Methoden empirischer Sozialforschung

Lineare Modelle

Varianzanalyse

1 metrische und 1 (oder mehr) qualitative Variable(n)

- Unterscheiden sich die Mittelwerte zweier Gruppen (bei metrischen Variablen)?
  2-Stichproben-t-Test, Wilcoxon-Rangsummentest

- Unterscheiden sich die Mittelwerte mehrerer Gruppen (bei metrischen Variablen)?
  Einfache Varianzanalyse,
  Kruskal-Wallis Rang-Varianzanalyse

- Gibt es Wechselwirkungen zwischen zwei (oder mehr) unabhängigen Variablen?
  Varianzanalyse (ANOVA)

Lineare Regression

2 (oder mehr) metrische Variablen.

- Von welcher Form ist der Zusammenhang zwischen zwei metrischen Variablen?
  Läßt sich der Wert einer Variablen anhand des Wertes einer zweiten vorhersagen?
  (Einfache) Lineare Regression

- Wie läßt sich die Abhängigkeit einer metrischen Variablen von mehreren anderen beschreiben?
  Läßt sich der Wert einer Variablen anhand der Werte anderer Variablen vorhersagen?
  (Multiple) Lineare Regression

Gebrauchtwagenpreise

Beispiel: Gebrauchtwagenpreise (USA)
Erstellung einer Richtpreisliste für Gebrauchtwagen, Untersuchung von 100 Ford Taurus (3 Jahre alt).

Fragen:

- Hängt der Preis von den gefahrenen Meilen ab?

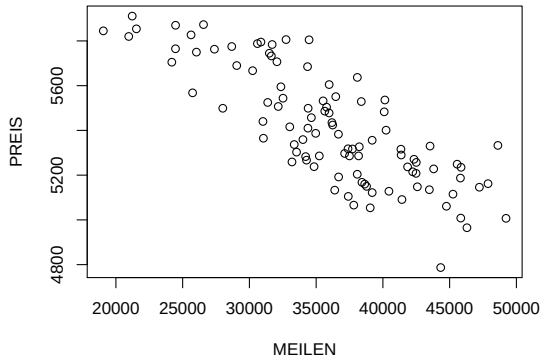
- Kann man den Gebrauchtwagenpreis aufgrund der gefahrenen Meilen vorhersagen?

Gebrauchtwagenpreise

```
R> SECONDHANDCAR <- read.table("secondhandcar.tab", header = TRUE)
R> dim(SECONDHANDCAR)

[1] 100  11

R> names(SECONDHANDCAR)

 [1] "PREIS"   "MEILEN"  "COLOR"   "I1"      "I2"      "P"       "O"
 [8] "RES.1"   "PRE.1"   "PRE.2"   "SERVICE"

R> attach(SECONDHANDCAR)
R> summary(PREIS)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4787    5225    5362    5411    5598    5911

R> summary(MEILEN)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  19060   32140   36210   36010   40290   49220
```

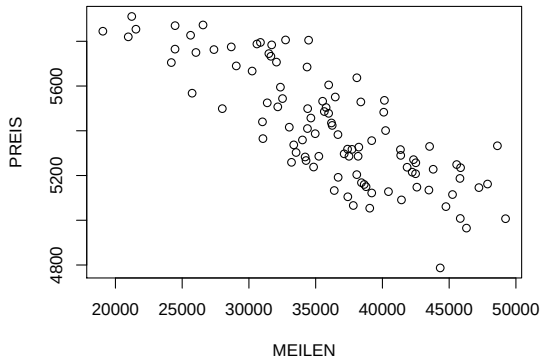
Gebrauchtwagenpreise

```
R> plot(MEILEN, PREIS)
```



Gebrauchtwagenpreise

```
R> plot(PREIS ~ MEILEN)
```



Lineare Regression

Regression: Beschreibung der Form des Zusammenhangs

Unterscheidung zur Korrelation:

- y-Variable abhängige Variable, Responsevariable
- x-Variable unabhängige Variable, erklärende Variable, Regressor

"Wenn → Dann" Beziehung:

wenn	dann
x	y
unabhängig	abhängig von x
erklärend	Response

Lineare Regression

Beispiele:

- Das Verkehrsministerium möchte die Beziehung zwischen **Straßenunebenheiten** und **Benzinverbrauch** bestimmen.
- Ein Händler, der seine Waren bei Fußballspielen verkauft, möchte die **Verkaufszahlen** auf die **Anzahl von Siegen** der Heimmannschaft beziehen.
- Ein Soziologe möchte die **Anzahl von Wochenenden**, die Studenten zu Hause verbringen, im Verhältnis zur **Entfernung** zwischen Herkunft- und Studienort untersuchen.

Die zu untersuchenden Forschungshypothesen bestimmen, welches die Response- und welche die erklärende(n) Variable(n) sind.

Lineare Regression

Unterscheidung zwischen Regression und Korrelation:

Wenn man die "Wenn → Dann" Beziehung auch umdrehen kann, dann sind beide Variablen gleichwertig, dann: *Korrelation*

Wenn man das nicht kann, dann: *Regression*

Lineare Regression

Gegeben: Datenpaare (x_i, y_i)

Gesucht: Prognosefunktion $y = g(x)$, die aus einem beobachteten Wert x eine möglichst gute Prognose y berechnet.

- $\widehat{y}_i = g(x_i)$ heißt **Schätzwert**,
- $y_i - \widehat{y}_i$ heißt **Prognosefehler**.

Was heißt *möglichst gut*? Minimiere die Quadratsumme der Prognosefehler

$$RSS = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2.$$

Lineare Regression

Wähle für $g(x) = \widehat{a} + \widehat{b}x$ eine lineare Funktion. Dann wird RSS minmal, wenn

$$\begin{aligned} \widehat{b} &= r \cdot \frac{s_y}{s_x} = \frac{s_{xy}}{s_x^2} \\ \widehat{a} &= \overline{y} - \widehat{b} \cdot \overline{x} \end{aligned}$$

Dies ist das **Prinzip der kleinsten Quadrate** (LSQ, KQ oder OLS).

$g(x) = \widehat{a} + \widehat{b}x$ heißt (empirische) Regressionsgerade und $\widehat{b}$ (empirischer) Regressionskoeffizient.

Lineare Regression

(Einfaches) lineares Regressionsmodell:

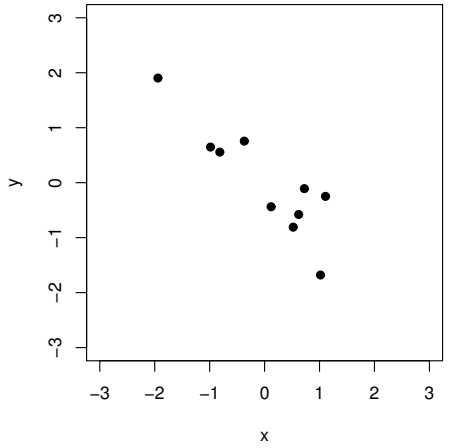
$$\begin{aligned} y_i &= a + b \cdot x_i + u_i \qquad (i = 1, \ldots, n) \\ \text{Preis} &= a + b \cdot \text{Meilen} + \text{Fehler} \end{aligned}$$

Der Prognosefehler $\widehat{u}_i = y_i - \widehat{y}_i$ heißt auch **Residuum**.

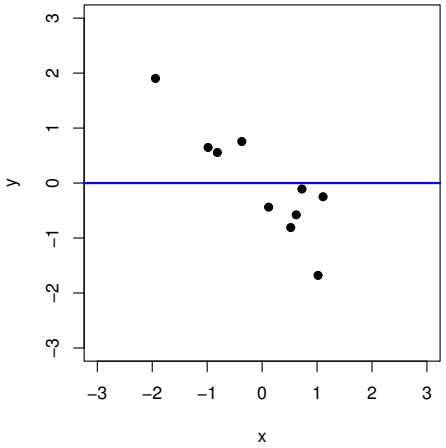
Die beobachtete Punktewolke ist (x_i, y_i) und $y_i = \widehat{a} + \widehat{b} \cdot x_i + \widehat{u}_i$.

