Methoden empirischer Sozialforschung

Lineare Modelle (2. Teil)

Kovarianzanalyse

1 metrische und mehrere metrische und kategoriale Variablen

- Wie läßt sich die Abhängigkeit einer metrischen Variablen von mehreren metrischen und kategorialen Variablen beschreiben?
  Läßt sich der Wert einer Variablen anhand der Werte anderer (metrischer und kategorialer) Variablen vorhersagen?
  Kovarianzanalyse (ANCOVA)

Truthahngewicht

Bei der Züchtung von Truthähnen aus zwei verschiedenen Regionen wird das Gewicht und das Alter der Tiere erhoben. Mögliche Fragen:

1. Hängt das Gewicht der Tiere vom Alter ab?
2. Sind Truthähne aus unterschiedlichen Region im Durchschnitt unterschiedlich schwer?
3. Ist der Zusammenhang zwischen Gewicht und Alter für Tiere aus beiden Regionen unterschiedlich?

Truthahngewicht

```
R> TRUTHAHN <- read.csv2(file = "truthahn.csv")
R> TRUTHAHN

   Gewicht Alter  Region
1     11.5    21 Region1
2     14.6    27 Region1
3     15.4    29 Region1
4     13.1    23 Region1
5     13.8    25 Region1
6     13.6    24 Region1
7     14.3    28 Region2
8      9.9    20 Region2
9     16.1    32 Region2
10    11.4    22 Region2
11    13.7    26 Region2
```

Truthahngewicht

```
R> dim(TRUTHAHN)

[1] 11  3

R> summary(TRUTHAHN)

    Gewicht          Alter          Region
 Min.   : 9.90   Min.   :20.00   Region1:6
 1st Qu.:12.30   1st Qu.:22.50   Region2:5
 Median :13.70   Median :25.00
 Mean   :13.40   Mean   :25.18
 3rd Qu.:14.45   3rd Qu.:27.50
 Max.   :16.10   Max.   :32.00

R> attach(TRUTHAHN)
```

Truthahngewicht

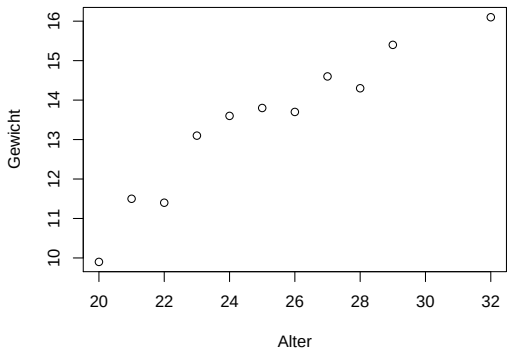
Frage: 1. Hängt das Gewicht der Tiere vom Alter ab?

Diese Frage können wir bereits beantworten: entweder durch Analyse der Korrelation oder durch einfache lineare Regression.

Da die Variable `Gewicht` von größerem Interesse ist als die Variable `Alter`, sehen wir hier erstere als abhängige und letztere als erklärende Variable an.

Truthahngewicht

```
R> plot(Gewicht ~ Alter)
```



Truthahngewicht

```
R> cor(Gewicht, Alter)

[1] 0.9495734

R> fm1 <- lm(Gewicht ~ Alter)
R> fm1

Call:
lm(formula = Gewicht ~ Alter)

Coefficients:
(Intercept)        Alter
     1.3778       0.4774
```

Truthahngewicht

```
R> summary(fm1)

Call:
lm(formula = Gewicht ~ Alter)

Residuals:
     Min       1Q   Median       3Q      Max
-1.02612 -0.46320  0.09646  0.40939  0.76422

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.37782    1.33584   1.031    0.329
Alter        0.47741    0.05255   9.086  7.9e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6074 on 9 degrees of freedom
Multiple R-Squared: 0.9017,        Adjusted R-squared: 0.8908
F-statistic: 82.55 on 1 and 9 DF,  p-value: 7.902e-06
```

