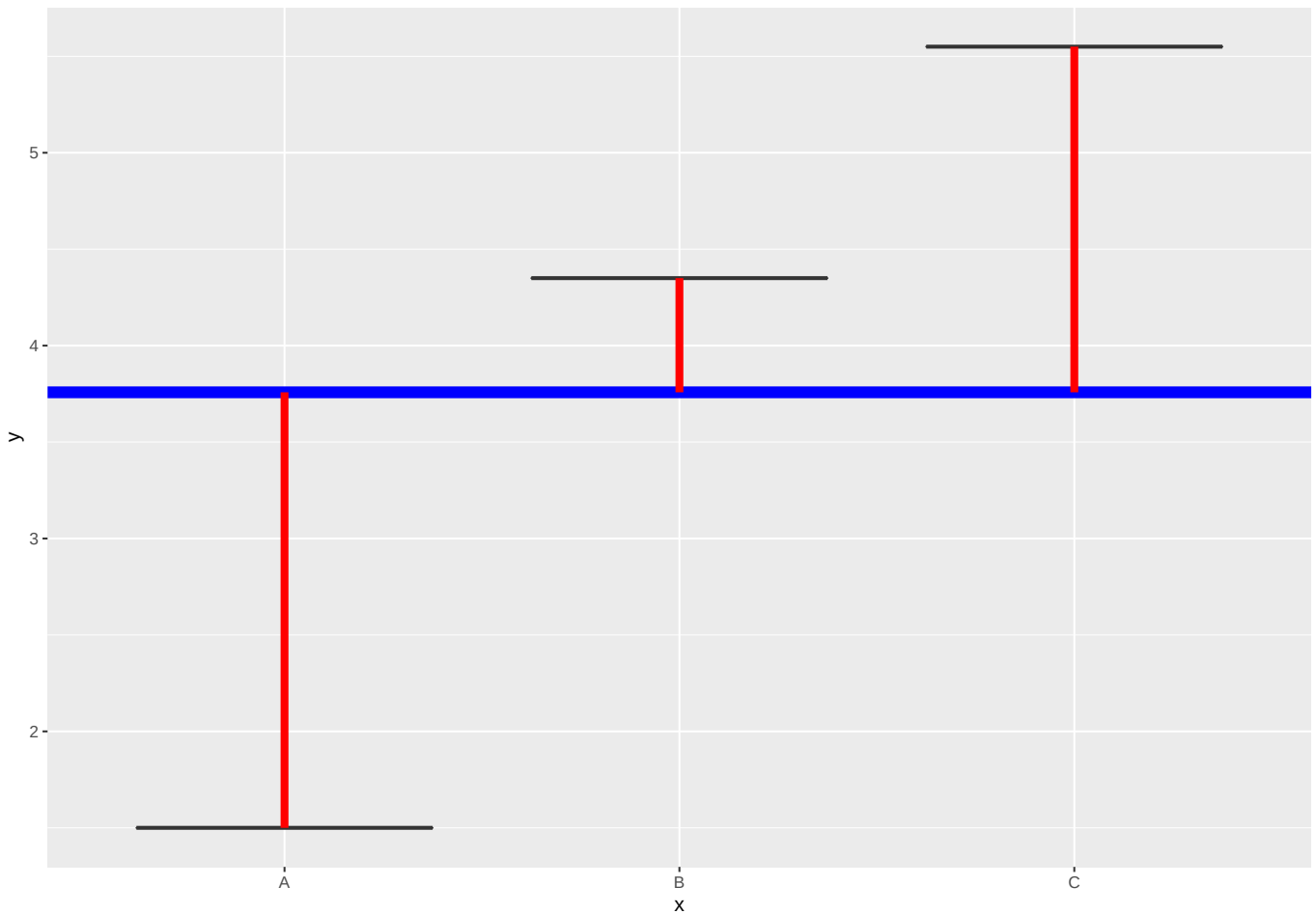
wobei wir hier die verschiedenen Kategorieausprägungen mit dem Index j versehen, bedeutet, dass α_j die mittlere Abweichung des Mittelwertes der Kategorie j vom Mittelwert aller Daten y misst.

ANOVA parameters visualised

$\mu = \bar{y}_{..} =$ Gesamtmittelwert

$\alpha_i =$ Abstand der Gruppenmittelwerte von Gruppe i vom Gesamtmittelwert

$\bar{y}_{i.} =$ Gruppenmittelwert von Gruppe i



ANOVA als Hypothesentest

Diese einfache Varianzanalyse wird auch als Test verwendet, um folgende Hypothesen zu testen:

H_0 : $\mu_1 = \mu_2 = \mu_3 = \cdots$

H_1 : Wenigstens in einer Kategorie ist der Mittelwert unterschiedlich

Das ist die allgemeine Methode, die Unterschiede der mittleren Werte von bekannten Kategorien zu ermitteln.

Die Voraussetzung ist, dass die **Daten**

annähernd symmetrisch,

unimodal und

ohne Ausreißer sind.

ANOVA models differences in means with moderate fit

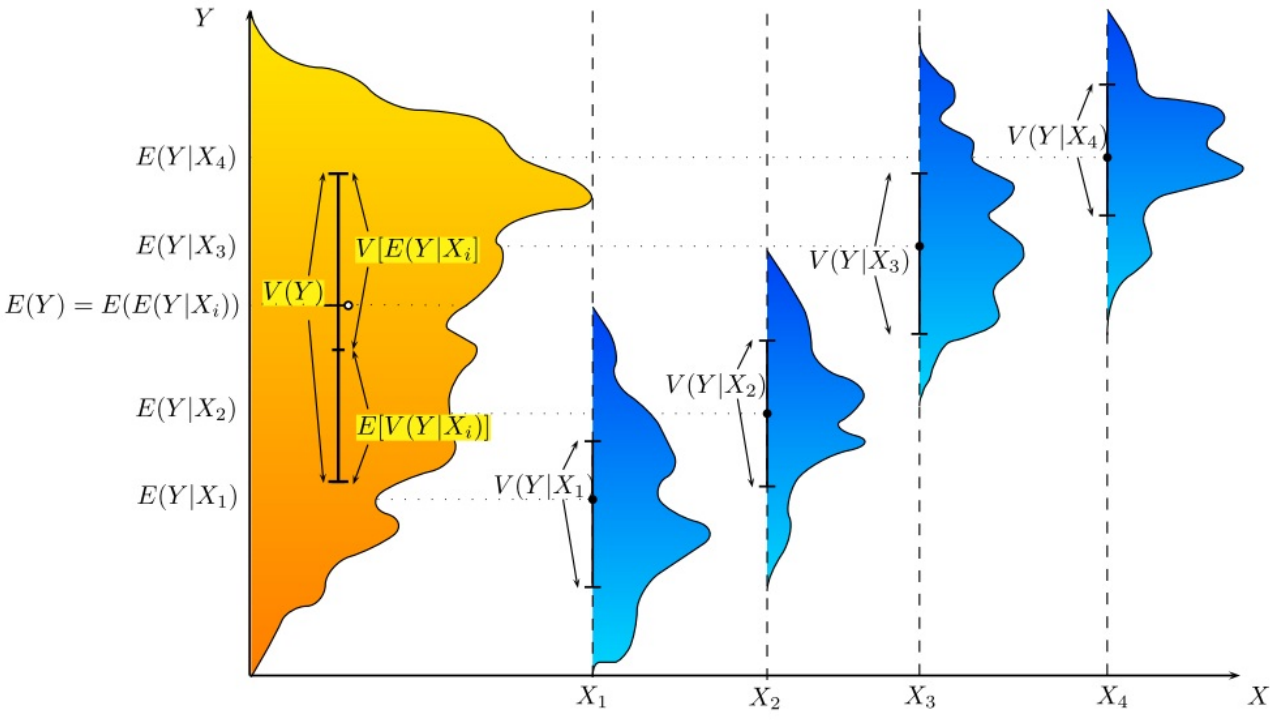


Figure 1: ANOVA : Fair fit

ANOVA models differences in means with bad fit

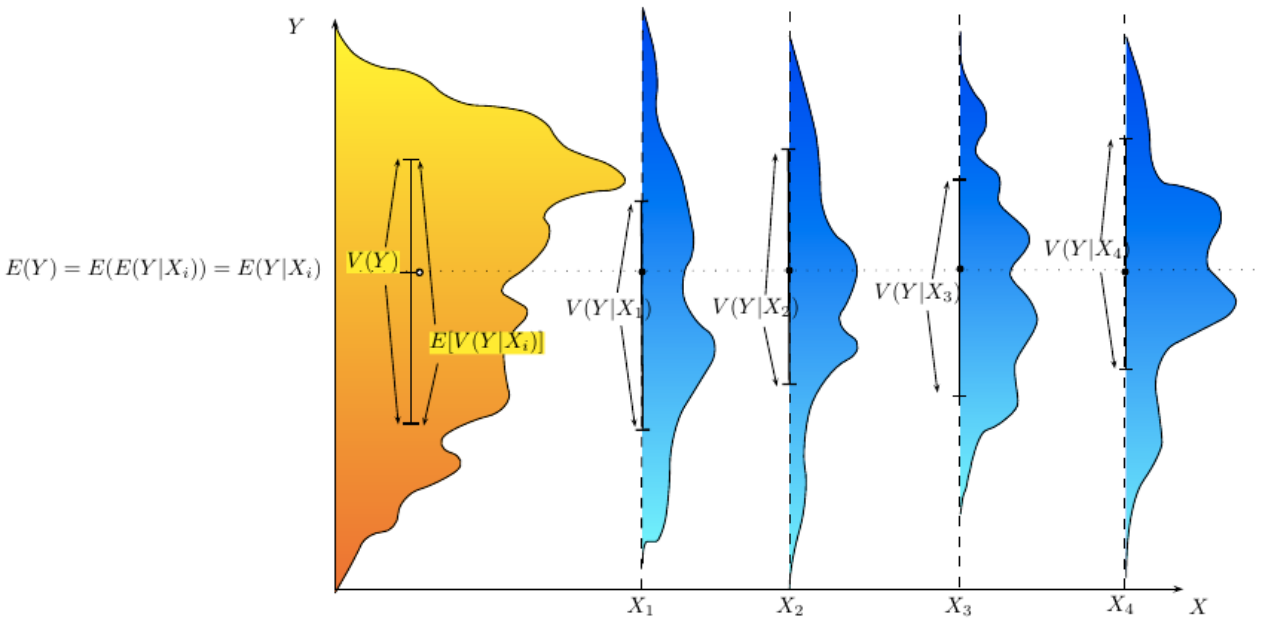


Figure 2: ANOVA : No fit

ANOVA models differences in means with good fit

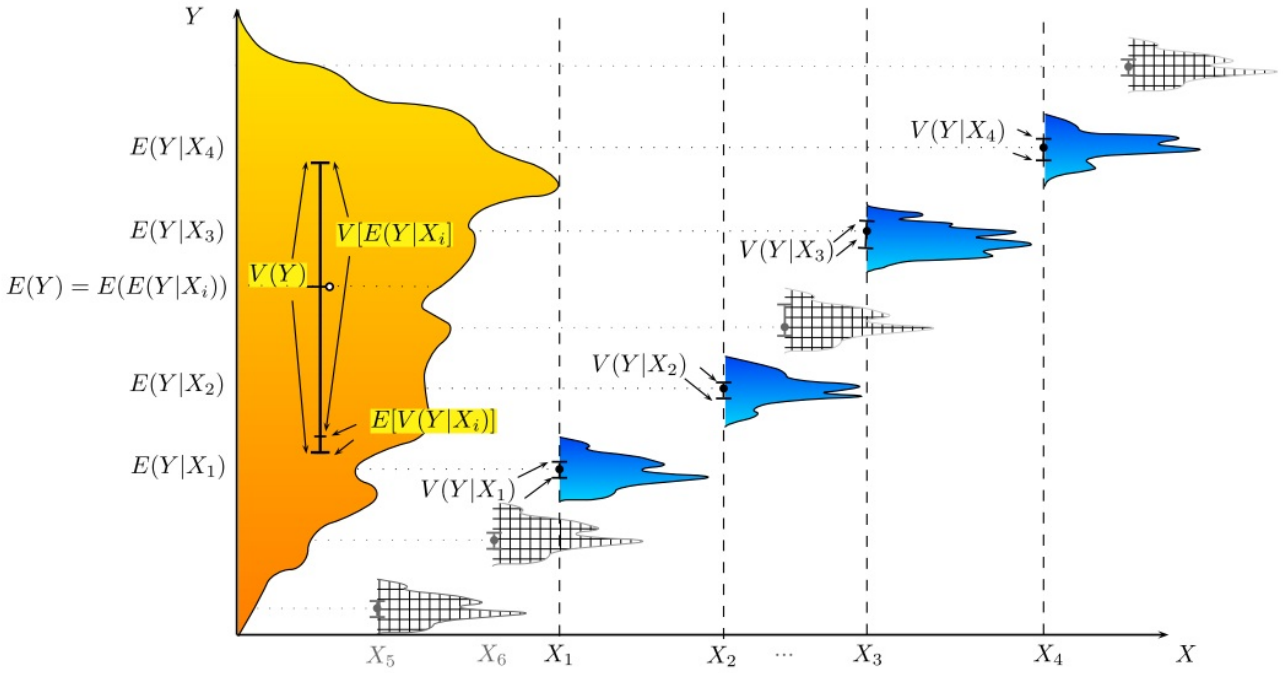


Figure 3: ANOVA : very good fit

Why is ANOVA an analysis of variance?