Die Punkte auf der Geraden sind $(x_i, \widehat{y}_i)$ und $\widehat{y}_i = \widehat{a} + \widehat{b} \cdot x_i$.

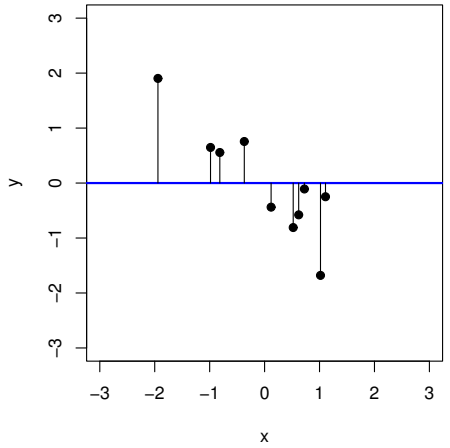
Kleinste Quadrate



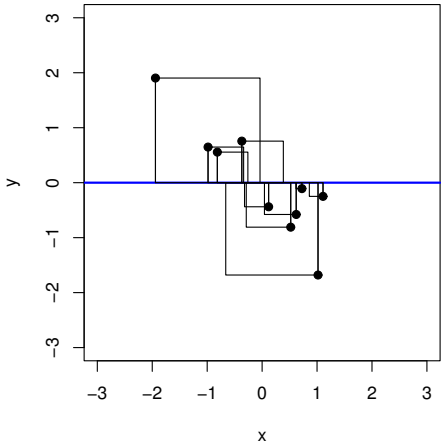
Kleinste Quadrate



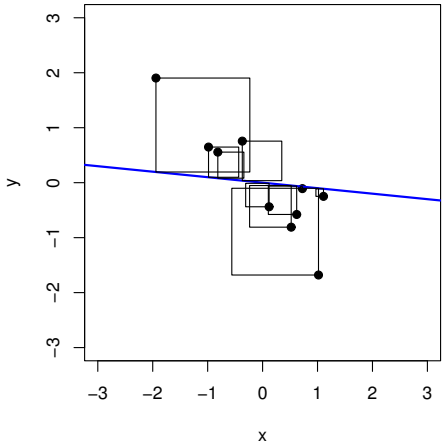
Kleinste Quadrate



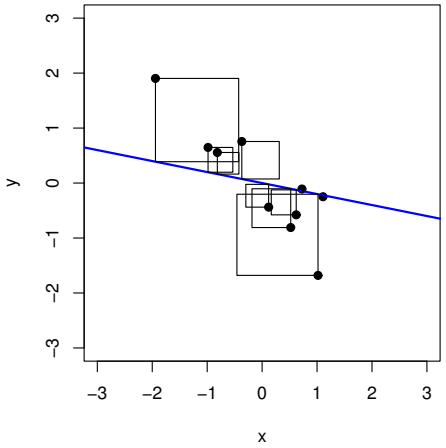
Kleinste Quadrate



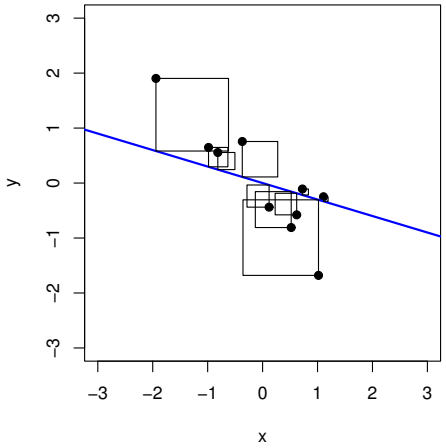
Kleinste Quadrate



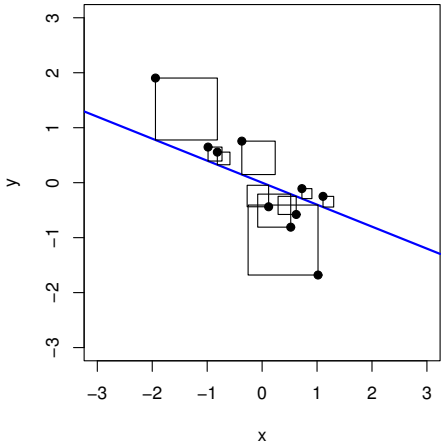
Kleinste Quadrate



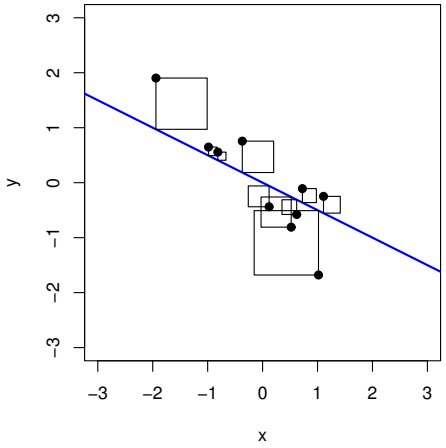
Kleinste Quadrate



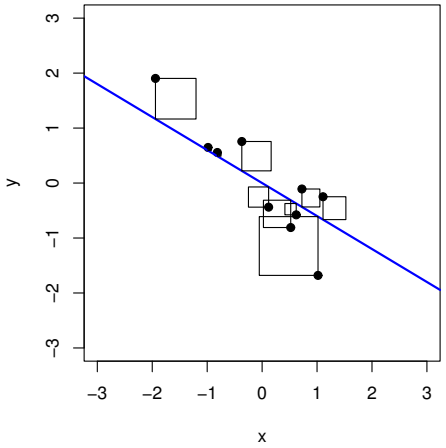
Kleinste Quadrate



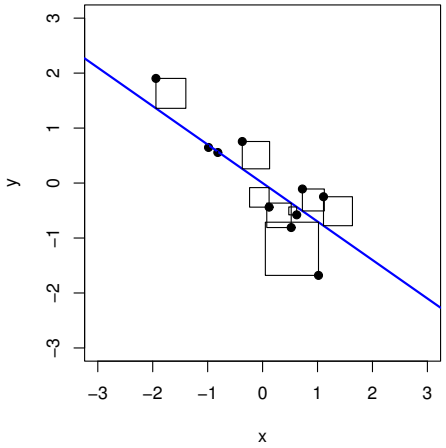
Kleinste Quadrate



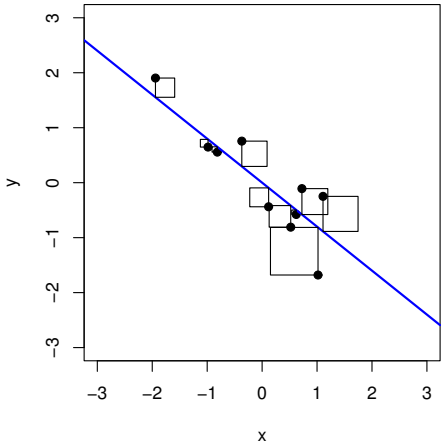
Kleinste Quadrate



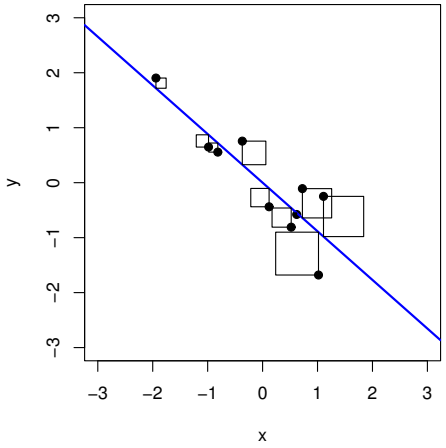
Kleinste Quadrate



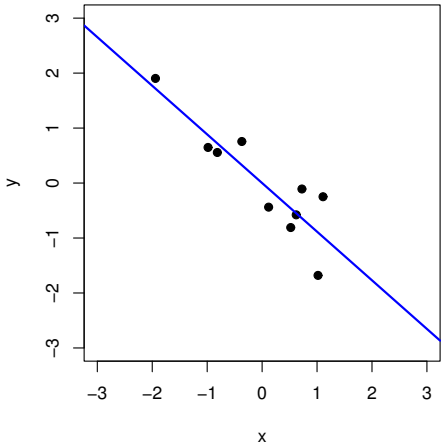
Kleinste Quadrate



Kleinste Quadrate



Kleinste Quadrate



Kleinste Quadrate

```
R> lm(PREIS ~ MEILEN)


Call:
lm(formula = PREIS ~ MEILEN)

Coefficients:
(Intercept)       MEILEN
 6533.38303     -0.03116
```