Truthahngewicht

```
R> anova(fm1)

Analysis of Variance Table

Response: Gewicht
          Df  Sum Sq Mean Sq F value    Pr(>F)
Alter      1 30.4591 30.4591  82.547 7.902e-06 ***
Residuals  9  3.3209  0.3690
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

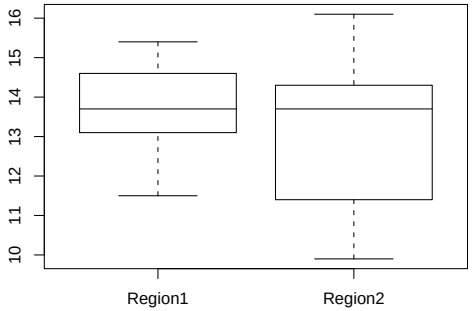
Truthahngewicht

Frage: 2. Sind Truthähne aus unterschiedlichen Region im Durchschnitt unterschiedlich schwer?

Auch diese Frage können wir bereits beantworten: entweder durch einen 2-Stichprobentest (t-Test, Wilcoxon-Rangsummentest) oder durch eine ANOVA.

Truthahngewicht

```
R> boxplot(Gewicht ~ Region)
```



Truthahngewicht

```
R> t.test(Gewicht ~ Region)

        Welch Two Sample t-test

data:  Gewicht by Region
t = 0.4801, df = 5.94, p-value = 0.6484
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.410986  3.584319
sample estimates:
mean in group Region1 mean in group Region2
             13.66667              13.08000
```

Truthahngewicht

```
R> fm2 <- lm(Gewicht ~ Region)
R> anova(fm2)

Analysis of Variance Table

Response: Gewicht
          Df Sum Sq Mean Sq F value Pr(>F)
Region     1  0.939   0.939  0.2572 0.6242
Residuals  9 32.841   3.649
```

Truthahngewicht

Frage: 3. Ist der Zusammenhang zwischen Gewicht und Alter für Tiere aus beiden Regionen unterschiedlich?

Für diese Fragen haben wir noch kein passendes Verfahren, aber zunächst können wir für beide Teilstichproben ein lineares Modell anpassen.

Truthahngewicht

```
R> fm3a <- lm(Gewicht ~ Alter, subset = Region == "Region1")
R> fm3a

Call:
lm(formula = Gewicht ~ Alter, subset = Region == "Region1")

Coefficients:
(Intercept)        Alter
     2.3143       0.4571

R> fm3b <- lm(Gewicht ~ Alter, subset = Region == "Region2")
R> fm3b

Call:
lm(formula = Gewicht ~ Alter, subset = Region == "Region2")

Coefficients:
(Intercept)        Alter
    0.06667      0.50833
```

Truthahngewicht

Doch auch diese getrennten Modell können in einem einzigen Modell angepaßt werden, das äquivalent zu den zwei Teilmodellen ist.

```
R> fm4 <- lm(Gewicht ~ Alter * Region)
R> fm4

Call:
lm(formula = Gewicht ~ Alter * Region)

Coefficients:
        (Intercept)                Alter         RegionRegion2
            2.31429              0.45714              -2.24762
Alter:RegionRegion2
            0.05119
```

Truthahngewicht

Die Achsenabschnitte der Modelle `fm3a` und `fm3b` sind klar unterschiedlich, aber die Steigungen (also die Abhängigkeit vom Alter) scheint ähnlich zu sein.

```
R> fm3a

Call:
lm(formula = Gewicht ~ Alter, subset = Region == "Region1")

Coefficients:
(Intercept)        Alter
     2.3143       0.4571

R> fm3b

Call:
lm(formula = Gewicht ~ Alter, subset = Region == "Region2")

Coefficients:
(Intercept)        Alter
    0.06667      0.50833
```

Truthahngewicht

Die Kodierung der Koeffizienten im Modell `fm4` macht dies noch deutlicher. Es gibt die beiden Regressionskoeffizienten für die Region 1 an, sowie die Differenz zu den Koeffizienten in Region 2.

```
R> coef(fm4)

        (Intercept)              Alter       RegionRegion2 Alter:RegionRegion2
         2.31428571         0.45714286         -2.24761905          0.05119048

R> coef(fm3a)

(Intercept)       Alter
  2.3142857   0.4571429

R> coef(fm3b) - coef(fm3a)

(Intercept)       Alter
-2.24761905  0.05119048
```

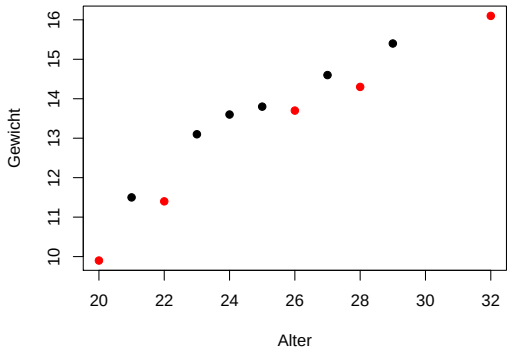
Truthahngewicht

Es scheint also so zu sein, daß die Achsenabschnitte in den beiden Regionen unterschiedlich sind, aber die Steigungen gleich.

Ach dieses Modell kann durch `lm` angepaßt werden.

Truthahngewicht

