UNIVERSITÄT LEIPZIG

# Eine Einführung in R: Das Lineare Modell II

**Bernd Klaus, Verena Zuber**
Institut für Medizinische Informatik, Statistik und Epidemiologie (IMISE),
Universität Leipzig

15. Dezember 2011

1. Modelldiagnose
2. Interpretation der Koeffizienten
   - Metrische erklärende Variablen
   - Kategoriale erklärende Variablen
   - Testen der Regressionskoeffizienten
   - Interaktionen
3. Variablenselektion
4. Prädiktion

# I. Modelldiagnose

# Wiederholung: Residuenanalyse

Frage: Sind die Voraussetzungen für das lineare Modell erfüllt?
Zu untersuchen sind:

1. **Anpassung des Modells an die Daten**:
   $\rightarrow$ Residuen gegen gefittete Wert $\hat{Y}$
2. **Normalverteilung des Fehlers**:
   $\rightarrow$ QQ-Plot: Quantile der Residuen gegen die theoretische NV
3. **Homoskedastizität des Fehlers**:
   $\rightarrow$ Standardisierte Residuen gegen gefittete Wert $\hat{Y}$,
   wenn die geeignet mit $H$ standardisierten Residuen abhängig von $\hat{Y}$ sind, deutet dies auf ungleiche Varianzen der Fehler hin

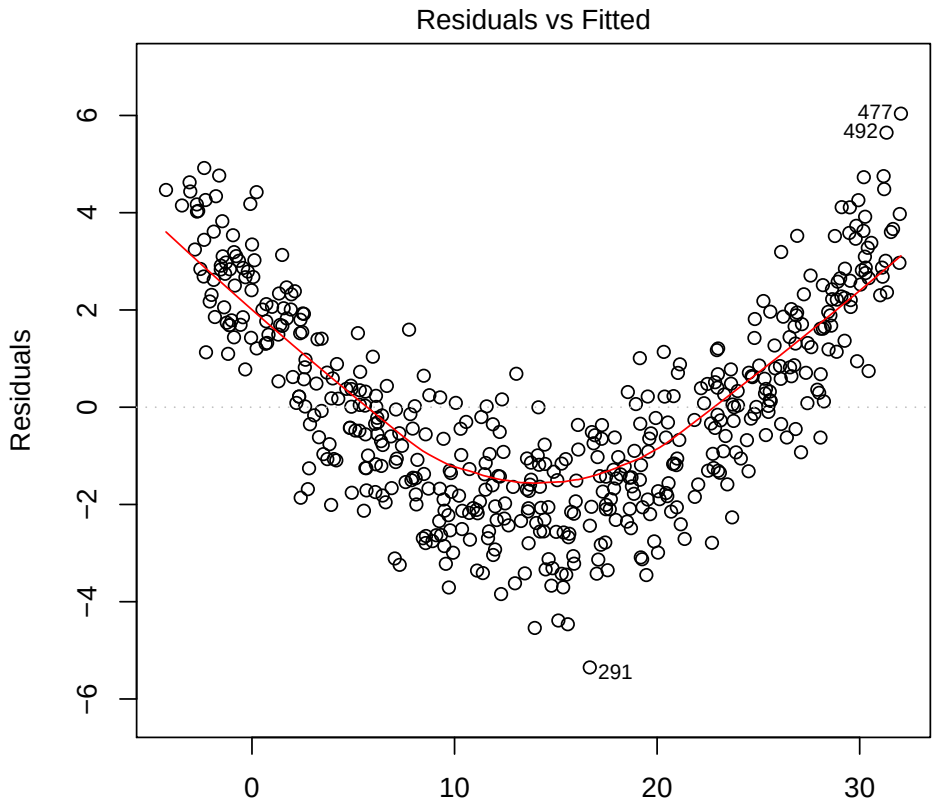
# Beispiele: Simulationen

```
h1<-seq(1,6,0.01)
X<-h1+rnorm(length(h1), mean=0, sd=0.1)
```