Kleinste Quadrate

Interpretation der Regressionskoeffizienten im Beispiel: Gebrauchtwagen

Empirische Regressionsgerade:

$$\text{Preis} \;=\; 6533.38 - 0.031 \cdot \text{Meilen}$$

Interpretation der Steigung b:
Der (geschätzte) Preis für einen Gebrauchtwagen sinkt um 0,031 Dollar pro gefahrener Meile, d.h. ungefähr 3 Dollar weniger für zusätzliche 100 gefahrene Meilen ($-0,031 \cdot 100$ Einheiten).

Interpretation des Achsenabschnitts a:
In vielen Beispielen (wie hier) nur hypothetisch: ein 3 Jahre alter Ford Taurus mit 0 gefahrenen Meilen würde einen Preis von 6533.38 Dollar erwarten lassen.

Kleinste Quadrate

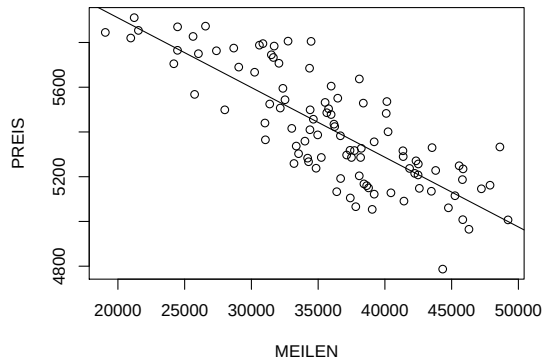
```
R> shcLM <- lm(PREIS ~ MEILEN)
R> class(shcLM)

[1] "lm"

R> names(shcLM)

 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"
```

Kleinste Quadrate

```
R> shcLM

Call:
lm(formula = PREIS ~ MEILEN)

Coefficients:
(Intercept)       MEILEN
 6533.38303     -0.03116

R> coef(shcLM)

  (Intercept)       MEILEN
6533.38303498   -0.03115774
```

Kleinste Quadrate

```
R> plot(PREIS ~ MEILEN)
R> abline(shcLM)
```



Prognose

Frage: Welchen Preis würde man für einen 3 Jahre alten Ford Taurus mit 40000 Meilen erwarten?

Antwort: Für Meilen $x = 40000$ in die Regressionsgleichung einsetzen.

$$\begin{aligned} \hat{y} &= \hat{a} - \hat{b} \cdot 40000 \\ &\approx 6533.38 - 0.031 \cdot 40000 \\ &\approx 5287.073 \end{aligned}$$

D.h. rund 5287 Dollar.

Prognose

```
R> predict(shcLM, data.frame(MEILEN = 40000))
```

```
[1] 5287.073
```

Testen im Regressionsmodell

Frage: Hat die Variable x *überhaupt* einen linearen Einfluß auf die abhängige Variable?

Anders: Ist der zugehörige Regressionskoeffizient (Steigung) von 0 verschieden?

Formal: $H : b = 0$ vs. $A : b \neq 0$

Unter der Nullhypothese:

$$y \quad = \quad a + 0 \cdot x \quad = \quad a$$

unabhängig von x.

Testen im Regressionsmodell

Lösung: Schätze die Streuung $\text{sd}(\hat{b})$ von $\hat{b}$, d.h. wie stark variiert die Schätzung von b dadurch, daß wir nur eine Stichprobe gezogen haben?

Bilde dann eine t-Teststatistik

$$T \quad = \quad \frac{\hat{b}}{\widehat{\text{sd}(\hat{b})}}$$

die unter der Nullhypothese (approximativ) t-verteilt ist.

Testen im Regressionsmodell

```
R> summary(shcLM)

Call:
lm(formula = PREIS ~ MEILEN)

Residuals:
      Min        1Q    Median        3Q       Max
-365.1605 -117.5073    0.6528   93.8729  345.6242

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.533e+03  8.451e+01   77.31   <2e-16 ***
MEILEN      -3.116e-02  2.309e-03  -13.49   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 151.6 on 98 degrees of freedom
Multiple R-Squared: 0.6501,        Adjusted R-squared: 0.6466
F-statistic: 182.1 on 1 and 98 DF,  p-value: < 2.2e-16
```

Diagnostik

Voraussetzungen für das lineare Regressionsmodell (wie bei der Korrelation):

- linearer Zusammenhang
- die Responsevariable ist metrisch
- y ist normalverteilt $\Leftrightarrow$ Residuen sind normalverteilt
- Achten auf Ausreißer!

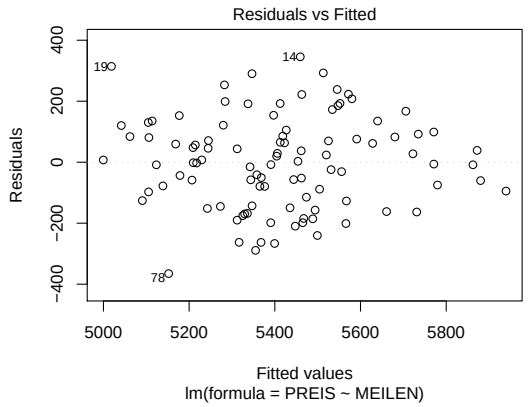
Diagnostik

Diese Voraussetzungen können grafisch geprüft werden:

- Linearität: **Residualplot** sollte kein erkennbares Muster haben. Plotte die geschätzten $\widehat{y}_i$ gegen die Residuen $\widehat{u}_i$ (oder auch x_i gegen $\widehat{u}_i$).
- Normalität: **QQ-Plot** für Residuen.
  Bei Normalverteilung sollten die Punkte nahe der Hauptdiagonale liegen.
- Ausreißer: Plots dafür wie groß der Einfluß einer Beobachtung auf die geschätzten Regressionskoeffizienten ist.

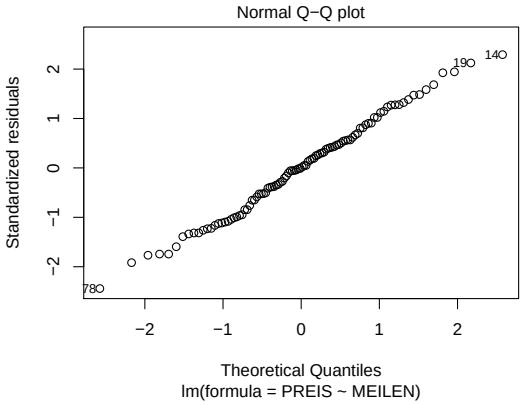
Diagnostik

```
R> plot(shcLM, which = 1)
```



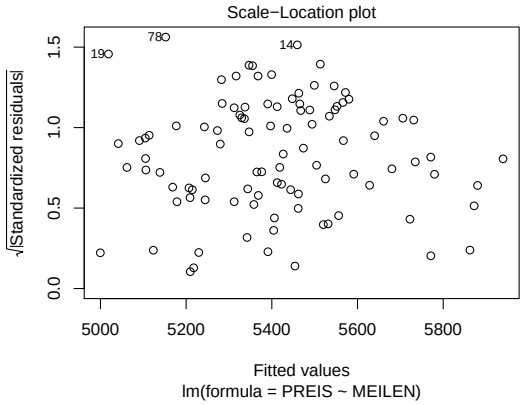
Diagnostik

```
R> plot(shcLM, which = 2)
```



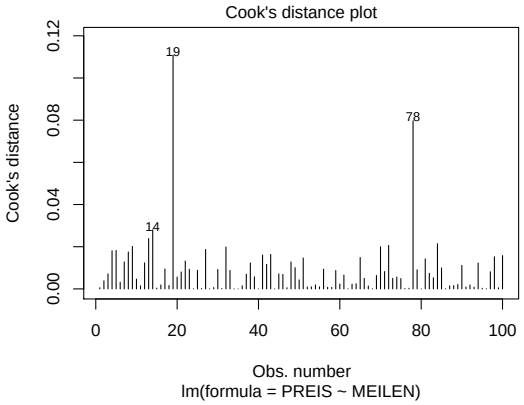
Diagnostik

```
R> plot(shcLM, which = 3)
```



Diagnostik

```
R> plot(shcLM, which = 4)
```



Diagnostik

Wie gut ist ein Regressionsmodell?