We have a "sum of squares law":

$$\underbrace{RSS_{total}}_{\text{total residual sum of squares}} = \underbrace{RSS_1}_{\text{within-sample-variance}} + \underbrace{RSS_0}_{\text{between-sample-variance}}$$

where RSS_1 and RSS_0 are uncorrelated.

Zweiweg Varianzanalyse (ANOVA)

Das Modell wird erweitert um eine zweite erklärende kategoriale Variable ("Zweiweg") mit Kategorien, m= 1, ..., M zusätzlich zu den Kategorien der ersten erklärenden Variable k= 1, ..., K

$$y_{i,k,m} = \mu + \alpha_k + \beta_m + \epsilon_{i,k,m}$$

wobei wir hier die verschiedenen Kategorieausprägungen mit dem Index k bzw. m für die 1. und 2. kategoriale Variable versehen.

α_k misst also die mittlere Abweichung des Mittelwertes der Kategorie k vom Mittelwert aller Daten y.

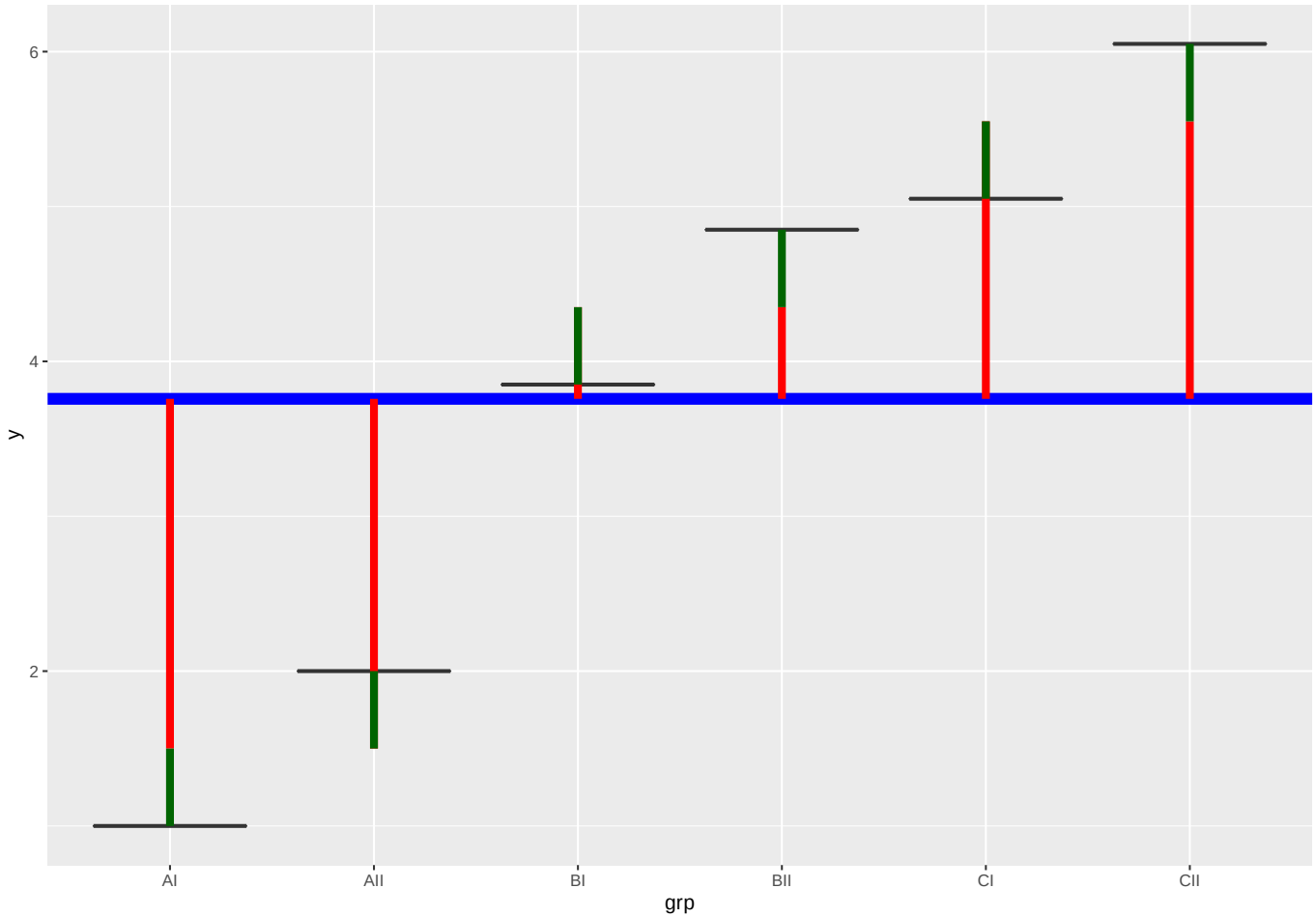
β_m misst also die mittlere Abweichung des Mittelwertes der Kategorie m von Mittelwerten der Unterkategorien $k = 1, \ldots, K$ aller Daten y.

Graphical Explanation of the parameters

$\mu = \bar{y}.. =$ Gesamtmittelwert

$\alpha_i =$ Abstand der Gruppenmittelwerte von x1-Gruppe i vom Gesamtmittelwert

$\beta_j =$ Abstand der Gruppenmittelwerte von x2-Gruppe j vom Gruppenmittelwerte α_i



Zweiweg Varianzanalyse (ANOVA)

Bei zwei Variablen mit Unterkategorien, deren Zählungen wir in Kreuztabellen und mit Mosaicplots dargestellt haben, sind also mehrere Szenarien möglich:

$y \sim 1$ Nur der Mittelwert der gesamten Daten wird als Mittelwert in allen Teilkategorien angenommen.

$y \sim X1$ Nur die erste kategoriale Variable X1 führt zu einer Aufteilung der Mittelwerte, die zweite Variable hat keinen Effekt.

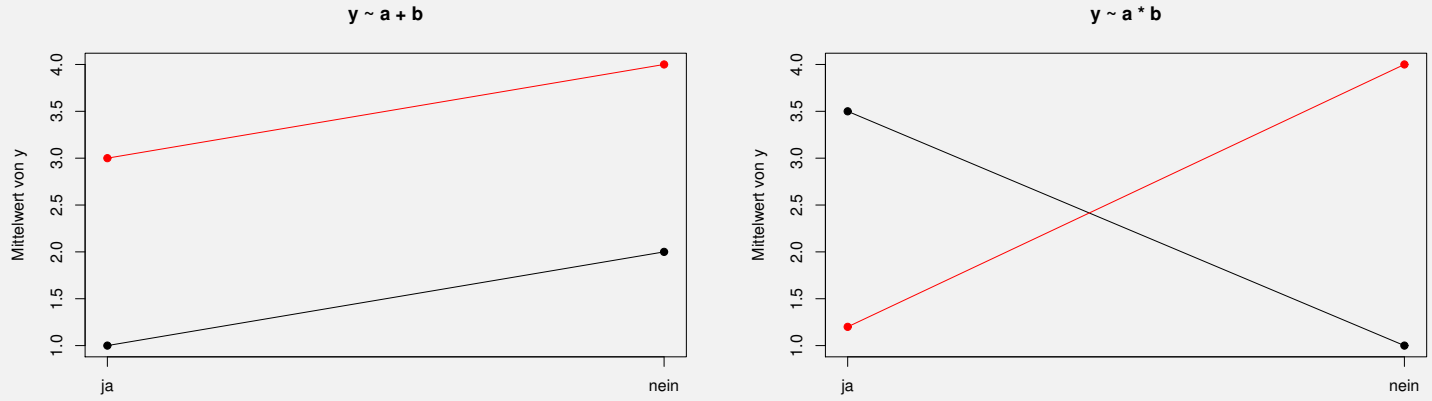
$y \sim X2$ Nur die zweite kategoriale Variable X2 führt zu einer Aufteilung der Mittelwerte, die erste Variable hat keinen Effekt.

$X1 + X2$ Beide kategorialen Variablen X1 und X2 führen zu einer Aufteilung der Mittelwerte in den unterschiedlichen Teilkategorien.

$X1 * X2$ Beide kategorialen Variablen X1 und X2 führen zu einer Aufteilung der Mittelwerte in den unterschiedlichen Teilkategorien und zusätzlich addieren sich die Effekte nicht

Zweiweg Varianzanalyse

Interaction Plot Der **Interaction Plot** visualisiert das Verhalten der Mittelwerte über Kategorien hinweg. Dabei werden die Richtungen von einer Kategorie zur nächsten als Linien durchgezogen. Die Unterkategorien der 2. Variable werde durch unterschiedliche Linientypen visualisiert.



Example for motivating models: Cellular phones

Customer ratings based on certain features of products are basic data in marketing. We use a cell phone example, where one feature is the design as provided by Apple's iPhones or Samsung's products as an alternative. The second variable is the availability of Apps for the products. Hypothetical categories are defined for marketing experiments, trying to estimate consumer's behaviour.

	X_1 (Design)	
X_2 (Apps)	Apple	Samsung
App Store		
Google Archives		

1^{st} case: no effect of factors X_1 and X_2

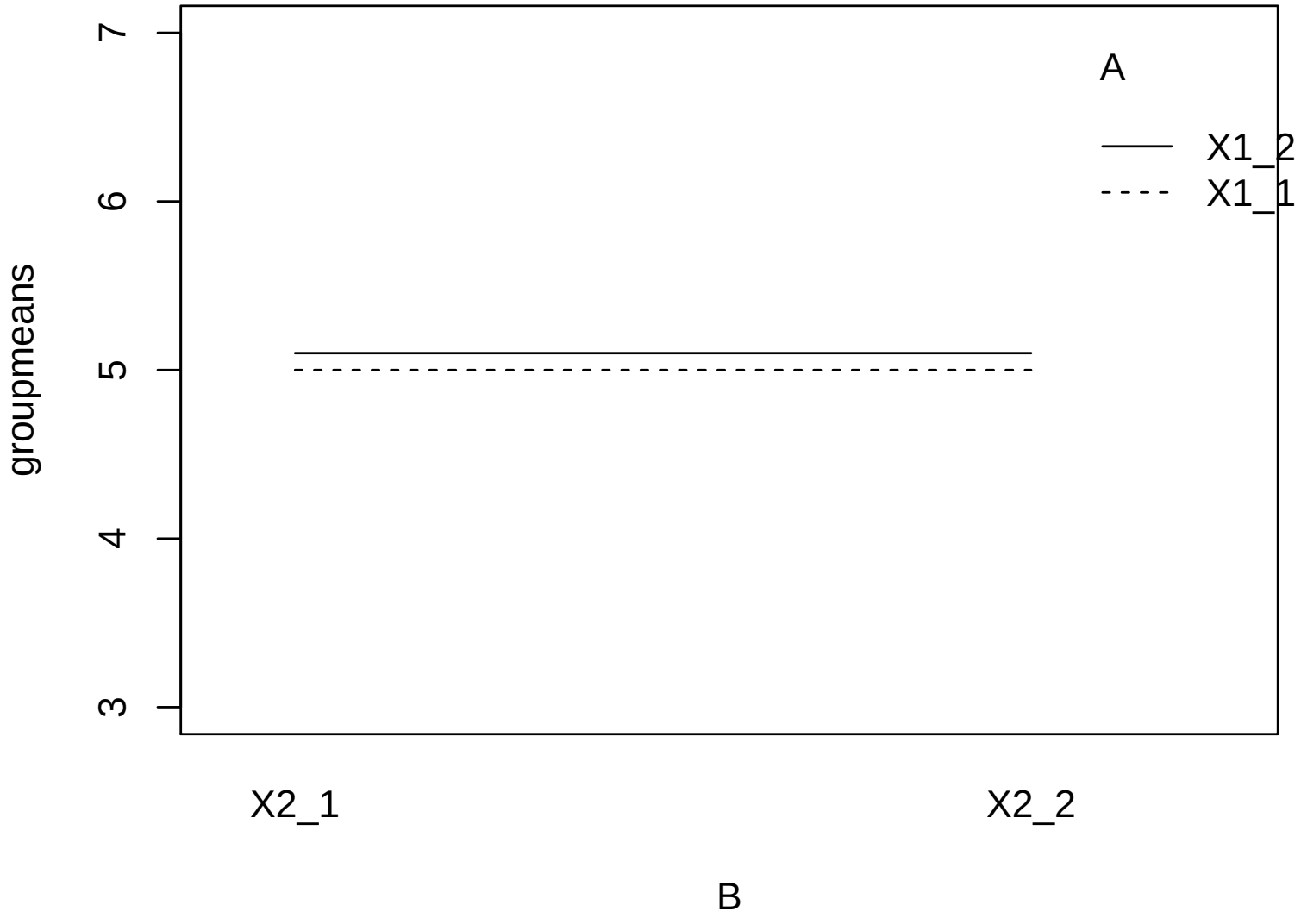
	X_1 (Design)	
X_2 (Apps)	Apple	Samsung
App Store	5	5
Google Archives	5	5