```
R> plot(Gewicht ~ Alter, pch = 19, col = as.numeric(Region))
```



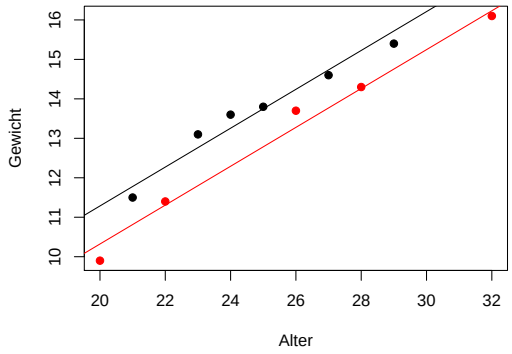
Truthahngewicht

```
R> fm3 <- lm(Gewicht ~ Alter + Region)
R> fm3

Call:
lm(formula = Gewicht ~ Alter + Region)

Coefficients:
  (Intercept)          Alter  RegionRegion2
       1.4362         0.4925        -0.9643

R> anova(fm3)

Analysis of Variance Table

Response: Gewicht
          Df  Sum Sq Mean Sq F value    Pr(>F)
Alter      1 30.4591 30.4591 298.773 1.278e-07 ***
Region     1  2.5053  2.5053  24.575  0.001111 **
Residuals  8  0.8156  0.1019
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Truthahngewicht

```
R> plot(Gewicht ~ Alter, pch = 19, col = as.numeric(Region))
R> abline(coef(fm3)[1:2])
R> abline(c(sum(coef(fm3)[c(1, 3)]), coef(fm3)[2]), col = 2)
```

Truthahngewicht



Truthahngewicht

Eigentlich waren wir an folgender Frage interessiert:

Frage: 3. Ist der Zusammenhang zwischen Gewicht und Alter für Tiere aus beiden Regionen unterschiedlich?

Diese Frage können wir nun statistisch umformulieren? Ist das Modell mit unterschiedlichen Steigungen für die Regionen `fm4` besser als das Modell mit demselben Zusammenhang (Steigung) für beide Regionen?

Truthahngewicht

Antwort: Nein, ein Vergleich beider Modelle mit einer Kovarianzanalyse zeigt, daß die Verbesserung nicht signifikant ist.

```
R> anova(fm3, fm4)

Analysis of Variance Table

Model 1: Gewicht ~ Alter + Region
Model 2: Gewicht ~ Alter * Region
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1      8 0.81558
2      7 0.74167  1   0.07391 0.6976 0.4312
```

Truthahngewicht

Auch eine Modellwahl per AIC gibt wählt `fm3` als das beste Modell aus.

```
R> AIC(fm1, fm2, fm3, fm4)

    df      AIC
fm1  3 24.04248
fm2  3 49.24837
fm3  4 10.59734
fm4  5 11.55239
```

Truthahngewicht

Die Modellwahl per AIC muß nicht unbedingt "von Hand" durchgeführt werden: alle relevanten/interessanten Modelle anpassen und dann das AIC ausrechnen.

Alternativ kann auch ein *Minimalmodell* und ein *Maximalmodell* angegeben werden, zwischen denen durch schrittweise Suche das beste Modell gesucht wird.

Dies findet nicht notwendigerweise das global beste Modell (speziell wenn es viele Erklärungsvariablen gibt), aber liefert bei einem guten Ausgangsmodell in aller Regel gute Ergebnisse.

Truthahngewicht

```
R> step(fm1, scope = list(lower = Gewicht ~ 1,
                          upper = Gewicht ~ Alter * Region))

Start:  AIC= -9.17
 Gewicht ~ Alter

         Df Sum of Sq     RSS     AIC
+ Region  1     2.505   0.816 -22.619
<none>                  3.321  -9.174
- Alter   1    30.459  33.780  14.342
```

Truthahngewicht

```
Step:  AIC= -22.62
 Gewicht ~ Alter + Region

               Df Sum of Sq     RSS     AIC
<none>                       0.816 -22.619
+ Alter:Region  1     0.074   0.742 -21.664
- Region        1     2.505   3.321  -9.174
- Alter         1    32.026  32.841  16.032

Call:
lm(formula = Gewicht ~ Alter + Region)

Coefficients:
  (Intercept)          Alter  RegionRegion2
       1.4362         0.4925        -0.9643
```