Regressionsmodell dient dazu, eine Responsevariable zu erklären bzw. vorherzusagen. Das heißt, daß man die Variation (Streuung) der abhängigen Variablen y durch die Variation einer unabhängigen Variablen x erklären möchte.

Wenn das Modell gut ist, sollte ein hoher Anteil der Varianz von y modelliert werden. Bestimmtheitsmaß:

$$R^2 = \frac{\text{VAR}(\hat{y})}{\text{VAR}(y)}$$

Diagnostik

R^2 ist also:

- der Anteil der erklärten Varianz an der Varianz von y,
- das Quadrat des Korrelationskoeffizienten (nur bei der einfachen Regression).

Diagnostik

```
R> summary(shcLM)

Call:
lm(formula = PREIS ~ MEILEN)

Residuals:
      Min        1Q    Median        3Q       Max
-365.1605 -117.5073    0.6528   93.8729  345.6242

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.533e+03  8.451e+01   77.31   <2e-16 ***
MEILEN      -3.116e-02  2.309e-03  -13.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 151.6 on 98 degrees of freedom
Multiple R-Squared: 0.6501,        Adjusted R-squared: 0.6466
F-statistic: 182.1 on 1 and 98 DF,  p-value: < 2.2e-16
```

Diagnostik

```
R> summary(shcLM)$r.squared

[1] 0.650132

R> cor(PREIS, MEILEN)

[1] -0.8063076

R> cor(PREIS, MEILEN)^2

[1] 0.650132
```

Multiple Regression

Oft lässt sich ein Modell verbessern, wenn man zusätzliche erklärende Variablen berücksichtigt.

Multiples Regressionsmodell:

$$y_i = b_0 + b_1 \cdot x_{i1} + b_2 \cdot x_{i2} + u_i$$

Beispiel: Gebrauchtwagen

Zusätzliche Variable: Anzahl Serviceüberprüfungen

$$\text{Preis} = b_0 + b_1 \cdot \text{Meilen} + b_2 \cdot \text{Service} + \text{Fehler}$$

Multiple Regression

```
R> shcLM2 <- lm(PREIS ~ MEILEN + SERVICE)
R> shcLM2

Call:
lm(formula = PREIS ~ MEILEN + SERVICE)

Coefficients:
(Intercept)       MEILEN      SERVICE
 6206.12836     -0.03146    135.83749
```

Multiple Regression

```
R> summary(shcLM2)

Call:
lm(formula = PREIS ~ MEILEN + SERVICE)

Residuals:
    Min      1Q  Median      3Q     Max
-86.079 -28.920   1.483  29.011  86.736

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.206e+03  2.497e+01  248.58   <2e-16 ***
MEILEN      -3.146e-02  6.319e-04  -49.79   <2e-16 ***
SERVICE      1.358e+02  3.903e+00   34.81   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41.48 on 97 degrees of freedom
Multiple R-Squared: 0.9741,        Adjusted R-squared: 0.9735
F-statistic:  1822 on 2 and 97 DF,  p-value: < 2.2e-16
```

Multiple Regression

Interpretation:

- beide Einflussgrößen `MEILEN` und `SERVICE` sind signifikant (p-Werte < 0.001),
- pro durchgeführtem Service erhöht sich der durchschnittliche Preis um rund 136 Dollar,
- der Koeffizient für `MEILEN` ändert sich nur leicht von -0.03116 auf -0.03146

Überprfung der Voraussetzungen wie im einfachen Regressionsmodell.

Multiple Regression

Das Bestimmtheitsmaß R^2 ist durch die zusätzliche erklärende Variable `SERVICE` von 0.650 auf 0.974 gestiegen.

Dies ist ein sehr gutes Modell, da 97.4% der Varianz der Variable `PREIS` durch die Variablen `MEILEN` und `SERVICE` erklärt werden können (maximal 100%).

Multiple Regression

Beachte:

- Durch Hinzunahme von neuen erklärenden Variablen kann das Modell nie schlechter werden. Das R^2 sinkt immer! Die Frage ist, ob es auch *signifikant* sinkt. (Hier: ja, siehe t-Tests.)

- Ein hohes R^2 deutet auf ein gutes Erklärungsmodell hin. Die Umkehrung muss nicht gelten! (Ein gutes Modell kann manchmal auch ein niedriges R^2 haben.)

Test auf Lageunterschied

Frage: Sind Lagemaße in zwei oder mehreren Gruppen unterschiedlich?

"Lagemaß" bedeutet meistens "Mittelwert".

Aber manchmal macht es keinen Sinn Mittelwerte zu verwenden, nur sinnvoll, wenn:

- Daten (in allen Gruppen) annaehernd normalverteilt, also zumindest unimodal und symmetrisch,
- die abhängige Variable metrisch ist.

Mögliche Alternative: rangbasierte Verfahren.

Test auf Lageunterschied

Bei dieser Art von Fragestellung unterscheiden wir:

- 2 Gruppen
  - Unterschiede in den Mittelwerten (Beispiel: Haushaltsarbeit von Teenagern)
    2-Stichproben-t-Test
  - Unterschiede in den Medianen (Beispiel: Telearbeit)
    Wilcoxon Rangsummentest
- 2 oder mehr Gruppen
  - Unterschiede in den Mittelwerten (Beispiel: Effektivität von Werbekampagnen)
    Einfache Varianzanalyse
  - Unterschiede in den Medianen (Beispiel: TV Konsum)
    Kruskal-Wallis-Test

Haushaltsarbeit von Teenagern

Frage: Helfen männliche und weibliche Jugendliche im Durchschnitt unterschiedlich lang im Haushalt mit?

```
R> TEENAGEWORK <- read.table("teenagework.tab", header = TRUE)
R> dim(TEENAGEWORK)

[1] 192   3

R> summary(TEENAGEWORK)

    STUNDEN            MUTTER    SEX
 Min.   : 0.000   arbeitet:122   m: 91
 1st Qu.: 4.000   daheim  : 70   w:101
 Median : 7.000
 Mean   : 7.031
 3rd Qu.: 9.000
 Max.   :19.000
```

Haushaltsarbeit von Teenagern

```
R> attach(TEENAGEWORK)
R> tapply(STUNDEN, SEX, mean)
```

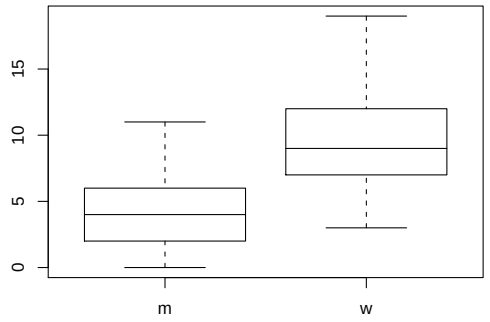
```
       m        w
4.439560 9.366337
```

```
R> tapply(STUNDEN, SEX, sd)
```

```
       m        w
2.729301 3.233335
```

Haushaltsarbeit von Teenagern

```
R> boxplot(STUNDEN ~ SEX)
```



2-Stichproben-t-Test

Frage: Gibt es einen Unterschied in den Mittelwerten von 2 Gruppen (bei einer metrischen Variable)?