1. Kein linearer, sondern quadratischer Zusammenhang:
   ```
   epsilon1<-rnorm(length(X), mean=0, sd=1)
   Y1<-X*X+epsilon1
   ```
2. Kein Normal-, sondern gleichverteilter Fehler:
   ```
   epsilon2<-runif(length(X), min=-1, max=1)
   Y2<-X+epsilon2
   ```
3. Die Fehler haben unterschiedliche Varianz, bzw sind abhängig von $Y$:
   ```
   epsilon3<-rnorm(length(X),
   mean=rep(0,length(X)), sd=X)
   Y3<-X+epsilon3
   ```

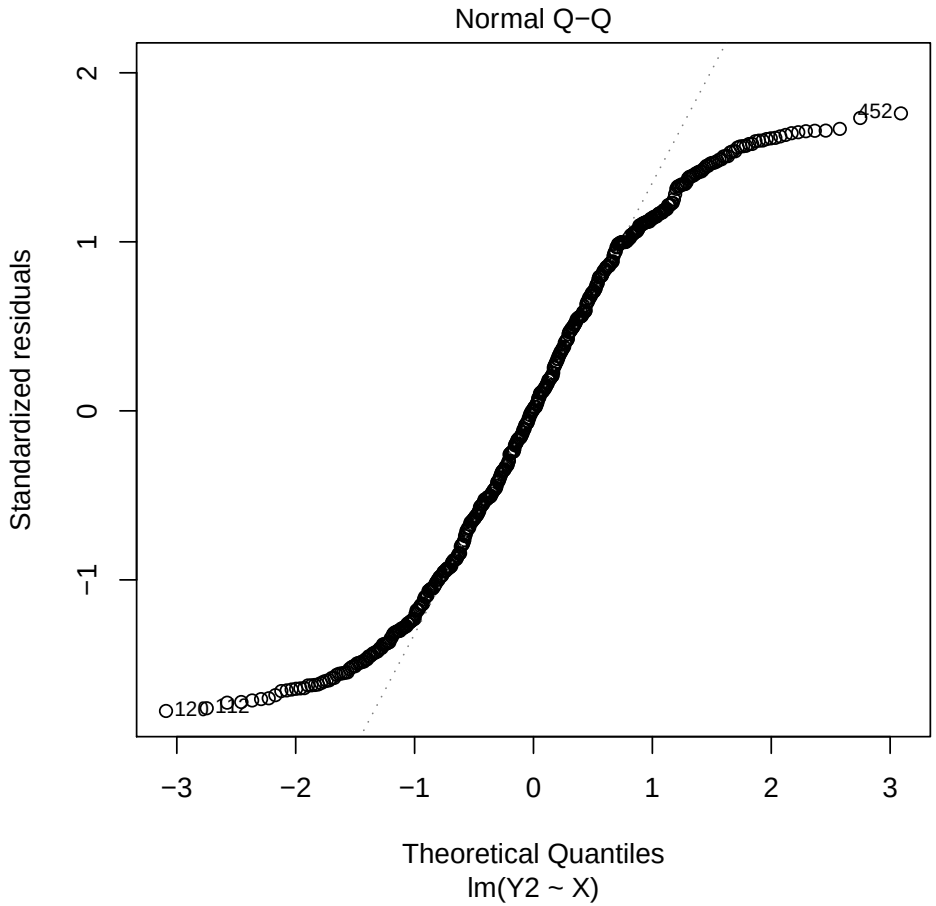
# Modelldiagnose in R I: Residuen gegen gefittete Werte

- Residuen gegen gefittete Werte $\hat{Y}$ zur Untersuchung der Anpassung des Modells an die Daten