Truthahngewicht

Zusammenfassend:

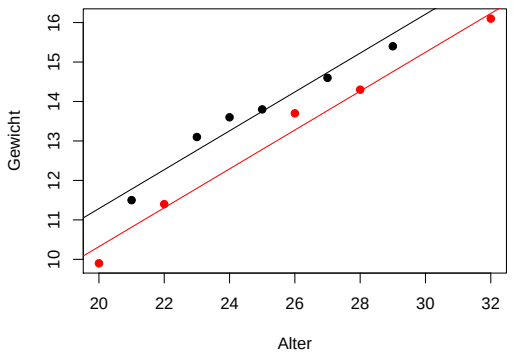
```
R> summary(fm3)

Call:
lm(formula = Gewicht ~ Alter + Region)

Residuals:
     Min       1Q   Median       3Q      Max
-0.42199 -0.20625  0.03800  0.21463  0.42300

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.43620    0.70225   2.045  0.07507 .
Alter         0.49250    0.02779  17.724 1.05e-07 ***
RegionRegion2 -0.96425    0.19451  -4.957  0.00111 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3193 on 8 degrees of freedom
Multiple R-Squared: 0.9759,        Adjusted R-squared: 0.9698
F-statistic: 161.7 on 2 and 8 DF,  p-value: 3.398e-07
```

Truthahngewicht

Im Durchschnitt unterscheidet sich das Gewicht von Truthähnen (gleichen Alters) in Region 1 und Region 2 um -0.96 lbs, ist also in Region 1 um gut 1 lb höher.

Der Zusammenhang zwischen Alter und Gewicht unterscheidet NA zwischen den beiden Regionen: Das Gewicht der Truthähne (in beiden Regionen) steigt im Schnitt um 0.49 Pfund pro Woche.

Truthahngewicht



Lineare Modelle in R

Die R-Funktion `lm()` paßt lineare Modelle an:

```
lm(formula, data, subset, weights, na.action, ...)
```

mit

- `formula` — eine symbolische Formel für das Modell,
- `data` — ein optionaler `data.frame`, der die Variablen enthält,
- `subset` — ein optionaler Vektor, der eine Teilmenge von Beobachtungen spezifiziert, die zur Anpassung verwendet werden sollen.

Lineare Modelle in R

`y ~ x` `y ~ 1 + x`	Einfache lineare Regression von y auf x. Der Achsenabschnitt steckt implizit in der ersten und explizit in der zweiten Formel.
`y ~ x - 1` `y ~ x + 0`	Einfache lineare Regression von y auf x durch den Ursprung (ohne Achsenabschnitt).
`log(y) ~ x1 + x2`	Multiple Regression der transformierten Variable $\log(y)$ auf x_1 und x_2 (mit implizitem Achsenabschnitt).
`y ~ 1 + x + I(x^2)` `y ~ poly(x, 2)`	Polynomiale Regression vom Grad 2 von y auf x.

Lineare Modelle in R

`y ~ a`	Einfache Varianzanalyse von y mit durch a festgelegten Klassen.
`y ~ a + x`	Einfache Kovarianzanalyse von y mit durch a festgelegten Klassen und Kovariable x.
`y ~ a + b`	Zwei-Weg-Varianzanalyse von y, ohne Interaktion.
`y ~ a * b` `y ~ a + b + a : b`	Zwei-Weg-Varianzanalyse von y nach a und b mit Interaktion.

Lineare Modelle in R

`y ~ (a + b + c)^2` `y ~ a*b*c - a:b:c`	Drei-Weg Experiment mit einem Modell, das die Haupteffekte und alle paarweisen Interaktionen (aber keine Drei-Weg-Interaktion) enthält.
`y ~ a * x`	Separate lineare Regressionen von y nach x getrennt für die Ausprägungen von a.

Lineare Modelle in R

`coefficients(lmobj)`	Regressionskoeffizienten extrahieren. Kurz: `coef(lmobj)`.
`deviance(lmobj)`	Fehlerquadratsumme.
`fitted.values(lmobj)`	Prognosewerte (fitted values). Kurz: `fitted(lmobj)`.
`plot(lmobj)`	Diagnostische Plots.
`predict(lmobj)`	Prognosewerte (auf den Originaldaten oder neuen Daten).
`print(lmobj)`	druckt den Funktionsaufruf und die Regressionskoeffizienten.
`residuals(lmobj)`	Residuen extrahieren.
`summary(lmobj)`	Ausführliche Zusammenfassung der Ergebnisse eines linearen Modells.