Formal: Zwei metrische Variablen y_1 und y_2, die annähernd normalverteilt sind mit Mittelwerten μ_1 und μ_2. Teste die Nullhypothese:

$$H: \mu_1 = \mu_2$$

gegen die Alternative

$$A: \mu_1 \neq \mu_2$$

oder stattdessen gegen eine der einseitigen Alternativen $\mu_1 \geq \mu_2$ oder $\mu_1 \leq \mu_2$.

2-Stichproben-t-Test

Lösung: Überprüfe ob die Differenz der empirischen Mittelwerte $\bar{y}_1 - \bar{y}_2$ "nahe" bei 0 liegt.

Berücksichtige aber die zufällige Variation der Mittelwertdifferenz durch Schätzung der Streuung $\text{sd}(\bar{y}_1 - \bar{y}_2)$.

Bilde dann eine t-Teststatistik

$$T = \frac{\bar{y}_1 - \bar{y}_2}{\widehat{\text{sd}(\bar{y}_1 - \bar{y}_2)}}$$

die unter der Nullhypothese (approximativ) t-verteilt ist.

Verschiedene Schätzungen für $\text{sd}(\bar{y}_1 - \bar{y}_2)$ können verwendet werden, je nachdem ob man annimmt, daß die Varianzen in beiden Stichproben gleich sind ($\text{sd}(y_1) = \text{sd}(y_2)$) oder nicht.

2-Stichproben-t-Test

```
R> t.test(STUNDEN ~ SEX)

        Welch Two Sample t-test

data:  STUNDEN by SEX
t = -11.4432, df = 189.219, p-value = < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.776057 -4.077495
sample estimates:
mean in group m mean in group w
       4.439560        9.366337
```

2-Stichproben-t-Test

```
R> t.test(STUNDEN ~ SEX, var.equal = TRUE)

        Two Sample t-test

data:  STUNDEN by SEX
t = -11.343, df = 190, p-value = < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.783532 -4.070021
sample estimates:
mean in group m mean in group w
       4.439560        9.366337
```

2-Stichproben-t-Test

Wie interpretiert man statistische Ergebnisse?
(am Beispiel des 2-Stichproben-t-Tests)

1. Technische (statistische) Interpretation

- Angabe der Methode
- Null- und Alternativhypothese
- Signifikanzniveau

2. Inhaltliche Interpretation

- Bezugnehmen auf die Fragestellung
- Angabe von Statistiken
- p-Wert

2-Stichproben-t-Test

Beispiel:

Berechnet wird ein t-Test. Es zeigt sich, daß die Nullhypothese "die durchschnittliche Anzahl der im Haushalt gearbeiteten Stunden pro Woche ist bei männlichen und weiblichen Teenagern gleich" zugunsten der Alternativhypothese "die Durchschnittszeit von Haus- haltsarbeit ist bei Burschen und Mädchen ungleich" zum 5%-Niveau verworfen werden muss.

Mädchen helfen demnach mit 9.36 Stunden im Durchschnitt signifikant länger im Haushalt mit als Burschen mit 4.43 Stunden ($p < 0.001$).

Wilcoxon Rangsummentest

Der 2-Stichproben-t-Test funktioniert dann besonders gut, wenn die Daten etwa symmetrisch und eingipflig und etwa normalverteilt sind.

Sind die Daten zwar symmetrisch und eingipflig, aber nicht besonders normalverteilt – etwa weil es Ausreißer gibt – dann gibt es – wie auch im 1-Stichprobenfall – einen anderen Test, den Wilcoxon Rangsummentest.

Idee: Betrachte die Ränge der Beobachtungen in der vereinigten Stichprobe.

Wilcoxon Rangsummentest

```
R> wilcox.test(STUNDEN ~ SEX)

        Wilcoxon rank sum test with continuity correction

data:  STUNDEN by SEX
W = 1110.5, p-value = < 2.2e-16
alternative hypothesis: true mu is not equal to 0
```

Telearbeit

Beispiel: Telearbeit und Arbeitszufriedenheit

In den USA arbeiten schon mehr als 10% der Vollbeschäftigten zuhause am Computer (Telearbeiter), d.h. sie sind mittels Modem mit einem Firmenrechner verbunden. In einer Studie sollte untersucht werden, ob diese Beschäftigten mit ihrer Arbeit zufrieden sind.

Frage: Unterscheiden sich Telearbeiter von Personen, die in einem Büro arbeiten bezüglich ihrer Arbeitszufriedenheit?
(1 = sehr unzufrieden, ... ,50 = sehr zufrieden)

Telearbeit

```
R> TELEARBEIT <- read.table("comphomeneu.tab", header = TRUE)
R> dim(TELEARBEIT)

[1] 200   2

R> names(TELEARBEIT)

[1] "ZUFRIEDENHEIT" "ORT"

R> attach(TELEARBEIT)
```

Telearbeit

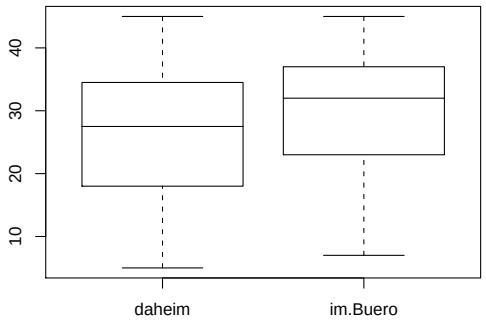
```
R> summary(TELEARBEIT)

 ZUFRIEDENHEIT         ORT
 Min.   : 5.00   daheim  :100
 1st Qu.:20.00   im.Buero:100
 Median :29.50
 Mean   :28.41
 3rd Qu.:36.25
 Max.   :45.00

R> table(ZUFRIEDENHEIT)

ZUFRIEDENHEIT
 5  6  7  8  9 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
 1  1  2  4  2  1  3  1  6  4  6  4  8  2  6  1  7  6  6  5 10  4  2  8  3  3
32 33 34 35 36 37 38 39 40 41 42 43 44 45
14 12  8  4  6  7  4  7  5  4  8  4  6  5
```

Telearbeit

```
R> tapply(ZUFRIEDENHEIT, ORT, median)

  daheim im.Buero
   27.5     32.0

R> tapply(ZUFRIEDENHEIT, ORT, IQR)

  daheim im.Buero
  16.25    14.00
```

Telearbeit

```
R> boxplot(ZUFRIEDENHEIT ~ ORT)
```



Telearbeit

```
R> wilcox.test(ZUFRIEDENHEIT ~ ORT)

        Wilcoxon rank sum test with continuity correction

data:  ZUFRIEDENHEIT by ORT
W = 4319, p-value = 0.09614
alternative hypothesis: true mu is not equal to 0
```

Telearbeit

Resultat:

Es besteht kein signifikanter Unterschied: Personen, die in einem Büro arbeiten, unterscheiden sich nicht in ihrer Arbeitszufriedenheit von solchen, die zu Hause arbeiten.

Der Median der Zufriedenheit liegt zwar in der Telearbeitergruppe mit 27.5 leicht unter dem der Büroarbeiter mit 32, jedoch kann der Wilcoxon Rangsummentest einen Unterschied in der mittleren Zufriedenheit zum 5%-Niveau nicht nachweisen ($p = 0.096$).

Varianzanalyse

Der 2-Stichproben-t-Test auf Lageunterschiede (bei gleichen Varianzen) kann auch schematisch als Varianzanalyse durchgeführt werden.

Dafür kann man das Modell für Lageunterschiede auch als lineares Modell formulieren:

$$\begin{aligned} y_i &= a + b \cdot x_i + u_i \qquad (i = 1, \ldots, n) \\ \text{Stunden} &= a + b \cdot \text{Geschlecht} + \text{Fehler} \end{aligned}$$

Die Variable "Geschlecht" wird dabei als Dummy- oder 0/1-Variable definiert: Ist die untersuchte Person weiblich? 0 = nein, 1 = ja.