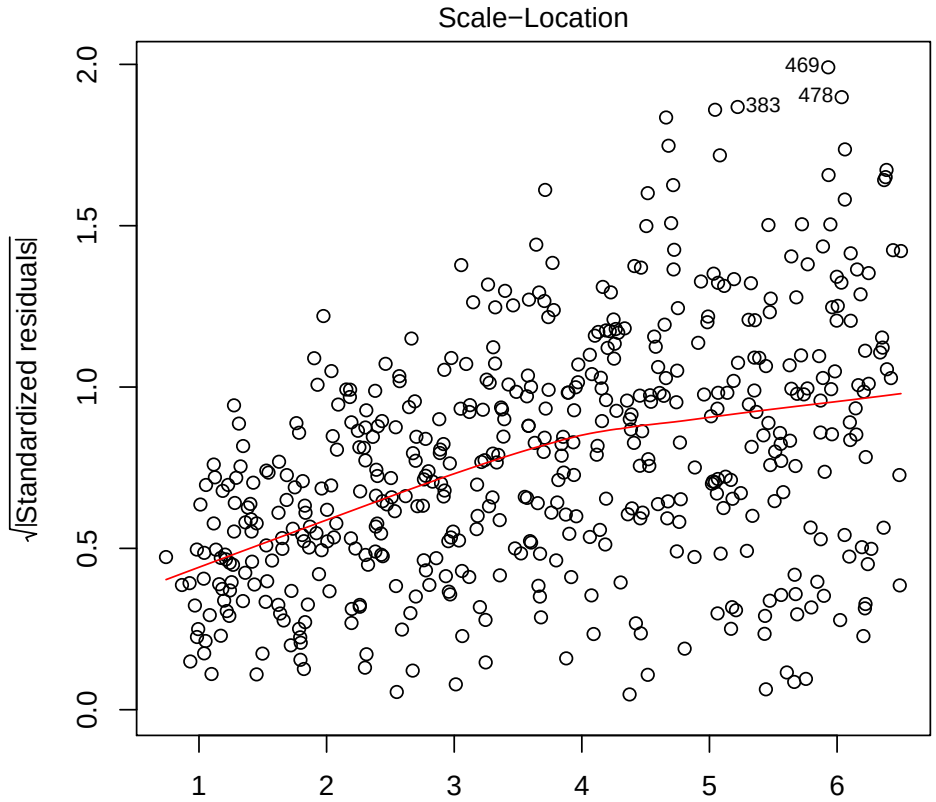
# Modelldiagnose in R II: Residuen-QQ

- Plot der studentisierten gegen die theoretischen (NV) Residuen zur Untersuchung der Normalverteilung des Fehlers



# Modelldiagnose in R III: Standardisierte Residuen gegen $\hat{Y}$

- Standardisierte, absolute Residuen gegen gefittete Werte $\hat{Y}$ zur Untersuchung der Homoskedastizität des Fehlers



# $R^2$ oder das adjustierte $R^2$

In der multiplen Regression wird zum $R^2$ meist auch das adjustierte $R^2$ ausgegeben:

$$R^2_{adjust} = R^2 - \frac{p - (1 - R)}{n - p - 1}$$

Dann können auch Modelle verglichen werden, die eine unterschiedliche Zahl $p$ an Variablen besitzen.

Weitere Kriterien sind Akaike Information Criterion (AIC), Baysian Information Criterion (BIC) und viele mehr!

# Interpretation der Koeffizienten

# Metrische erklärende Variablen

- Der Regressionskoeffizient $\beta_j$ einer Variable $X_j$ gibt deren Einfluss auf die Zielgröße $Y$ unter gleichzeitiger Kontrolle der anderen Variablen an.
- Bei einer metrischen Variable $X_j$ gilt:
  (Gegeben die übrigen $(p-1)$ Variablen werden festgehalten)
  Wenn sich $X_j$ um eine Einheit erhöht, so verändert sich $Y$ um $\beta_j$ Einheiten.
- Beispiel: Datensatz "airquality"

$$\texttt{Ozone}_i = -145.7 + 2.27847 \cdot \texttt{Temp}_i + 0.05711 \cdot \texttt{Solar.R}_i + \varepsilon_i$$

- "Wenn die Temperatur (bei gegebener Sonneneinstrahlung) um eine Einheit steigt, steigt die Ozonkonzentration um ca. 2.3 Einheiten"