All combinations yield the same outcome.

Thus, neither X_1 nor X_2 are required for the model (**null-model**).

```
lm(Y ~ 1)
```

1^{st} case: no effect of factors X_1 and X_2



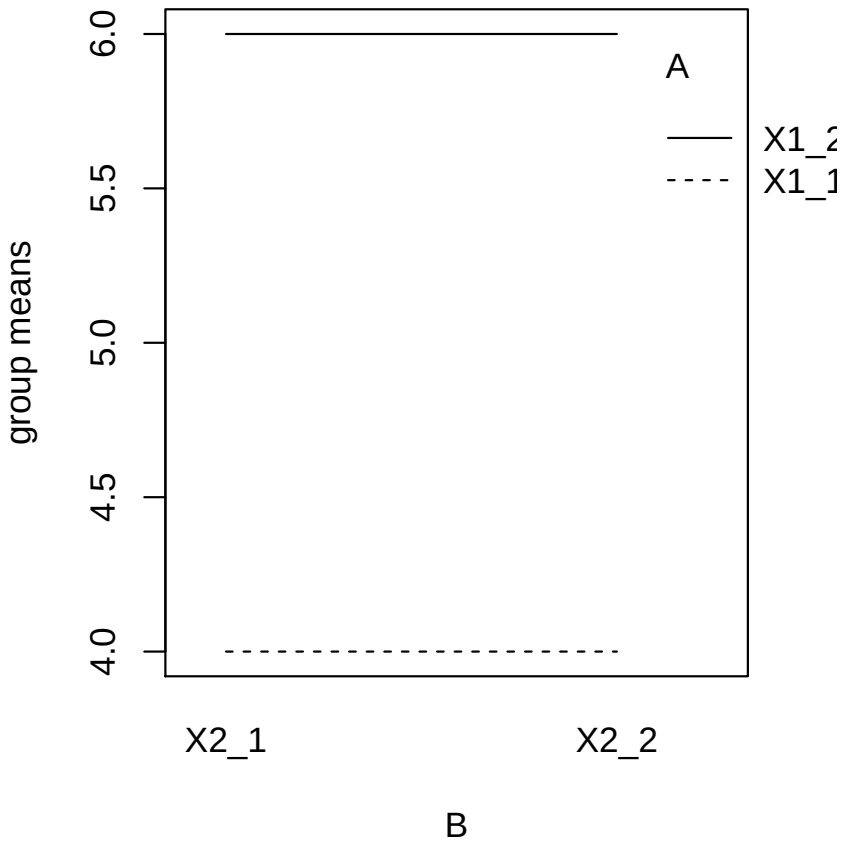
2^{nd} case: effect of X_1 or X_2 only

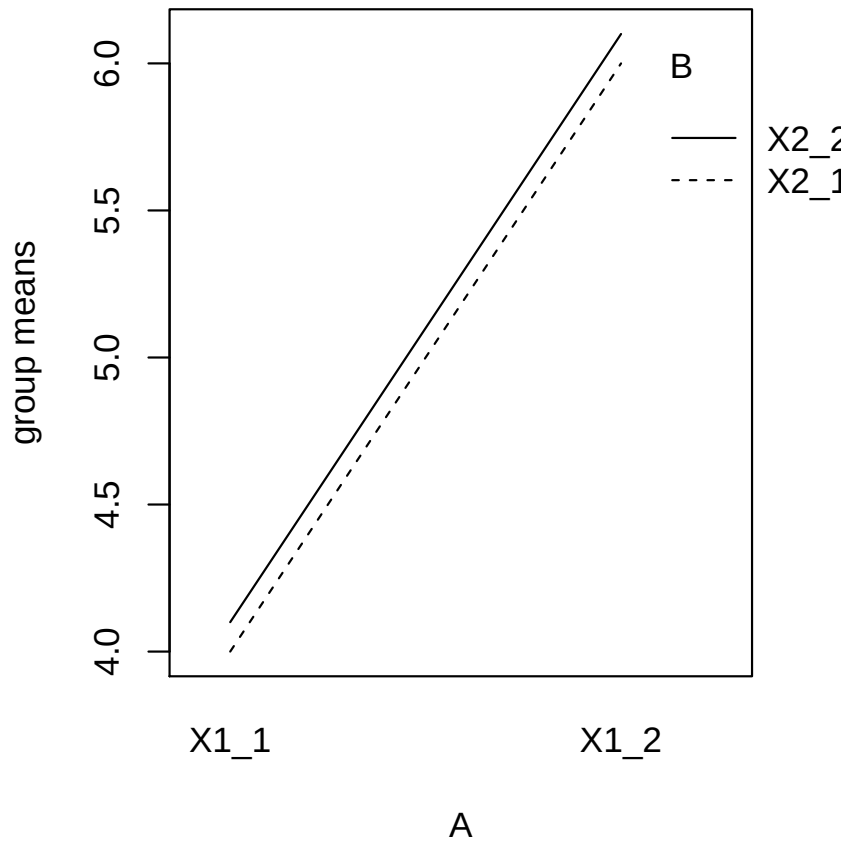
	X_1 (Design)	
X_2 (Apps)	Apple	Samsung
App Store	4	6
Google Archives	4	6

Changes on means depend only on one of the categorical variables, not the other. The expected values thus do not change when the second factor is changed.

```
lm(Y ~ X_1)
lm(Y ~ X_2)
```

2^{nd} case: effect of X_1 or X_2 only





3^{rd} case: additive effects $X_1 + X_2$

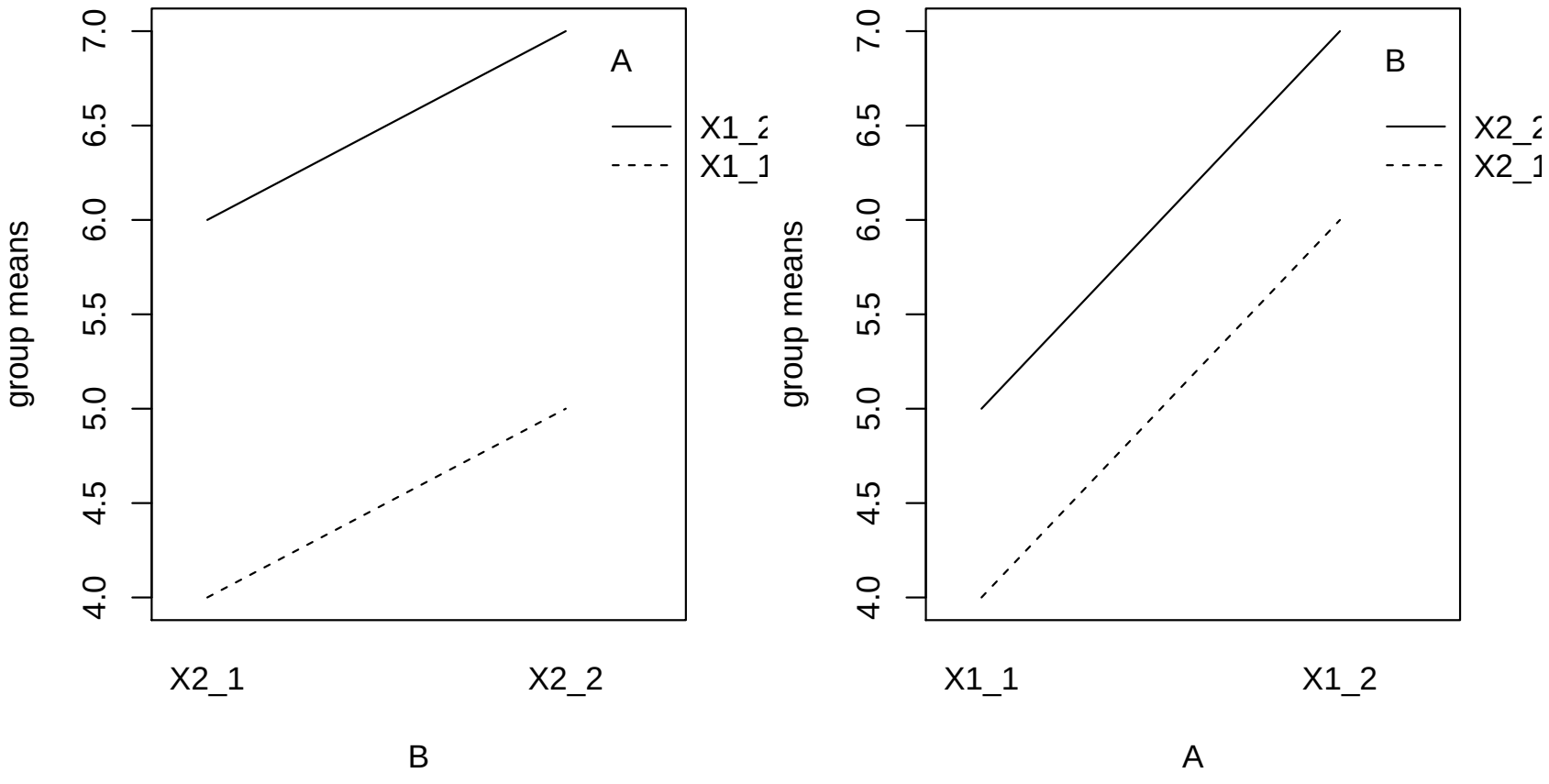
	X_1 (Design)	
X_2 (Apps)	Apple	Samsung
App Store	4	6
Google Archives	5	7

Mean of joint categories depend on both factor X_1 (design) and factor X_2 (Apps) independently.