Varianzanalyse

Der Mittelwert bei den Jungen ist damit $a + b \cdot 0 = a$ und bei den Mädels $a + b \cdot 1 = a + b$.

Ein Test auf Gleichheit der Mittelwerte reduziert sich dann also auf einen Test ob $b = 0$.

Varianzanalyse

Dieses Modell kennen wir aber schon: einfache lineare Regression, wobei der Regressor nur die Werte 0 und 1 annimmt.

Der Test des Regressionskoeffizienten b kann dann wie vorher durchgeführt werden.

In R muss die 0/1-Variable nicht selbst ausgerechnet werden: `lm` weiß, was zu tun ist, wenn ein Faktor als erklärende Variable übergeben wird.

Varianzanalyse

```
R> teleLM1 <- lm(STUNDEN ~ SEX)
R> summary(teleLM1)

Call:
lm(formula = STUNDEN ~ SEX)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3663 -2.3663 -0.4029  1.6337  9.6337

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.4396     0.3150   14.09   <2e-16 ***
SEXw          4.9268     0.4343   11.34   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.005 on 190 degrees of freedom
Multiple R-Squared: 0.4038,        Adjusted R-squared: 0.4006
F-statistic: 128.7 on 1 and 190 DF,  p-value: < 2.2e-16
```

Varianzanalyse

```
R> t.test(STUNDEN ~ SEX, var.equal = TRUE)


        Two Sample t-test

data:  STUNDEN by SEX
t = -11.343, df = 190, p-value = < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.783532 -4.070021
sample estimates:
mean in group m mean in group w
       4.439560        9.366337
```

Varianzanalyse

Wir erinnern uns: zur Schätzung der Koeffizienten eines Regressionmodells wird die Fehlerquadratsumme (RSS) minimiert.

Durch Hinzunahme weiterer Variablen kann diese Fehlerquadratsumme nur sinken. Deshalb ist die Fehlerquadratsumme RSS_2 eines *komplexeren* Modells (mit $p+q$ Koeffizienten) immer kleiner als die eines *einfachen* Modells (mit p Koeffizienten).

Frage: "Lohnt" sich das komplexere Modell?
D.h. ist RSS_2 signifikant kleiner als RSS_1?

Varianzanalyse

Antwort: Überpfüfung anhand der Teststatistik

$$\frac{(RSS_1 - RSS_2)/q}{RSS_2/(p+q)}$$

Diese ist (approximativ) F-verteilt mit q und $p+q$ Freiheitsgraden.

Hier ist unser Referenzmodell (mit RSS_1) das Modell, das unterstellt, daß der Mittelwert in beiden Gruppen gleich ist.

Varianzanalyse

```
R> anova(teleLM1)

Analysis of Variance Table

Response: STUNDEN
           Df  Sum Sq Mean Sq F value    Pr(>F)
SEX         1 1161.95 1161.95  128.66 < 2.2e-16 ***
Residuals 190 1715.86    9.03
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R> t.test(STUNDEN ~ SEX, var.equal = TRUE)$statistic^2

       t
128.6643
```

Varianzanalyse

```
R> summary(teleLM1)

Call:
lm(formula = STUNDEN ~ SEX)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3663 -2.3663 -0.4029  1.6337  9.6337

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.4396     0.3150   14.09   <2e-16 ***
SEXw          4.9268     0.4343   11.34   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.005 on 190 degrees of freedom
Multiple R-Squared: 0.4038,        Adjusted R-squared: 0.4006
F-statistic: 128.7 on 1 and 190 DF,  p-value: < 2.2e-16
```

Varianzanalyse

```
R> teleLM0 <- lm(STUNDEN ~ 1)
R> anova(teleLM0, teleLM1)

Analysis of Variance Table

Model 1: STUNDEN ~ 1
Model 2: STUNDEN ~ SEX
  Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
1    191 2877.8
2    190 1715.9   1    1162.0 128.66 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Varianzanalyse

```
R> anova(teleLM1)

Analysis of Variance Table

Response: STUNDEN
           Df  Sum Sq Mean Sq F value    Pr(>F)
SEX         1 1161.95 1161.95  128.66 < 2.2e-16 ***
Residuals 190 1715.86    9.03
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Varianzanalyse

Dieselben Ideen und Methoden lassen sich auch für den Vergleich anderer Modelle anwenden.

Auch für den Mittelwertvergleich zwischen mehr als nur 2 Gruppen wie im folgenden Beispiel.

Apfelsaftkonzentrat

Fruchtsafthersteller bringt neues Produkt, Apfelsaftkonzentrat, auf den Markt, bei dem aus kleiner Menge mit Wasser 1l Apfelsaft hergestellt werden kann:

Dies bietet 3 Vorteile:

- Bequemlichkeit: leicht, wenig Platz im Kühlschrank
- Qualität: wird aus echten pfeln hergestellt
- Preis: niedriger als herkömmlicher Saft

In 3 "Versuchsstädten" wird Produkt beworben und jeweils nur ein Aspekt in der Werbung betont.

Frage: Unterscheiden sich die durchschnittlichen Verkaufszahlen von Apfelsaftkonzentrat je nach Werbeinhalt?

Apfelsaftkonzentrat

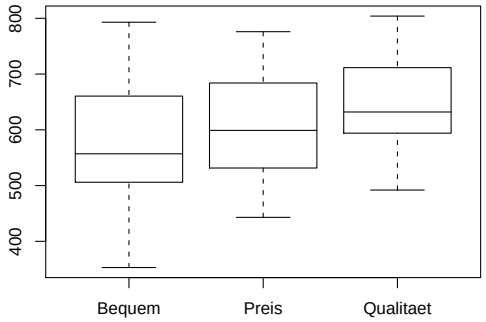
```
R> APPLE <- read.table("apple.tab", header = TRUE)
R> dim(APPLE)

[1] 60  2

R> summary(APPLE)

    VERKAUF            INHALT
 Min.   :353.0   Bequem   :20
 1st Qu.:531.8   Preis    :20
 Median :610.0   Qualitaet:20
 Mean   :613.1
 3rd Qu.:689.5
 Max.   :804.0

R> attach(APPLE)
```

Apfelsaftkonzentrat

```
R> tapply(VERKAUF, INHALT, mean)

   Bequem     Preis Qualitaet
   577.55    608.65    653.00

R> tapply(VERKAUF, INHALT, median)

   Bequem     Preis Qualitaet
      557       599       632

R> tapply(VERKAUF, INHALT, sd)

   Bequem     Preis Qualitaet
103.80268  93.11412  85.07705
```

Apfelsaftkonzentrat

```
R> boxplot(VERKAUF ~ INHALT)
```



Apfelsaftkonzentrat

```
R> appleLM <- lm(VERKAUF ~ INHALT)
R> anova(appleLM)

Analysis of Variance Table

Response: VERKAUF
          Df Sum Sq Mean Sq F value  Pr(>F)
INHALT     2  57512   28756   3.233 0.04677 *
Residuals 57 506983    8894
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Apfelsaftkonzentrat

Resultat:

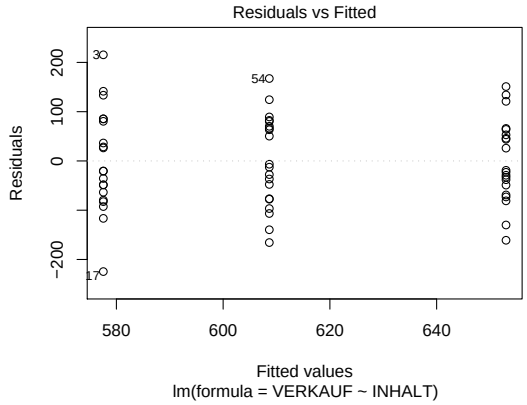
Es besteht ein signifikanter Unterschied in der Wirksamkeit der drei Werbeinhalte ($p = 0.047$).

Die Betonung der Qualitt des Produkts führt vergleichsweise zu den höchsten Verkaufszahlen

Anmerkung: Auch hier sollten die Annahmen des Verfahrens zumindest grafisch überprüft werden.

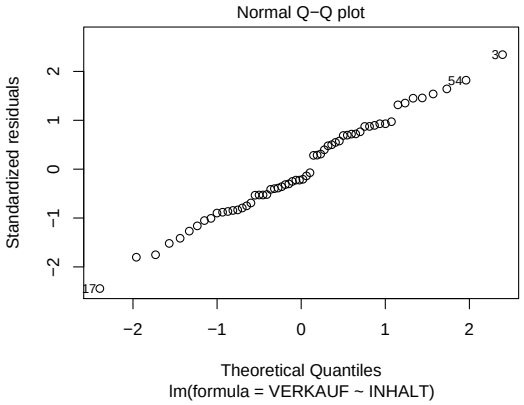
Apfelsaftkonzentrat