# Kategoriale erklärende Variablen

- Warnung: Kategoriale Variablen $X_j$ können nicht wie metrische interpretiert werden!
- Vorgehensweise:
  1. Wähle eine Kategorie als Referenz
  2. Führe binäre Variablen ("Dummyvariablen") ein, die angeben, ob eine Beobachtung in die Referenzkategorie oder in eine andere Kategorie fällt
  3. Wenn $k$ Kategorien vorliegen, müssen $k-1$ Dummyvariablen konstruiert werden
  4. Interpretation:
     (Gegeben die übrigen $(p-1)$ Variablen werden festgehalten)
     Wenn eine Beobachtung nicht in die Referenzkategorie fällt, so verändert sich $Y$ um $\beta_j$ Einheiten
- Deswegen ist es in R essentiell, kategoriale Variablen als Faktoren zu führen (dann berechnet R die Dummyvariablen automatisch)

# Beispieldaten: "Work"

Untersuchung verschiedener Einflussfaktoren (`COMP`, `RTW`, `PVT`) auf den prozentualen Anteil der Beschäftigten im öffentlichen Sektor `DENS`, die in einer Gewerkschaft organisiert sind, in verschiedenen amerikanischen Bundesstaaten.

- Metrische Variablen:
  - `DENS`: *Percent of public sector employees in unions, 1982*
  - `PVT`: *Percent of private sector employees in unions, 1982*
- Kategoriale Variablen:
  - `COMP`: *State bargaining laws cover public employees (1) or not (0)* (Referenzkategorie: Keine Rechte)
  - `RTW`: *State right-to-work law (1) or not (0)*

Zunächst ist folgendes lineares Modell von Interesse:

$$\texttt{DENS}_i = \beta_0 + \alpha_{RTW_i} + \beta_1 \cdot \texttt{PVT}_i + \varepsilon_i$$

# Beispieldaten: "Work"

- `Work <- read.table("Work.csv", header = TRUE)`
- Umwandlung von `RTW` in einen Faktor
  `Work$RTW <- as.factor(Work$RTW)`
- `test <- lm(DENS ~ RTW + PVT, data = Work)`

Ausgabe in `R`:

```
 Coefficients:
(Intercept)    RTW1      PVT
35.3881      -10.8599   0.1418
```

Interpretation von $\alpha_{RTW}$: Gibt es ein "Recht auf Arbeit", so verringert sich der Anteil der im öffentlichen Dienst in einer Gewerkschaft organisierten Beschäftigten um ca. 11%

# Testen der Regressionskoeffizienten

Der standardisierte Regressionskoeffizient ist $t$-verteilt mit einer Freiheitsgradzahl, die sich aus dem Stichprobenumfang $n$ und der Variablenzahl $p$ bestimmt:

$$T_j = \frac{\hat{\beta}_j}{\mathrm{SD}(\hat{\beta}_j)} \quad \sim t(n-p-1)$$

$\mathrm{SD}(\hat{\beta}_j)$: Standardabweichung von $\hat{\beta}_j$

- $H_0 : \beta = 0$ **ablehnen, falls** $|T_j| > t_{1-\alpha/2}(n-p-1)$
- $H_0 : \beta > 0$ **ablehnen, falls** $T_j < t_{\alpha}(n-p-1)$
- $H_0 : \beta < 0$ **ablehnen, falls** $T_j > t_{1-\alpha}(n-p-1)$

R gibt in der `summary` sowohl die $\beta$s (`estimate`), deren **Standardabweichung** (`Std. Error`) und $t$-**Statistik** (`t value`) und $p$-**Wert** ( `Pr(>|t|)`) an.

# Interaktionen

- Das Modell "test" besitzt nur ein $R^2$ von 0.25
- Wahrscheinlich sind wichtige Einflussfaktoren noch nicht berücksichtigt !
- Wir untersuchen daher das Modell:

$$\texttt{DENS}_i = \beta_0 + \alpha_{RTW_i} + \alpha_{COMP_i} + \alpha_{COMP_i * RTW_i} + \beta_1 \cdot \texttt{PVT}_i + \varepsilon_i$$

- Der Koeffizient $\alpha_{COMP_i * RTW_i}$ beschreibt eine multiplikative Interaktion der Faktoren `COMP` und `RTW`
- D.h. dieser Effekt besteht, wenn gleichzeitig Recht auf Arbeit UND Tarifverhandlungen der Gewerkschaftsmitglieder im öffentlichen Dienst gegeben sind.