Base model: independent additive effects of factors X_1 **and** X_2.

```
lm(Y ~ X_1 + X_2)
```

3^{rd} case: additive effects $X_1 + X_2$



4^{th} case: Interaction between X_1 and X_2

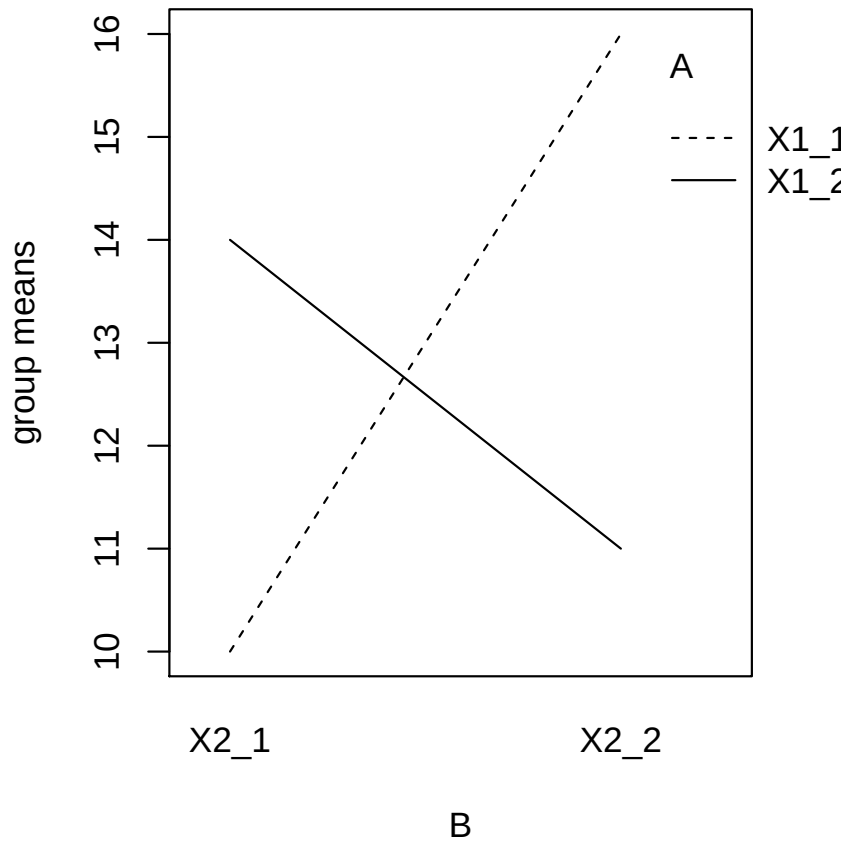
	X_1 (Design)	
X_2 (Apps)	Apple	Samsung
App Store	4	6.5
Google Archives	5.5	4.5

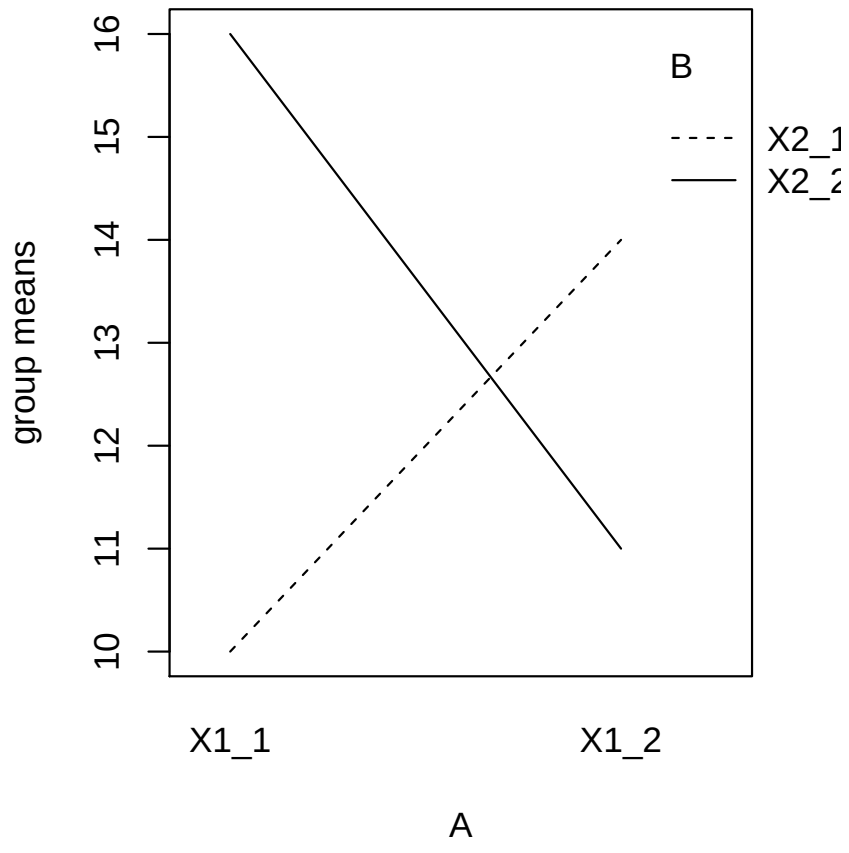
The change on group means depends on both factors X_1 and X_2 which do interact and thus yield different results in combination than each margin.

The corresponding model is the full model with interactions.

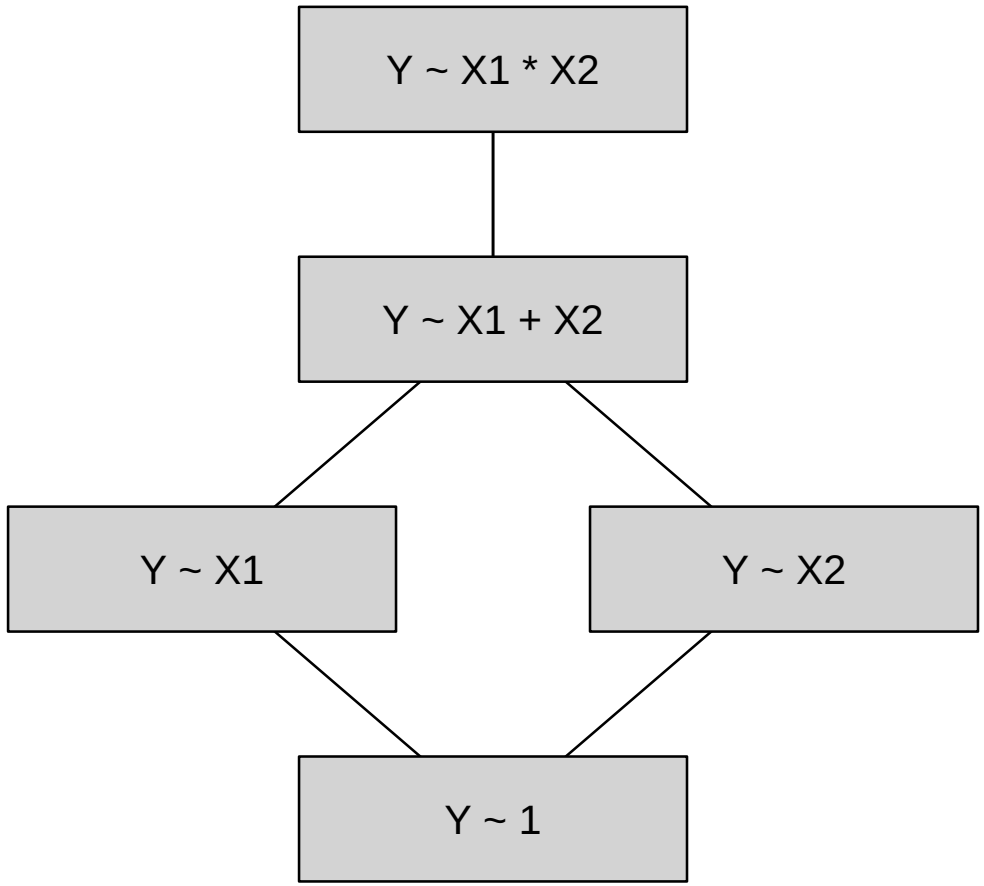
```
lm(Y ~ X_1 * X_2)
```

4^{th} case: Interaction between X_1 and X_2





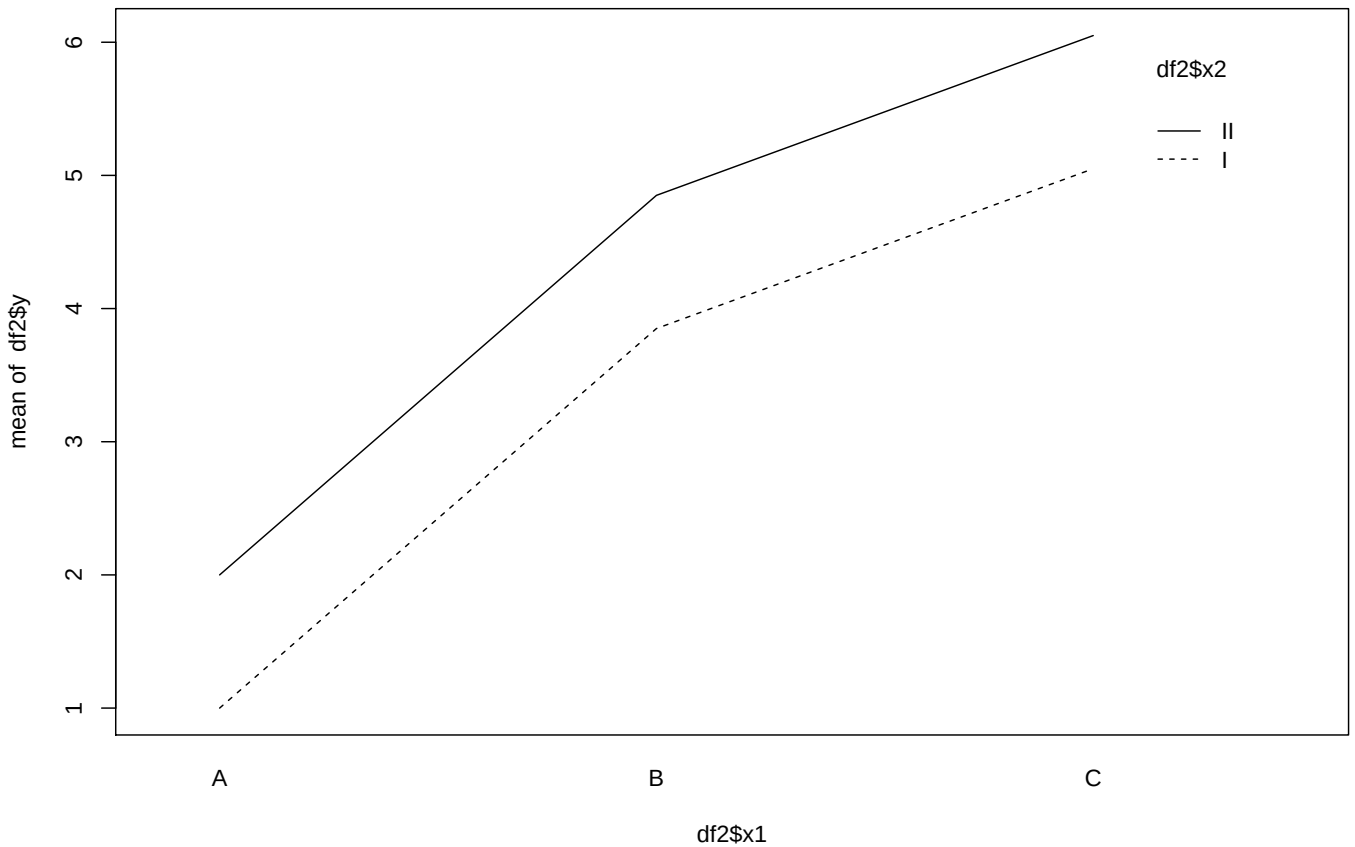
Zweiweg Varianzanalyse



ANOVA Models in R

First, we take a look at the interaction plot

```
interaction.plot(x.factor = df2$x1, trace.factor = df2$x2, response = df2$y)
```



ANOVA fit in R for additive model

ANOVA models are special cases of linear models (see linear regression) and fitted with the same command

```
# fitting additive ANOVA model
anovamodel_additive <- lm(y~x1+x2,data = df2)
anovamodel_additive
##
## Call:
## lm(formula = y ~ x1 + x2, data = df2)
##
## Coefficients:
## (Intercept)          x1B          x1C         x2II
##        1.00         2.85         4.05         1.00
```

ANOVA fit in R for interaction model

```
# fitting ANOVA model with interaction
anovamodel_interaction <- lm(y~x1*x2,data = df2)
anovamodel_interaction
##
## Call:
## lm(formula = y ~ x1 * x2, data = df2)
##
## Coefficients:
## (Intercept)          x1B          x1C         x2II     x1B:x2II     x1C:x2II
##   1.000e+00    2.850e+00    4.050e+00    1.000e+00   -1.803e-15   -1.003e-15
```

ANOVA model comparison in R

```
# comparing the ANOVA models
# with ANOVA for model comparison
anova(anovamodel_additive,anovamodel_interaction)
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x2
## Model 2: y ~ x1 * x2
##   Res.Df        RSS Df  Sum of Sq      F Pr(>F)
## 1      8 3.7964e-30
## 2      6 2.1646e-30  2 1.6318e-30 2.2616 0.1854
# the more complex model is NOT
# significantly better than the simple one
# we therefore keep the simpler model
```

Motivation

We start out from the multiple linear model

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon,$$

where we assume now we have a large amount of possible predictors $x_i, \ldots, x_p$ and consider a nested model

$$y = \beta_0 + \beta_{i_1} x_{i_1} + \ldots + \beta_{i_k} x_{i_k} + \epsilon,$$

where $k < p$ and $i_1, \ldots, i_k$ is a subset of $1, \ldots, p$.