```
R> plot(appleLM, which = 1)
```



Apfelsaftkonzentrat

```
R> plot(appleLM, which = 2)
```



Kruskal-Wallis-Test

Als rangbasierte Alternative zur einfachen Varianzanalyse (ANOVA) gibt es den rangbasierten Kruskal-Wallis-Test.

```
R> kruskal.test(VERKAUF ~ INHALT)

        Kruskal-Wallis rank sum test

data:  VERKAUF by INHALT
Kruskal-Wallis chi-squared = 6.075, df = 2, p-value = 0.04796
```

Fernsehkonsum

Beispiel: Fernsehkonsum von Kindern

In einer amerikanischen Studie wurde das TV-Verhalten von Kindern untersucht. Erfaßt wurde unter anderem die durchschnittliche tägliche Dauer des Fernsehens (nach *Television in the Home 1998. Annenberg Public Policy Center*)

Frage: Gibt es einen Unterschied in der Dauer des Fersehkonsums von Kindern unterschiedlichen Alters?

Fernsehkonsum

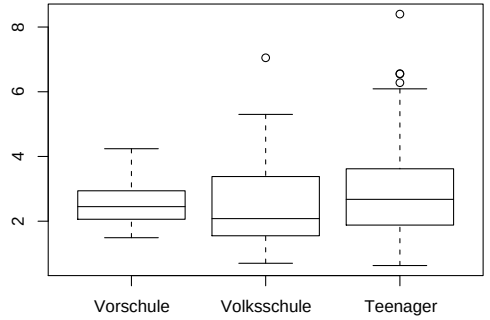
```
R> TV <- read.table("tv.tab", header = TRUE)
R> dim(TV)

[1] 150   2

R> summary(TV)

    STUNDEN             GRUPPE
 Min.   :0.630   Teenager   :50
 1st Qu.:1.860   Volksschule:50
 Median :2.410   Vorschule  :50
 Mean   :2.717
 3rd Qu.:3.345
 Max.   :8.400

R> TV$GRUPPE <- factor(TV$GRUPPE, levels = c("Vorschule", "Volksschule",
+     "Teenager"), ordered = TRUE)
R> attach(TV)
```

Fernsehkonsum

```
R> tapply(STUNDEN, GRUPPE, mean)

  Vorschule Volksschule    Teenager
     2.6054      2.5526      2.9942

R> tapply(STUNDEN, GRUPPE, median)

  Vorschule Volksschule    Teenager
      2.450       2.080       2.675

R> tapply(STUNDEN, GRUPPE, sd)

  Vorschule Volksschule    Teenager
   0.708526    1.349131    1.711256

R> tapply(STUNDEN, GRUPPE, IQR)

  Vorschule Volksschule    Teenager
      0.825       1.810       1.675
```

Fernsehkonsum

```
R> boxplot(STUNDEN ~ GRUPPE)
```



Fernsehkonsum

```
R> kruskal.test(STUNDEN ~ GRUPPE)

        Kruskal-Wallis rank sum test

data:  STUNDEN by GRUPPE
Kruskal-Wallis chi-squared = 2.3113, df = 2, p-value = 0.3148
```

Es kann kein Unterschied im mittleren TV-Konsum der drei Altersgruppen (Vorschule, Volksschule, Teenager) nachgewiesen werden ($p = 0.315$).

Haushaltsarbeit von Teenagern

In dem Beispiel zur Haushaltsarbeit von Teenagern haben wir uns bisher nur die Variablen `STUNDEN` und `SEX` untersucht. Es gibt aber noch die dritte Variable `MUTTER`.

```
R> summary(TEENAGEWORK)

    STUNDEN           MUTTER    SEX
 Min.   : 0.000   arbeitet:122   m: 91
 1st Qu.: 4.000   daheim  : 70   w:101
 Median : 7.000
 Mean   : 7.031
 3rd Qu.: 9.000
 Max.   :19.000
```

Haushaltsarbeit von Teenagern

```
R> tapply(STUNDEN, SEX, mean)

       m        w
4.439560 9.366337

R> tapply(STUNDEN, MUTTER, mean)

arbeitet   daheim
6.852459 7.342857

R> tapply(STUNDEN, list(SEX, MUTTER), mean)

   arbeitet   daheim
m  3.051724 6.878788
w 10.296875 7.756757
```

Haushaltsarbeit von Teenagern

Mit bisherigen Methoden lassen sich zwei Fragen beantworten:

- Unterscheiden sich Burschen von Mädchen bzgl. der Mitarbeit im Haushalt?
- Gibt es einen Unterschied in der geleisteten Haushaltsarbeit von Teenagern zwischen Familien, in denen die Mutter berufstätig ist und solchen, in denen Mutter zu Hause ist?

Diese nennt man *Haupteffekte*.

Hier wird jeweils nach einem einzelnen Unterschied in der abhängigen Variable (Zeit für Haushaltsarbeit) gefragt, unabhängig von der jeweils anderen Variablen. Entspricht marginaler Information bei kategorialen Daten.

Haushaltsarbeit von Teenagern

Eine andere Frage ist:

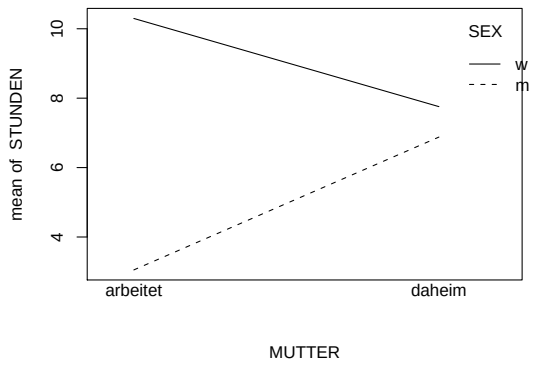
- Ist der Unterschied zwischen männlich und weiblich anders, je nachdem ob die Mutter zu Hause ist oder berufstätig ist?

Dieses nennt man einen *Wechselwirkungseffekt* oder *Interaktion*.

Effekte, bei denen mehr als eine unabhängige Variable involviert ist, nennt man Wechselwirkung – entspricht der gemeinsamen Information bei kategorialen Daten.

Haushaltsarbeit von Teenagern

```
R> interaction.plot(MUTTER, SEX, STUNDEN)
```



Haushaltsarbeit von Teenagern

Auch Wechselwirkungsmodelle können als lineare Modelle formuliert werden.

Berechnung wieder mit `lm`.

Testen wieder mit `anova`.

Das einfachste denkbare Modell ist das, wo für alle Beobachtungen derselben Mittelwert unterstellt wird.