# Interaktionen - Umsetzung in R

- Das erweiterte Modell

$$\texttt{DENS}_i = \beta_0 + \alpha_{RTW_i} + \alpha_{COMP_i} + \alpha_{COMP_i * RTW_i} + \beta_1 \cdot \texttt{PVT}_i + \varepsilon_i$$

- Aufruf der Funktion `lm()`
- `testI <- lm(DENS ~ COMP*RTW + PVT, data = Work)`

Ausgabe in R:

```
Coefficients:
(Intercept)    COMP1      RTW1       PVT     COMP1:RTW1
  27.31371   14.92008   -0.58751   0.04727   -18.38723
```

# Interaktionen - Interpretation der Ergebnisse

- `COMP1`: Gibt es ein Recht auf Tarifverhandlungen der Gewerkschaftsmitglieder im öffentlichen Dienst, aber KEINES auf Arbeit, STEIGT der Anteil der öffentlich Beschäftigten, die gewerkschaftlich organisiert sind, um 14.9%.
- `RTW1`: Gibt es ein Recht auf Arbeit, aber KEINES auf Tarifverhandlungen der Gewerkschaftsmitglieder im öffentlichen Dienst, SINKT der Anteil der öffentlich Beschäftigten, die gewerkschaftlich organisiert sind, marginal um 0.6%.
- `COMP1:RTW1`: Gibt es aber sowohl ein Recht auf Arbeit, als auch auf Tarifverhandlungen der Gewerkschaftsmitglieder im öffentlichen Dienst, so SINKT der Anteil der öffentlich Beschäftigten, die gewerkschaftlich organisiert sind, um ca. $18.4\% + 0.6\% - 14.9\% = 3.5\%$

# Variablenselektion

# Variablenselektion im Linearen Modell

- Ziel der Regressionsanalyse ist es oft, ein möglichst "gutes" Modell mit möglichst "wenig" Variablen zu erhalten (vergleiche **Occam's Razor**)
- Es gibt zwei weit verbreitete Varianten zur Variablenselektion in der linearen Regression:

1. **Test der Regressionkoeffizienten wie oben beschrieben**
   - $\Rightarrow$ Test einzelner Variablen auf Signifikanz
2. **Schrittweise Selektion von verschachtelten Modellen mittels F-test**
   - $\Rightarrow$ Test auf signifikante zusätzliche Varianzerklärung
   - verschachtelt heißt, dass das "volle" Modell das "reduzierte" komplett enthalten muss

# Beispieldaten

Datensatz `cystfibr` aus dem R- Paket `ISwR`. Er enthält Daten von Mukoviszidose-Patienten (Chronische Lungenkrankheit).

- Körpermaße
  - **age** age in years.
  - **sex** 0: male, 1:female.
  - **height** height (cm).
  - **weight** weight (kg).
  - **bmp** body mass (% of normal).
- Lungenfunktionsmaße
  - **fev1** forced expiratory volume.
  - **rv** residual volume.
  - **frc** functional residual capacity.
  - **tlc** total lung capacity.
  - **pemax** maximum expiratory pressure.

Zielgrösse ist **pemax**, die durch die anderen Variablen erklärt werden soll.