Model selection approaches

Typically, one selects a model which is a subset out of the q predictors by one of several standard approaches:

- Testing the model coefficients independently to be equal to 0

  Student's t test is included in the R summary output

```
##                   Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)      49.99303171 6.18640648  8.0811101 4.312200e-10
## Fertility        -0.52070092 0.07868790 -6.6172931 5.139985e-08
## Agriculture      -0.22879685 0.03906287 -5.8571441 6.374386e-07
## Catholic          0.08333381 0.02178681  3.8249658 4.275083e-04
## Infant.Mortality  0.28436994 0.30040325  0.9466274 3.492434e-01
```

This has the disadvantage that the influence of removing the least relevant explanatory variable on the other model coefficients and their significance is disregarded.

Model selection approaches

A different notion of model selection is given by step-wise model selection based on **information criteria** by minimising said criteria.

- Akaike Information Criterion (AIC)

$$\mathrm{AIC} = -2\ln(\hat{L}(\theta)) + 2k$$

Because this is the negative log-likelihood function penalised by the number of model parameters q minimising the AIC is equivalent to maximising the likelihood function which is the statistical basis for contruction of many tests and models.

For linear regression AIC is equivalent to Mallow's C_p.

- Bayesian Information Criterion (BIC)

$$\mathrm{BIC} = -2\ln(\hat{L}(\theta)) + ln(n)k$$

Assumes a prior probability of each candidate model $\frac{1}{\#\text{candidate models}}$ which is combined with the likelihood function which comprises the information of the data. BIC puts a larger penalty on model size (for n>7) than AIC and thus selects smaller models.

Information Criteria based Model selection

```
step(lm(Education~.-Examination,data=swiss))
```

```
## Start:  AIC=160.13
## Education ~ (Fertility + Agriculture + Examination + Catholic +
##     Infant.Mortality) - Examination
##
##                    Df Sum of Sq    RSS    AIC
## - Infant.Mortality  1     24.46 1170.8 159.12
## <none>                          1146.3 160.13
## - Catholic          1    399.32 1545.7 172.17
## - Agriculture       1    936.34 2082.7 186.19
## - Fertility         1   1195.15 2341.5 191.69
##
## Step:  AIC=159.12
## Education ~ Fertility + Agriculture + Catholic
##
##                Df Sum of Sq    RSS    AIC
## <none>                      1170.8 159.12
## - Catholic      1    410.71 1581.5 171.25
## - Agriculture   1   1079.89 2250.7 187.84
## - Fertility     1   1289.36 2460.2 192.02

##
## Call:
## lm(formula = Education ~ Fertility + Agriculture + Catholic,
##     data = swiss)
##
## Coefficients:
## (Intercept)    Fertility  Agriculture     Catholic
##     53.8505      -0.4888      -0.2380       0.0844
```

lineare Regression mit Nebenbedingungen

Anstatt das lineare Modell

$$y_i = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_k x_{k,i} + \varepsilon_i$$

mithilfe der kleinsten Quadrate Methode anzupassen, wird eine Optimierung mit Nebenbedingungen durchgeführt.

- **Ridge Regression** mit der Nebenbedingung

$$\sum_{i=1}^{n} \beta_i^2 < c \text{ für ein } c > 0$$

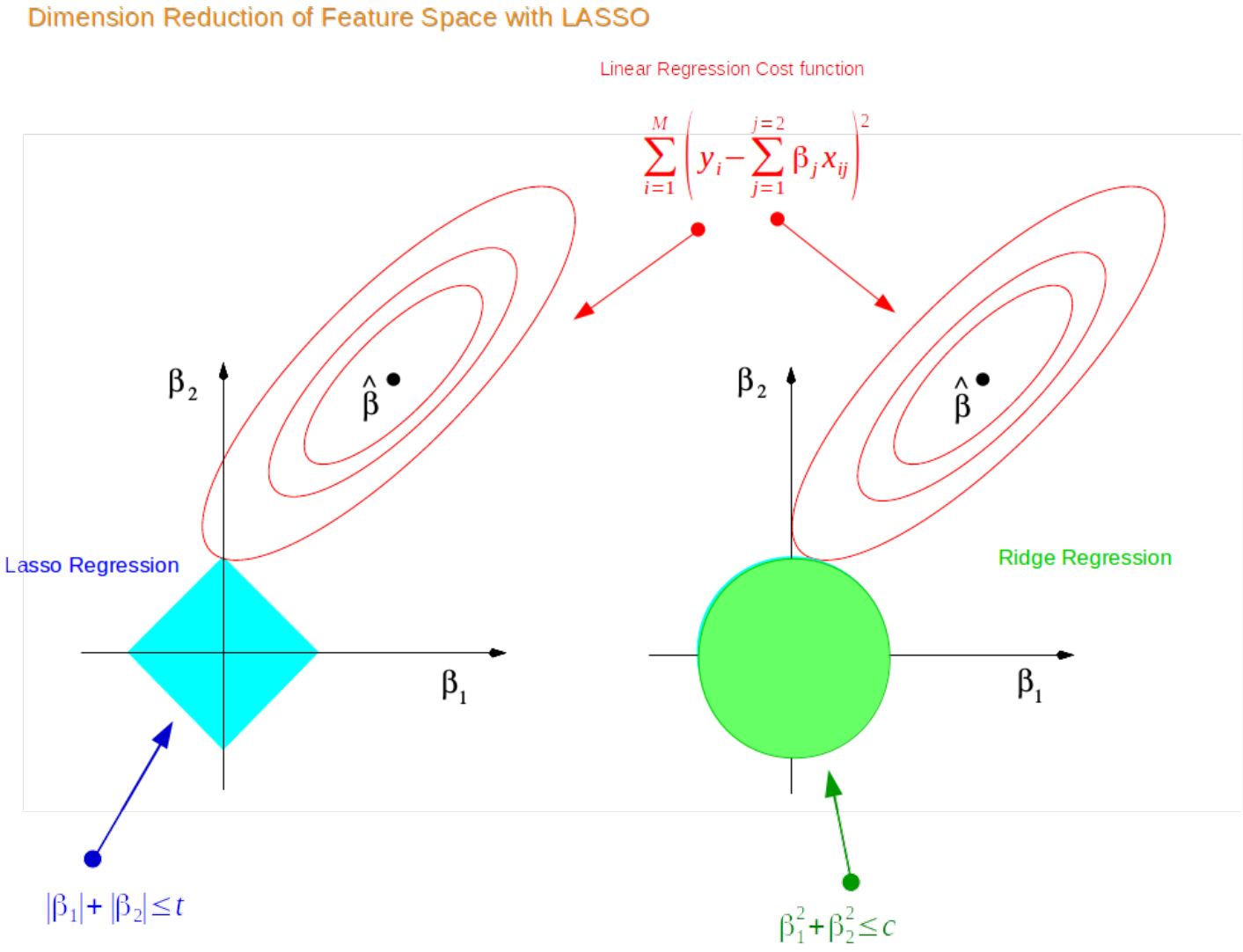
- **LASSO** **l**east **a**bsolute **s**hrinkage and **s**election **o**perator

$$\sum_{i=1}^{n} |\beta_i| < c \text{ für ein } c > 0$$

- sowie ihre Kombination **elastic nets** in Machine Learning

$$(1-\alpha) \cdot \sum_{i=1}^{n} \beta_i^2 + \alpha \cdot \sum_{j=1}^{n} |\beta_j| < c \text{ für ein } c > 0, p \in [0;1]$$

lineare Regression mit Nebenbedingungen



Quelle: Scikit

Ellipsen

$$x^2 + y^2 = r^2$$

$$(x/a)^2 + (y/b)^2 = r^2$$

lineare Regression mit Nebenbedingungen

Diese Methoden

- **Ridge Regression**
- **LASSO Regression**
- sowie ihre Kombination **elastic nets**

sind einfache Techniken, um Modellkomplexität und Überanpassung zu verhindern, die mit einfacher linearer Regression auftreten können.

Elastic Net in R

```
library(glmnet)

glmnet(x,y,family, alpha)
```

family beschreibt die Art, wie die abhängige Variable y modelliert werden soll.

family="gaussian" ist numerisch wie lineare Regression.

family= "binomial" ist binär wie bei logistischer Regression.

alpha ist der "elasticnet mixing parameter", with $0 \leq \alpha \leq 1$.

alpha=1 ist LASSO, alpha=0 Ridge Regression.

Example for Ridge Regression and LASSO in R

```
library(glmnet)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## Loading required package: foreach
```

```
##
## Attaching package: 'foreach'
```

```
## The following objects are masked from 'package:purrr':
##
##     accumulate, when
```

```
## Loaded glmnet 2.0-18
```

Ridge Regression in R

The connection between linear regression and Ridge regression is via the penalty term of the squared coefficients.

```
fitLM <- lm(y~X)
round(coef(fitLM),2)
```

```
## (Intercept)          X1          X2          X3          X4          X5
##        2.01        3.04        3.94        4.92       -0.12        0.24
##          X6          X7          X8          X9         X10         X11
##       -0.12       -0.16       -0.15       -0.06       -0.17       -0.08
##         X12         X13         X14         X15         X16         X17
##       -0.14        0.43       -0.08       -0.03        0.02        0.00
##         X18         X19         X20
##       -0.21       -0.18       -0.31
```

```
# root sum of squared coefficients
sqrt(sum(coef(fitLM)[-1]^2))
```

```
## [1] 7.044342
```

Ridge Regression in R

The package **glmnet** can perform ridge regression in R using the function `glmnet`. To obtain a ridge regression there, the parameter `alpha` needs to be set to zero.

```
lambda.grid <- 10^seq(10, -2,length=100)
fitRR <- glmnet(x=X, y=y, alpha=0, lambda=lambda.grid)
dim(coef(fitRR))
```

```
## [1]  21 100
```

Lasso in R

The function `glmnet` can also compute the lasso. For that purpose the parameter `alpha` needs to be set to one.

```
fitL <- glmnet(x=X, y=y, alpha=1, lambda=lambda.grid)
dim(coef(fitL))
```

```
## [1]  21 100
```

Ridge Regression in R II

```
fitRR$lambda[50]
```

```
## [1] 11497.57
```

```
round(coef(fitRR)[,50],2)
```

```
## (Intercept)          V1          V2          V3
##        2.01        0.00        0.00        0.00
##          V6          V7          V8          V9
##        0.00        0.00        0.00        0.00
##         V12         V13         V14         V15
##        0.00        0.00        0.00        0.00
##         V18         V19         V20
##        0.00        0.00        0.00
```

```
sqrt(sum(coef(fitRR)[-1,50]^2))
```

```
## [1] 0.004567586
```

Lasso in R II

Unlike Ridge regression LASSO shrinks the irrelevant coefficients to 0.

```
fitL$lambda[50]
```

```
## [1] 11497.57
```

```
round(coef(fitL)[,50],2)
```