```
R> fm0 <- lm(STUNDEN ~ 1)
```

Haushaltsarbeit von Teenagern

Wir haben nun schon gesehen, daß die mittlere Haushaltsarbeitszeit für Jungs und Mädels unterschiedlich ist. Dies entspricht dem Modell mit dem Haupteffekt `SEX`:

```
R> fm1 <- lm(STUNDEN ~ SEX)
R> anova(fm0, fm1)

Analysis of Variance Table

Model 1: STUNDEN ~ 1
Model 2: STUNDEN ~ SEX
  Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
1    191 2877.8
2    190 1715.9   1    1162.0 128.66 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R> t.test(STUNDEN ~ SEX, var.equal = TRUE)$statistic^2

       t
128.6643
```

Haushaltsarbeit von Teenagern

Wenn zusätzlich noch der Haupteffekt für die Variable `MUTTER` in das Modell mitaufgenommen werden soll:

```
R> fm2 <- lm(STUNDEN ~ SEX + MUTTER)
```

Aber eigentlich sind wir interessiert an der Frage, ob es eine Wechselwirkung zwischen `MUTTER` und `SEX` gibt. Folgende Spezifikationen sind in R äquivalent:

```
R> fm3 <- lm(STUNDEN ~ SEX + MUTTER + SEX:MUTTER)
R> fm3 <- lm(STUNDEN ~ SEX * MUTTER)
```

Haushaltsarbeit von Teenagern

Welcher Test durchgeführt werden muss, hängt von der Fragestellung ab.

Soll die Hypothese getestet werden, daß die Mittelwerte für alle Kombinationen von `SEX` und `MUTTER` gleich sind, gegen die Alternative, daß sie alle unterschiedlich sind, dann:

```
R> anova(fm0, fm3)

Analysis of Variance Table

Model 1: STUNDEN ~ 1
Model 2: STUNDEN ~ SEX * MUTTER
  Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
1    191 2877.8
2    188 1256.5   3    1621.3 80.858 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Haushaltsarbeit von Teenagern

Geht man hingegen aus irgendwelchen Gründen davon aus, daß beide Haupteffekte relevant sind, kann man fragen:

Besteht zusätzlich zu den Haupteffekten auch noch ein Wechselwirkungseffekt?

```
R> anova(fm2, fm3)

Analysis of Variance Table

Model 1: STUNDEN ~ SEX + MUTTER
Model 2: STUNDEN ~ SEX * MUTTER
  Res.Df     RSS  Df Sum of Sq     F    Pr(>F)
1    189 1706.00
2    188 1256.53   1    449.47 67.25 3.663e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Haushaltsarbeit von Teenagern

Um ein möglichst gutes Erklärungsmodell für die Daten zu finden, kann man auch einfach alle Modelle miteinander vergleichen und das “beste” nehmen.

Dabei vergleicht man üblicherweise “geschachtelte” (“genestete”) Modelle, also solche die inkrementell einfacher oder komplexer werden.

Haushaltsarbeit von Teenagern

```
R> anova(fm0, fm1, fm2, fm3)

Analysis of Variance Table

Model 1: STUNDEN ~ 1
Model 2: STUNDEN ~ SEX
Model 3: STUNDEN ~ SEX + MUTTER
Model 4: STUNDEN ~ SEX * MUTTER
  Res.Df     RSS  Df Sum of Sq       F    Pr(>F)
1    191 2877.81
2    190 1715.86   1   1161.95 173.849 < 2.2e-16 ***
3    189 1706.00   1      9.86   1.475    0.2261
4    188 1256.53   1    449.47  67.250 3.663e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Haushaltsarbeit von Teenagern

```
R> anova(fm3)

Analysis of Variance Table

Response: STUNDEN
           Df  Sum Sq Mean Sq F value    Pr(>F)
SEX         1 1161.95 1161.95 173.849 < 2.2e-16 ***
MUTTER      1    9.86    9.86   1.475    0.2261
SEX:MUTTER  1  449.47  449.47  67.250 3.663e-14 ***
Residuals 188 1256.53    6.68
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Haushaltsarbeit von Teenagern

Problem: Wenn man mehrere Tests durchführt, sind die p-Werte nicht mehr genauso interpretierbar.

Bei *einem* Test:
Die Fehlerwahrscheinlichkeit wird mit dem Signifikanzniveau (üblicherweise 5%) kontrolliert. Wenn die Nullhypothese verworfen wird, da $p \leq 0.05$, beträgt die Irrtumswahrscheinlichkeit höchstens p.

Bei *mehreren* Tests:
Wenn ich drei Tests durchführe und mich jedesmal mit 5% Wahrscheinlichkeit irre, dann beträgt die Wahrscheinlichkeit, dass ich *keinmal* falsch liege:

$$(1 - 0.05) \cdot (1 - 0.05) \cdot (1 - 0.05) \quad \approx \quad 0.857$$

d.h. die Irrtumswahrscheinlichkeit ist etwa 14.3% und nicht 5%!

Haushaltsarbeit von Teenagern

Ausweg: Informationskriterien

Idee: Wenn ich ein einfaches und ein komplexes Modell habe, das die Daten (etwa) gleichgut erklären, ist immer das einfache zu bevorzugen.

Deshalb haben Informationskriterien folgende Form:

$$IC \quad = \quad RSS + \text{Strafterm},$$

wobei:

- die Fehlerquadratsumme RSS **sinkt** je mehr Koeffizienten in das Modell aufgenommen werden,
- der Strafterm **steigt** je mehr Koeffizienten in das Modell aufgenommen werden.

Haushaltsarbeit von Teenagern

Das "beste" Modell bzgl. eines Informationskriteriums IC ist das Modell, welches das IC minimiert, d.h. wo der Trade-Off zwischen zusätzlicher Anzahl an Parametern und Reduktion der RSS am besten ist.

Das bekannteste ist das Akaike Informationskriterium (AIC).

Haushaltsarbeit von Teenagern

Die RSS kann aus einem `lm`-Objekt mit `deviance` extrahiert werden. Sie sinkt, je mehr Parameter in das Modell aufgenommen werden:

```
R> deviance(fm0)

[1] 2877.812

R> deviance(fm1)

[1] 1715.863

R> deviance(fm2)

[1] 1706.005

R> deviance(fm3)

[1] 1256.530
```

Haushaltsarbeit von Teenagern

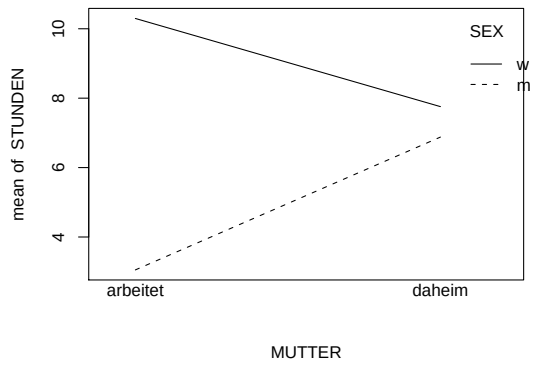
Das AIC wird allerdings beim komplexesten Modell am kleinsten, es erklärt die Daten offenbar am besten:

```
R> AIC(fm0, fm1, fm2, fm3)
```

```
    df       AIC
fm0  2 1068.6721
fm1  3  971.3862
fm2  4  972.2799
fm3  5  915.5663
```

Haushaltsarbeit von Teenagern

```
R> interaction.plot(MUTTER, SEX, STUNDEN)
```



Haushaltsarbeit von Teenagern

Resultat:

Alle diese Methoden liefern das Ergebnis, was am einfachsten aus dem Interaktionsplot ablesbar war:

Es gibt einen signifikanten Interaktionseffekt ($p < 0.001$): Mädchen und Jungen helfen nahezu gleich lang im Haushalt mit wenn die Mutter zuhause ist (durchschnittlich 7.8 bzw 6.9 Stunden pro Woche). Wenn die Mutter aber berufstätig ist, arbeiten Mädchen erheblich länger im Haushalt mit (10.3 Stunden) als Jungen (3.1 Stunden).

Demnach unterscheiden sich Burschen von Mädchen bezüglich der Mithilfe im Haushalt vor allem dann, wenn die Mutter einer Beschäftigung nachgeht.