# 1. Selektion mittels $t$-Test I

## Vollständiges Modell - adjusted $R^2 = 0.42$

pemax $\sim$ age+sex+height+weight+bmp+fev1+rv+frc+tlc

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 176.0582 | 225.8912 | 0.779 | 0.448 |
| age | -2.5420 | 4.8017 | -0.529 | 0.604 |
| sex | -3.7368 | 15.4598 | -0.242 | 0.812 |
| height | -0.4463 | 0.9034 | -0.494 | 0.628 |
| weight | 2.9928 | 2.0080 | 1.490 | 0.157 |
| bmp | -1.7449 | 1.1552 | -1.510 | 0.152 |
| fev1 | 1.0807 | 1.0809 | 1.000 | 0.333 |
| rv | 0.1970 | 0.1962 | 1.004 | 0.331 |
| frc | -0.3084 | 0.4924 | -0.626 | 0.540 |
| tlc | 0.1886 | 0.4997 | 0.377 | 0.711 |

# 1. Selektion mittels $t$-Test II

- $\Rightarrow$ Kein einzelner Prediktor signifikant!
- Problem: Korrelation unter den Kovariablen.
- z.B. `cor(frc,tlc) = 0.7`, `cor(age, height) = 0.9`
- $\Rightarrow$ Körpermaße und Lungenfunktionsmaße untereinander stark korreliert!
- wähle einige Repräsentanten aus jeder Gruppe aus!

**Reduziertes Modell**

pemax $\sim$ age+bmp+sex+fev1+tlc+rv

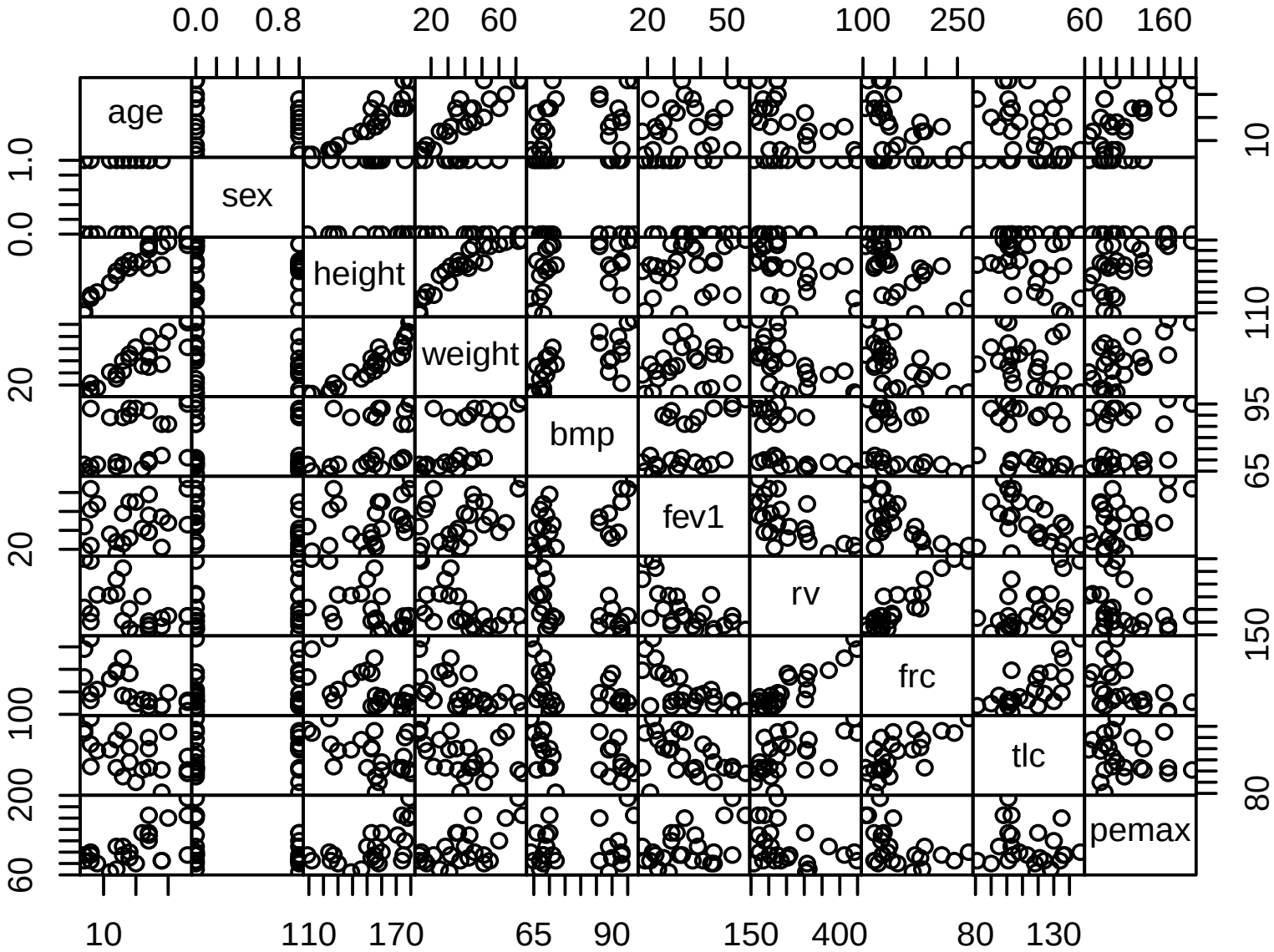
# Einschub: Korrelationsstruktur erkennen