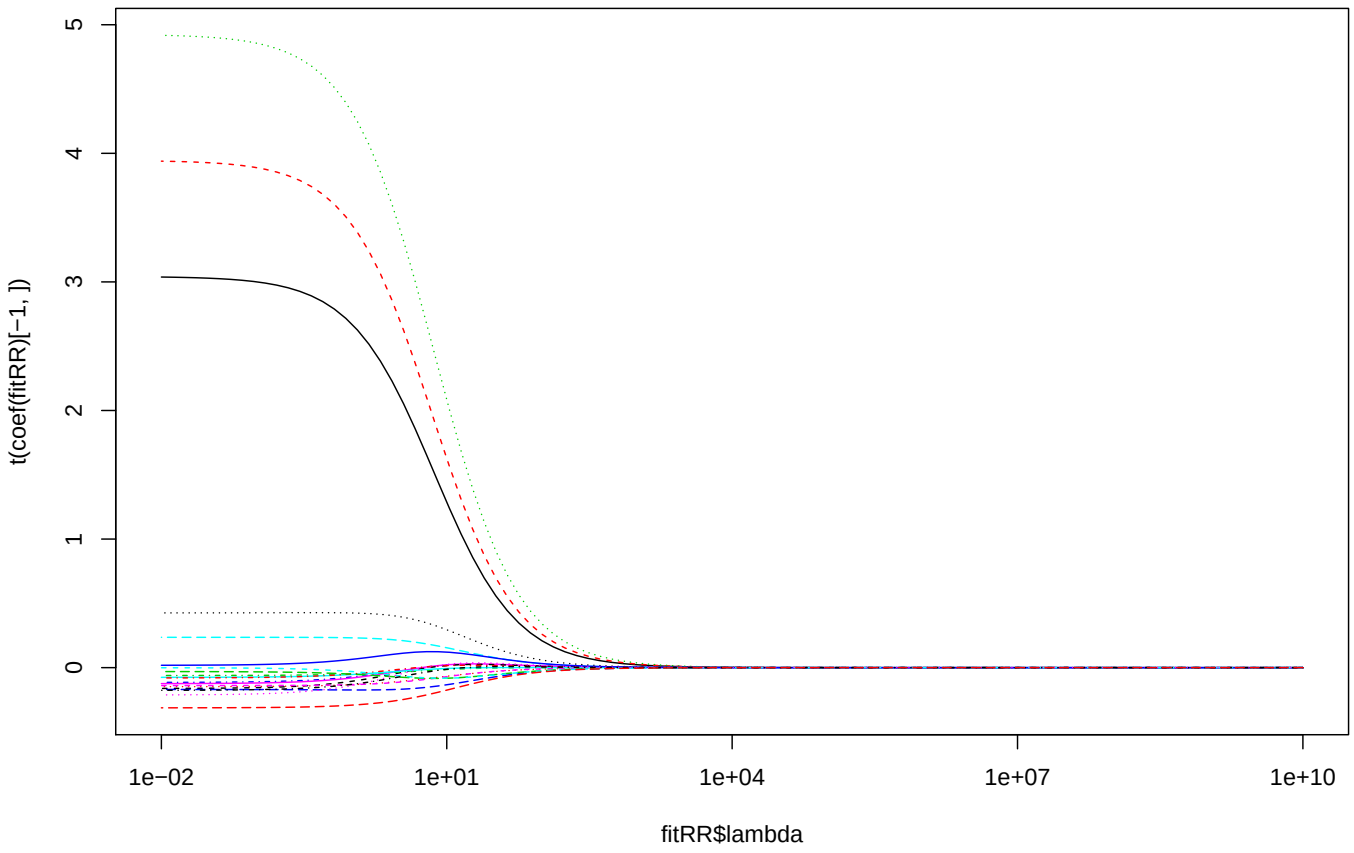
```
## (Intercept)          V1          V2          V3          V4          V5
##        2.01        0.00        0.00        0.00        0.00        0.00
##          V6          V7          V8          V9         V10         V11
##        0.00        0.00        0.00        0.00        0.00        0.00
##         V12         V13         V14         V15         V16         V17
##        0.00        0.00        0.00        0.00        0.00        0.00
##         V18         V19         V20
##        0.00        0.00        0.00
```

```
sqrt(sum(coef(fitL)[-1,50]^2))
```

```
## [1] 0
```

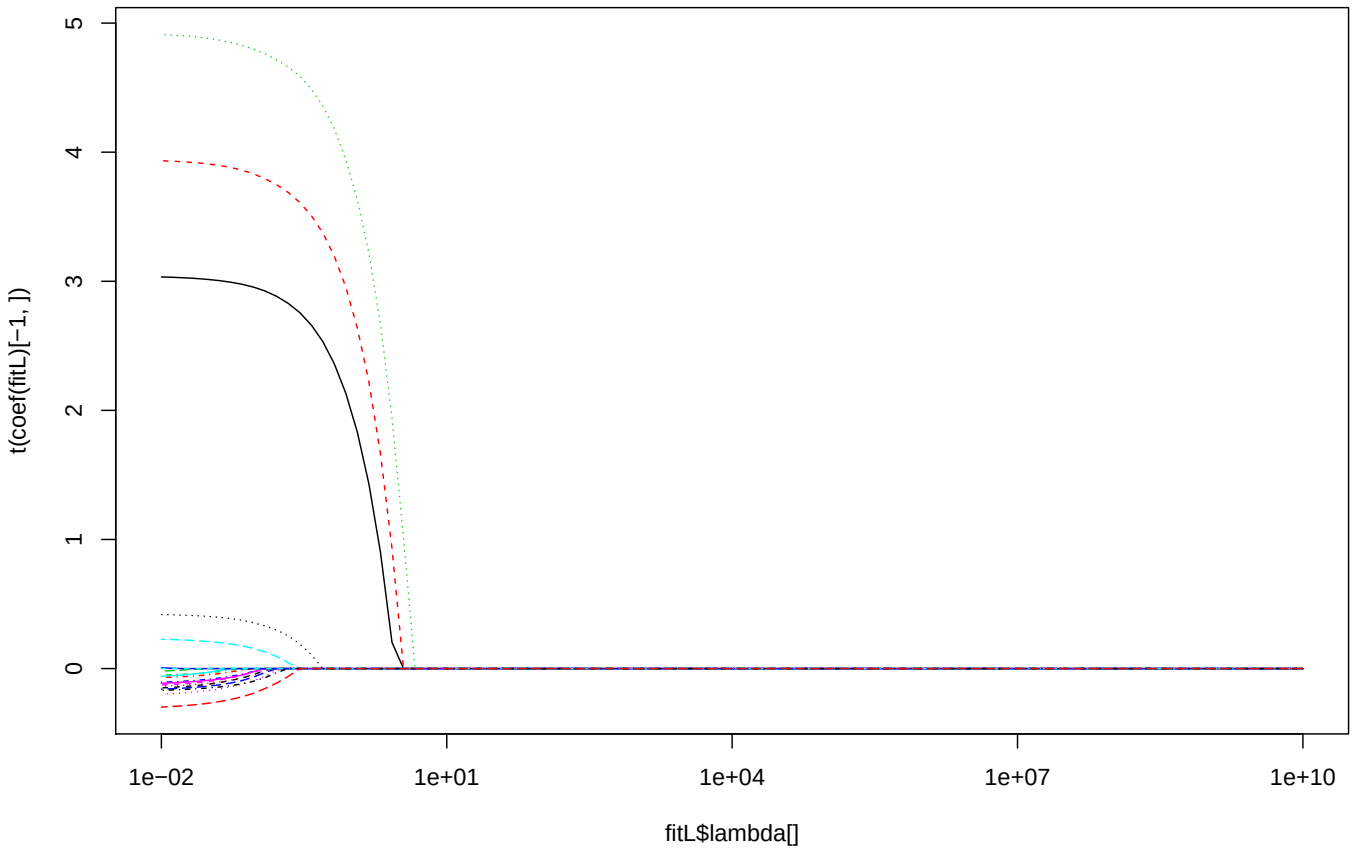
Ridge Regression in R

```
matplot(fitRR$lambda, t(coef(fitRR)[-1,]), type="l",log="x")
```



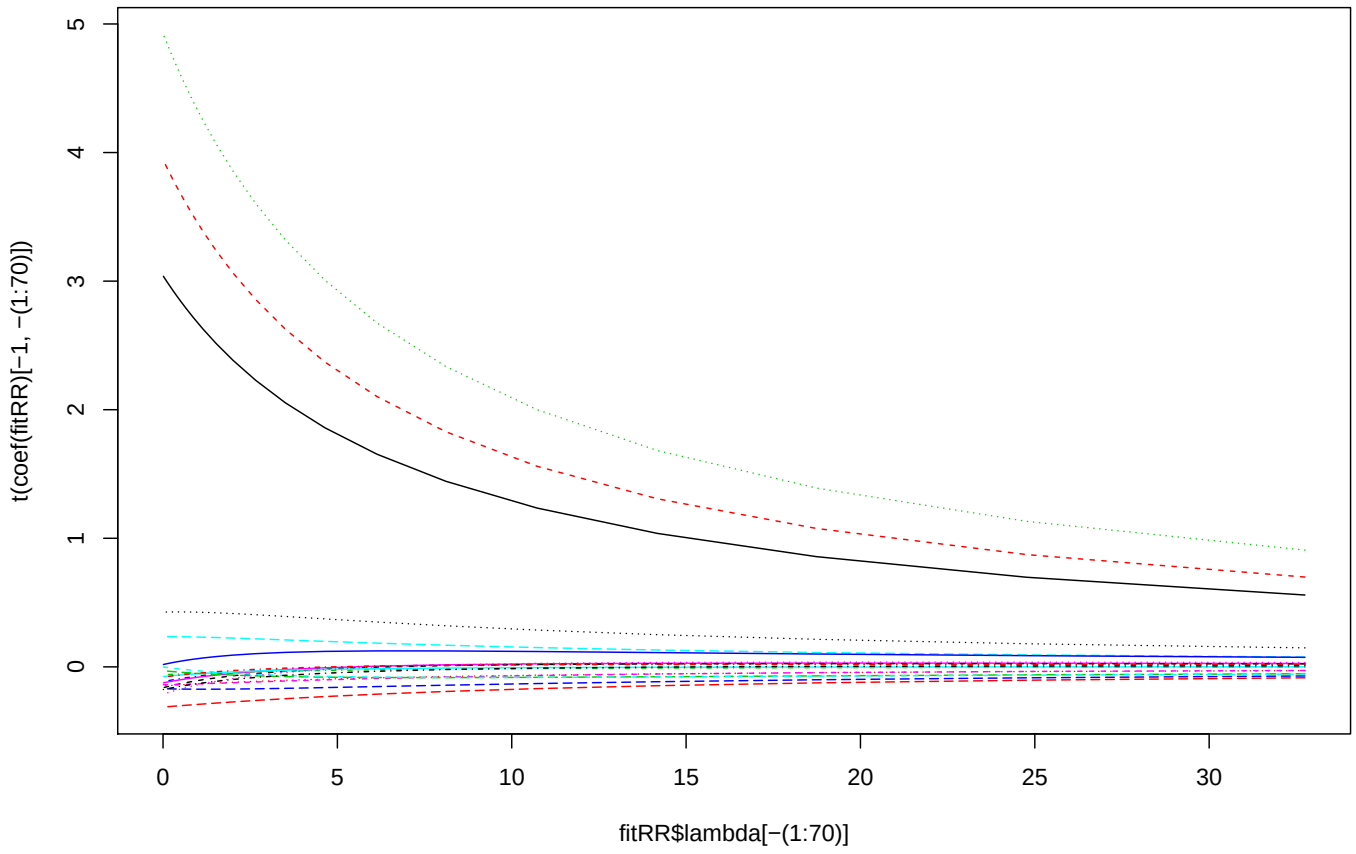
Lasso in R

```
matplot(fitL$lambda[], t(coef(fitL)[-1,]), type="l",log="x")
```