- Mittels `pairs` lässt sich ein guter Eindruck von der Korrelationsstruktur bekommen
- `cor` gibt paarweise Korrelationen oder die Korrelationsmatrix aus

## Korrelationstruktur untersuchen

```
pairs(cystfibr, gap=0, cex.labels=3)
cor(cystfibr)
```

# Einschub: Pairs-Plot für `cystfibr`



# 1. Selektion mittels $t$-Test III

## Reduziertes Modell

pemax age+bmp+sex+fev1+tlc+rv

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | -83.5254 | 82.8081 | -1.009 | 0.32650 | |
| age | 5.0380 | 1.2956 | 3.889 | 0.00108 | ** |
| bmp | -0.4030 | 0.5638 | -0.715 | 0.48397 | |
| sex | 4.9907 | 12.9130 | 0.386 | 0.70367 | |
| fev1 | 1.9313 | 0.7800 | 2.476 | 0.02345 | * |
| tlc | 0.4651 | 0.4046 | 1.149 | 0.26543 | |
| rv | 0.1135 | 0.1007 | 1.127 | 0.27450 | |

Entferne schrittweise alle Variablen mit hohen $p$-Werten!, zuerst **sex**

# 1. Selektion mittels $t$-Test IV

Dieses Vorgehen ergibt

**Endgültiges Modell**

pemax $\sim$ age+fev1+rv

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | -83.5254 | 82.8081 | -1.009 | 0.32650 | |
| age | 4.54157 | 1.19442 | 3.802 | 0.00104 | ** |
| fev1 | 1.57425 | 0.60316 | 2.610 | 0.01635 | * |
| rv | 0.16122 | 0.08998 | 1.792 | 0.08761 | . |

adjusted $R^2 = 0.46$, also besser als beim vollen Modell!

# 1. Selektion mittels $F$-Test I

- Wir benutzen das reduzierte Modell nach Berücksichtigung von der Korrelationen innerhalb der **Körpermaße** und **Lungenfunktionsmaße**

## Reduziertes Modell

anova(lm(pemax$\sim$ age+bmp+sex+fev1+rv+tlc))

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) | |
|---|---|---|---|---|---|---|
| age | 1 | 10098.5 | 10098.5 | 15.7840 | 0.0008921 | *** |
| bmp | 1 | 0.2 | 0.2 | 0.0002 | 0.9877060 | |
| sex | 1 | 963.8 | 963.8 | 1.5065 | 0.2354890 | |
| fev1 | 1 | 1944.2 | 1944.2 | 3.0388 | 0.0983542 | . |
| rv | 1 | 1464.5 | 1464.5 | 2.2891 | 0.1476481 | |
| tlc | 1 | 845.2 | 845.2 | 1.3211 | 0.2654344 | |

# 1. Selektion mittels $F$-Test II

- Die $F$-Test-Tabelle entspricht dem schrittweisen entfernen der Variablen von unten nach oben!
- Der $F$-Wert von `tlc` ergibt sich alternativ als