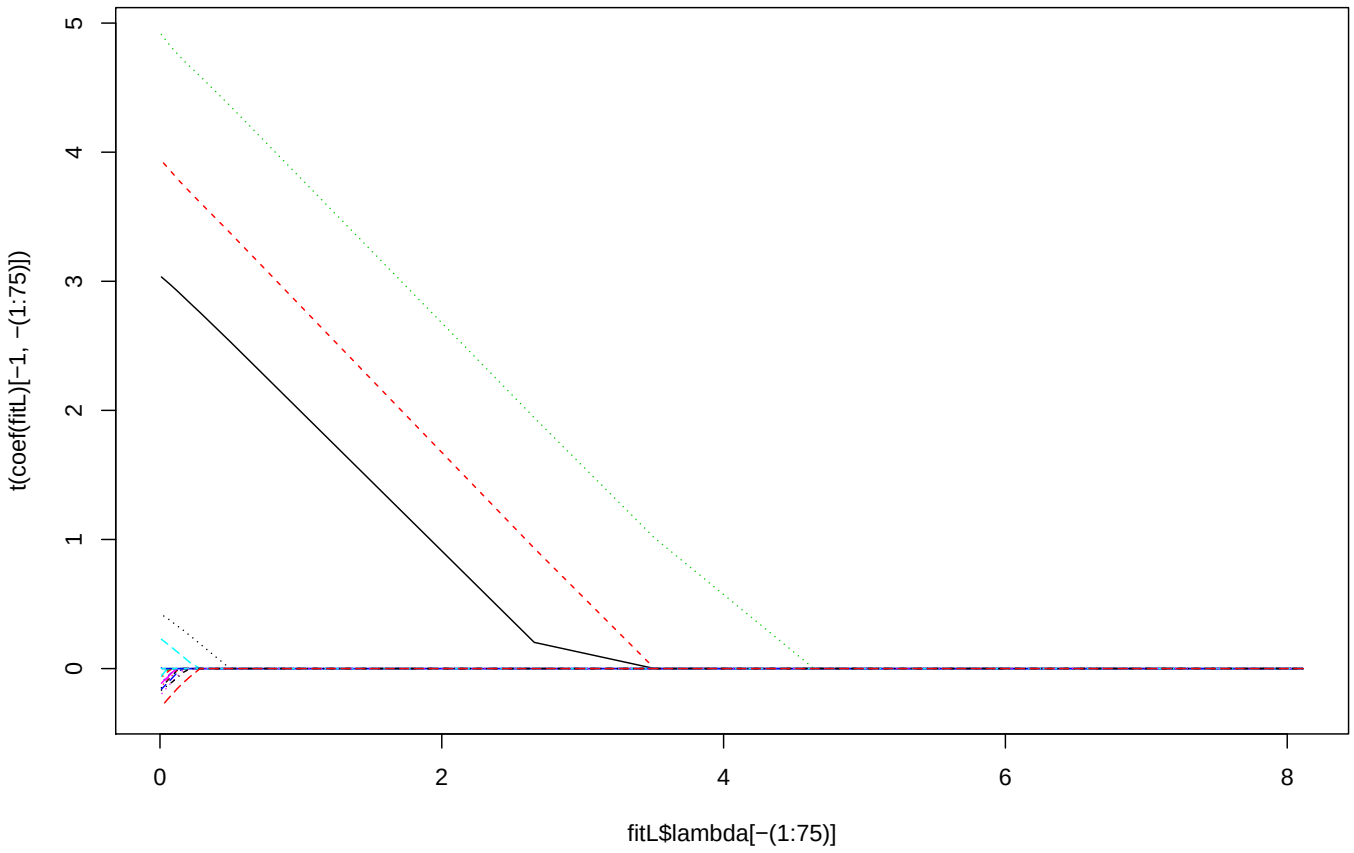
Ridge Regression in R

```
matplot(fitRR$lambda[-(1:70)], t(coef(fitRR)[-1,-(1:70)]), type="l")
```



Lasso in R

```
matplot(fitL$lambda[-(1:75)], t(coef(fitL)[-1,-(1:75)]), type="l")
```



Verallgemeinerte Regression - binäre Klassifikation

Wir möchten eine ähnliche Methode wie für numerische y-Werte verwenden, um binäre Outcomes anzupassen, also Klassifikation in 2 Kategorien zu betreiben. Binäre Daten werden mithilfe der Binomialverteilung modelliert, es ist also

$$f(y \mid p) \quad = \quad p^y \cdot (1-p)^{1-y}$$

wobei p die Erfolgswahrscheinlichkeit ist und y die Werte 0 (Misserfolg) oder 1 (Erfolg) annehmen kann.

Der Erwartungswert für den Ausgang y ist $E(y) = p$. Also in $p \cdot 100\%$ der Fälle erwartet man sich für y den Wert 1, in $(1-p) \cdot 100\%$ der Fälle erwartet man sich für y den Wert 0.

Verallgemeinerte Regression - binäre Klassifikation

Um nun ein Modell für den Erwartungswert $E(y_i) = \mu_i = p_i$ einer binären Variable zu formulieren, soll wieder ein linearer Prädiktor $x_i^\top \beta$ verwendet werden.

Problem: Während der lineare Prädiktor $x_i^\top \beta$ prinzipiell alle reellen Werte annehmen kann, liegt der Erwartungswert (= Erfolgswahrscheinlichkeit) $\mu_i = p_i$ immer im Intervall $[0, 1]$.

Lösung: Verwende eine Link-Funktion g, die das Intervall $[0, 1]$ auf die reellen Zahlen abbildet, so daß

$$g(\mu_i) \quad = \quad x_i^\top \beta$$

Frage: Was ist eine *geeignete* Link-Funktion?

Exkurs: Vierfeldertafeln

In der explorativen Datenanalyse wurde bereits der Zusammenhang von zwei kategorialen Merkmalen mit Hilfe ihrer Kontingenztafel und dem zugehörigen Mosaikplot untersucht

Beispiel: `Vorsorgeuntersuchung` und `Geschlecht`

Geschlecht vs. Vorsorgeuntersuchung	nein	ja
Frauen	273	183
Männer	627	217

Um zu untersuchen, ob sich die Wahrscheinlichkeit für Männer und Frauen unterscheidet, zur Vorsorgeuntersuchung zu gehen, sind die relativen Häufigkeiten bedingt unter dem Geschlecht geeigneter.

Geschlecht vs. Kauf	nein	ja	Summe
Frauen	0.599	0.401	1
Männer	0.743	0.257	1

Chancen - Odds

Um diese vier bedingten relativen Häufigkeiten noch weiter zu komprimieren, betrachtet man anstatt der **Wahrscheinlichkeiten** die zugehörigen **Chancen**. Die Chance eines Ereignisses ist die Wahrscheinlichkeit für sein Eintreten geteilt durch die Gegenwahrscheinlichkeit. Im englischen heissen diese **Odds**.

$$\text{Odds}(\text{Vorsorgeuntersuchung}) = \frac{P(\text{Vorsorgeuntersuchung})}{1 - P(\text{Vorsorgeuntersuchung})}$$

Bei den Frauen beträgt die Chance auf die Vorsorgeuntersuchung also $0.401/0.599 = \frac{2/5}{3/5} = 0.67$, also in etwa 2:3.

Bei den Männern hingegen beträgt die Chance auf die Vorsorgeuntersuchung $0.257/0.743 = \frac{1/4}{3/4} = 0.346$ also in etwa 1:3.

Chancenverhältnis

Wenn man auch noch diese beiden Zahlen ins Verhältnis setzt, so nennt man das Ergebnis Chancenverhältnis oder Odds Ratio.

Hier besagt das Chancenverhältnis von $0.346/0.67 = 0.516$, daß die Chancen auf die Vorsorgeuntersuchung bei Männern nur rund halb so groß sind wie bei Frauen.

Das Chancenverhältnis kann auch leicht aus der Originaltabelle berechnet werden.

Geschlecht vs. Vorsorgeuntersuchung	nein	ja
Frauen	273	183
Männer	627	217

$$\mathrm{OddsRatio} = \frac{273 \cdot 217}{627 \cdot 183} = 0.516$$

Logistische Regression

Als Link-Funktion in der logistischen Regression wird die sogenannte Logit-Funktion

$$g(\mu) \quad = \quad \log\left(\frac{\mu}{1-\mu}\right)$$

verwendet. Diese berechnet die logarithmierten Chancen.

Dadurch wird die logistische Regression

$$E(y_i) = \mu_i = p_i$$

mit Linkfunktion

$$g(\mu_i) \quad = \quad x_i^\top \beta$$

als Modell für die binäre Variable y angepasst.

Anders gesagt, passt man das lineare Modell für die Fraktion der Wahrscheinlichkeiten an

Logistische Regression

Das Ergebnis der logistischen Regressionsfunktion

$$\mathit{logit}(y) = \log\left(\frac{\mathbb{P}[y_i = 1|\mathbb{X}]}{\mathbb{P}[y_i = 0|\mathbb{X}]}\right) = \alpha + \beta \cdot x_i + \varepsilon_i$$

ist die Ermittlung der Wahrscheinlichkeiten in Gruppe 0 oder 1 zu fallen:

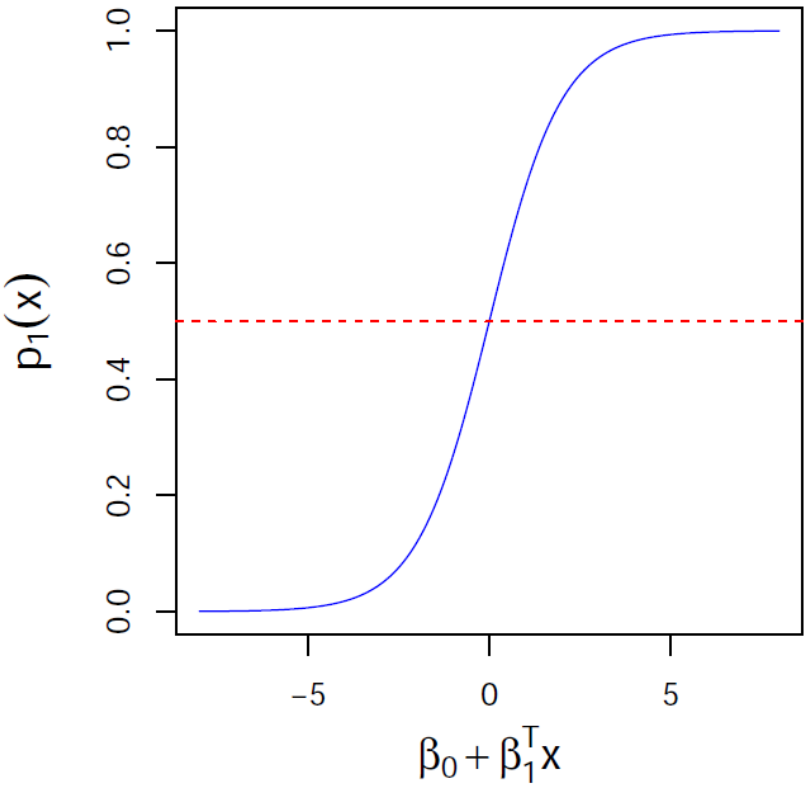
$$\mathbb{P}[y = 1|\mathbb{X}] = \frac{\mathit{exp}(\alpha + \beta \cdot x)}{1 + \mathit{exp}(\alpha + \beta \cdot x)}$$

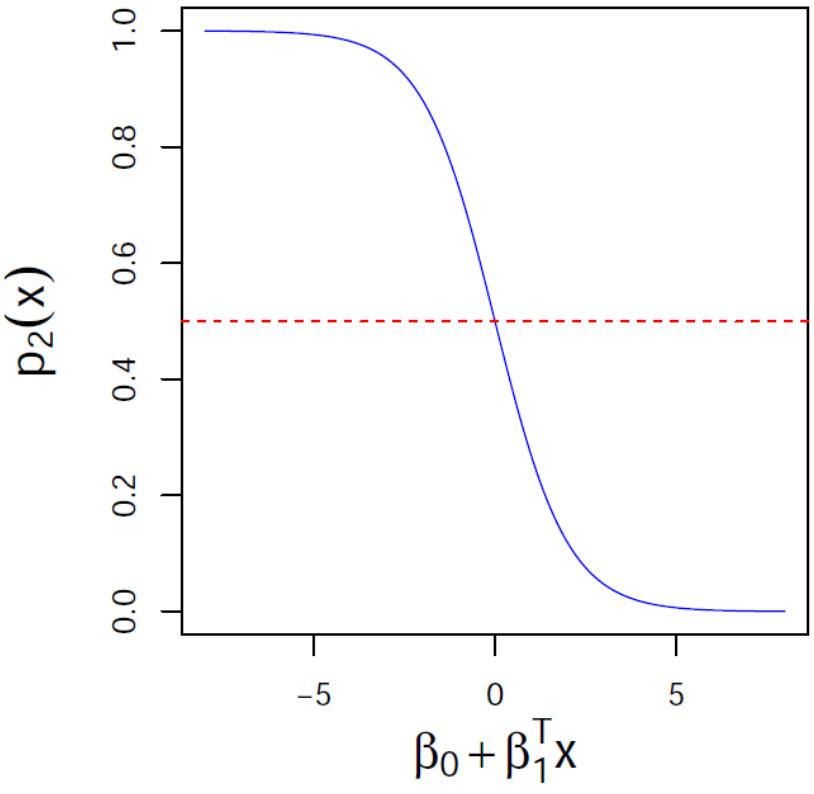
und

$$\mathbb{P}[y = 0|\mathbb{X}] = \frac{1}{1 + \mathit{exp}(\alpha + \beta \cdot x)}$$

welche die logistische Funktion beschreiben.

Logistische Regression





Beispiel: Prüfungsdaten - logistische Regression

Hours	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50
Pass	0	0	0	0	0	0	1	0	1	0
Hours	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass	1	0	1	0	1	1	1	1	1	1

```
glm(Pass~Hours,family=binomial(link ="logit"))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.08	1.76	-2.32	0.02
Hours	1.50	0.63	2.39	0.02

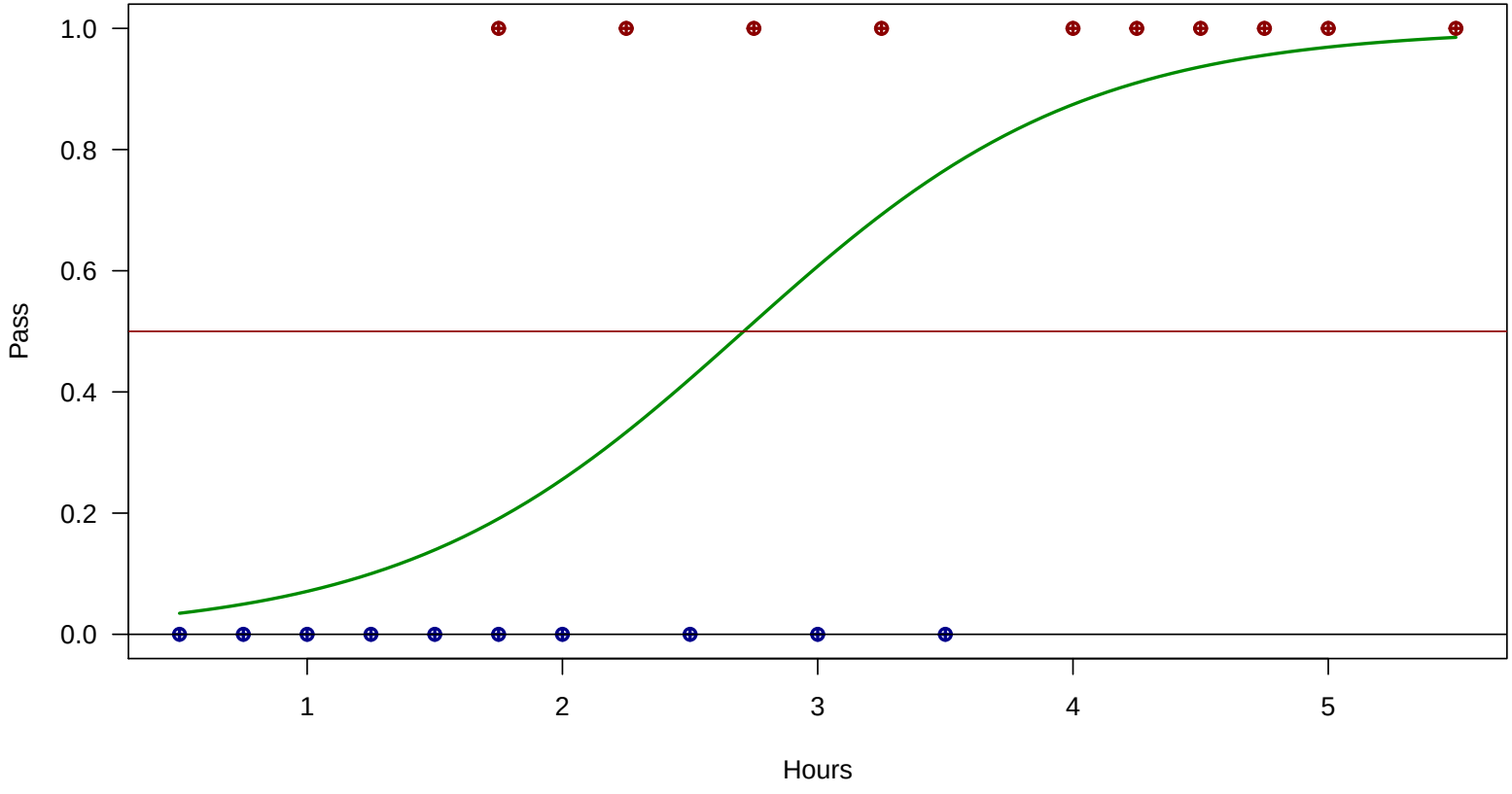
$0.017 =$ Odds-Ratio, die Prüfung zu schaffen, wenn 0 Stunden gelernt wird

$4.502 =$ Faktor, um den der Odds-Ratio, die Prüfung zu schaffen, steigt, für jede Stunden, die gelernt wird

$exp(-4.078 + 1.505 \cdot Hours) =$ Odds-Ratio, die Prüfung zu schaffen, wenn Hours Stunden gelernt wird

Beispiel: Prüfungsdaten - logistische Regressionskurve

$$\frac{1}{1 + exp(-(-4.078 + 1.505 \cdot Hours))} = \text{Wahrscheinlichkeit, die Prüfung mit Hours Stunden Lernen zu bestehen (= probability of passing the exam)}$$



Eindimensionale Modelle im Überblick

Wir betrachten die uns bekannten Modelle mit einer Einflussvariable

$y \sim 1$	triviales Modell (lm) (1 Mittelwert für die gesamten Daten)
$y \sim a$	Einweg-Varianzanalyse (lm) (2 Mittelwerte für 2 Kategorien der Variable a)
$y \sim x$	Lineare Regression (lm) (1 Regressionsgerade $y = \alpha + \beta \cdot x$)
$y \sim x$	Logistische Regression (glm(,family=binomial("logit"))) (Klassifikation der binären Variable y mithilfe von Daten x)

Zweidimensionale Modelle im Überblick

Wir betrachten die uns bekannten Modelle mit zwei Einflussvariablen

$y \sim a1 + a2$	Zweiweg-Varianzanalyse (lm) (4 Mittelwerte für je 2 Kategorien der Variable a1 und je 2 Kategorien der Variable a2)
$y \sim a1 * a2$	Zweiweg-Varianzanalyse (lm) (4 Mittelwerte für je 2 Kategorien der Variable a1 und je 2 Kategorien der Variable a2 plus Interaktionen zwischen a1 und a2)
$y \sim x1 + x2$	Lineare Regression (lm) (1 Regressionsebene $y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2$)
$y \sim x1 + x2$	Logistische Regression (glm(,family=binomial("logit"))) (Klassifikation der binären Variable y mithilfe von Daten x1 und x2)

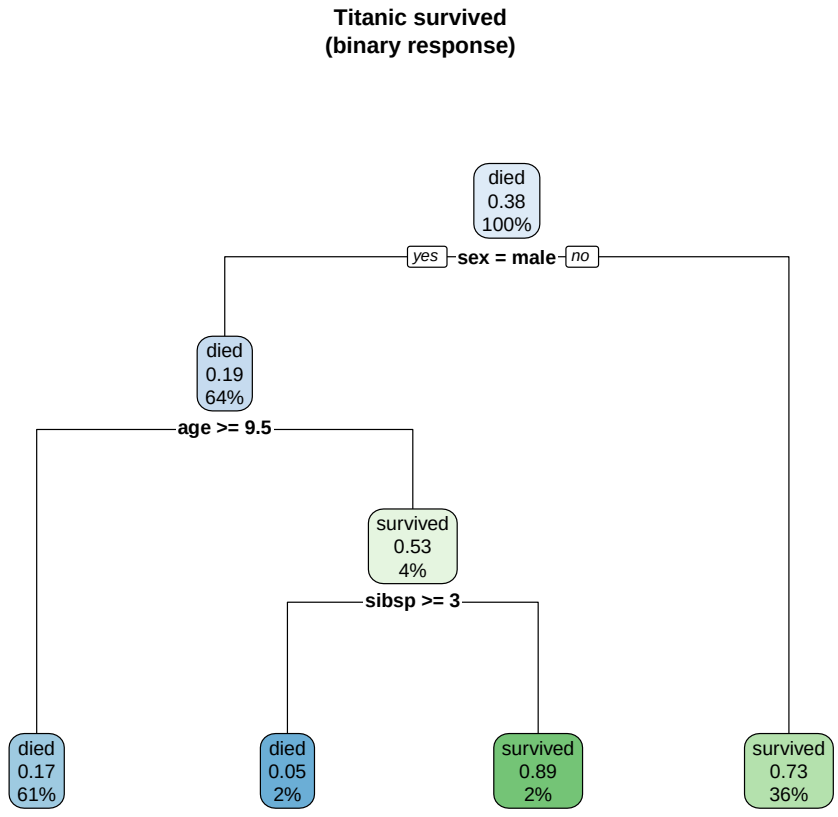
Regression Trees

Classification and Regression Tree (CART)
Anstatt ein Modell anzupassen, wird bei Baum-basierten Methoden in Untergruppen partitioniert. Die rekursive Partitionierung basiert auf Schwellwerten aus den zufälligen Untergruppen, sodass die Residuenquadratsumme der Aufteilung minimiert wird.
Dementsprechend werden Indikatoren, in welchen Untergruppenbereich Beobachtungen fallen, aus dem Baum vorhergesagt.

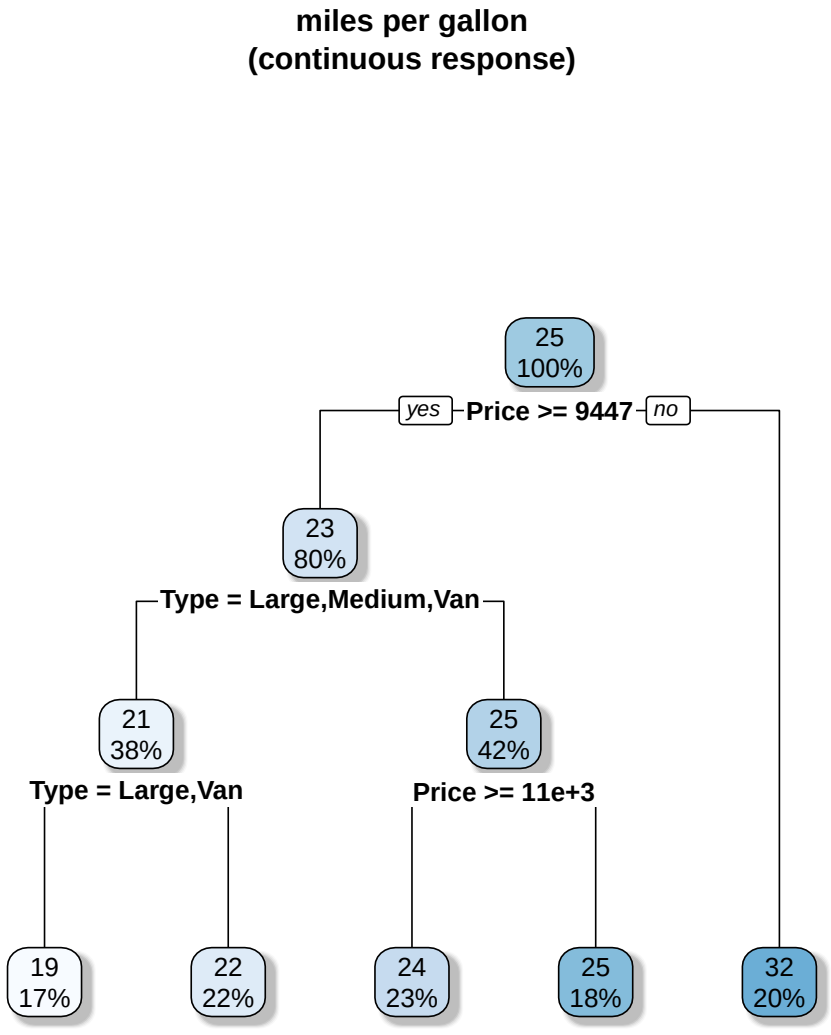
```
library(rpart); rpart(y~x1+x2,daten, method)
```

method ="class" for a classification tree OR "anova" for a regression tree

Regression Tree



Regression Tree



Regression Tree

Vorteile:

- einfach und direkt interpretierbar ohne mathematisches Modell
- Prädiktionen sind schnell und einfach
- mit teilweise fehlenden Daten immer noch verwendbar bis zur Verzweigung, wo Daten fehlen
- einfache, schneller Algorithmus

Nachteile:

- hohe Varianz zwischen unterschiedlichen Berechnungen
- sehr hoher Fehler in der Prädiktion