**$F$-Wert von `tlc`**

anova(lm(pemax ~ age+bmp+sex+fev1+tlc), lm(pemax~ age+bmp+sex+fev1))

- Wir gehen nach Korrelationsgruppen vor und entfernen erst `tlc`, dann `bmp` und `sex`
- Dies ergibt, dass nur age und `fev1` im Modell verbleiben sollten

# 1. Selektion mittels $F$-Test III

Ergebnis:

**Modell nur mit age und fev**

```
summary(lm(pemax age+fev1))
```

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 27.8637 | 20.1497 | - 1.383 | 0.18059 | ** |
| age | 3.4735 | 1.0858 | 3.199 | 0.00414 | ** |
| fev1 | 0.8917 | 0.4906 | 1.818 | 0.08275 | . |

adjusted $R^2 = 0.40$

# Variablenselektion im Linearen Modell - Fazit

- Es gibt **keinen** "Königsweg" bei der Variablenselektion
- Berücksichtigung von Korrelation sehr wichtig
- Variablenselektion = eigener Forschungszweig der Statistik, hier wurden nur die zwei wichtigsten Methoden gezeigt
- Nicht so sehr auf "Signifikanzen" achten!
- $\Rightarrow$ führt man viele Tests gleichzeitig durch, findet man fast zwangsläufig Signifikanzen! (Stichwort: multiples Testen)

# Prädiktion

# Vorhersage im Linearen Modell

- Gegeben: Lineares Modell mit Regressionskoeffizienten, die auf Basis bestehender Daten ermittelt wurden
- Neu: Eine neue Beobachtung $X_{n+1}$ deren Zielgröße $Y_{n+1}$ unbekannt ist
- Ziel: Vorhersage der unbekannten Zielgröße $Y_{n+1}$

**Vorgehensweise zur Vorhersage:**

- Bilde eine Vorhersageregel (*prediction rule*) aus dem gegeben Modell
- Setze die Werte der neuen Beobachtung $X_{n+1}$ in diese Vorhersageregel ein und berechne die Vorhersage $\hat{Y}_{n+1}$

# Anwendungsbeispiele

- Verwendung von sogenannten “Biomarkern” (wie bestimmte Genexpressionswerte, Variationen der DNA, oder bestimmte Ausprägungen der Proteinstruktur) zur Vorhersage von Krebsarten, Alter bis Ausbruch von Alzheimer, uvm.
- Neurowissenschaften: Vorhersage von Handlungen, Emotionen auf Basis von bestimmten Mustern in der Gehirnaktivität.
- Aus der Wirtschaft, Prognose aus Zeitreihen zur Konjunkturlage.
- Komplexe räumlich-zeitliche Modelle zu Klimaprognosen.

# Beispiel: *Airquality*-Daten

- Gegeben: Lineares Modell mit Regressionskoeffizienten aus dem Datensatz "airquality"

$$\texttt{Ozone}_i = -145.7 + 2.27847 \cdot \texttt{Temp}_i + 0.05711 \cdot \texttt{Solar.R}_i + \varepsilon_i$$

- Neu: 3 neue Beobachtungen $X_{n+1}$ `newdata`, deren Zielgröße $Y_{n+1}$ unbekannt sind

| Ozone | Solar.R | Temp |
|---|---|---|
| ? | 80 | 110 |
| ? | 80 | 112 |
| ? | 80 | 114 |

- Ziel: Vorhersage der unbekannten Zielgröße $Y_{n+1}$

# Beispiel: *Airquality*-Daten

- Berechnung der Vorhersageregel `air`:
  `air <- lm( formula= Ozone ~ Temp + Solar.R, data= airquality)`
- Vorhersage der Ozonwerte für `newdata` mit Hilfe des Modells `air` mit dem R-Aufruf: `predict(air,newdata)`
- Dies ergibt folgende Vorhersagen:

| Ozone | Solar.R | Temp |
|---|---|---|
| 109.4970 | 80 | 110 |
| 114.0539 | 80 | 112 |
| 118.6108 | 80 | 114 |