Arbeitsblatt 4

Multiple lineare Regression (Einführung)

Aufgabe 1: Werbeausgaben

Der Datensatz `Advertisement.rda` enthält die Werbeausgaben (in 1000 US Dollar) für TV, Radio und Zeitung für 100 verschiedene Absatzmärkte. In `Sales` sind die Verkaufszahlen (pro 1000 Artikel) für jeden Absatzmarkt gegeben.

Einlesen der Daten

```r
load("data/Advertisement.rda")
```

(a)

Passen Sie ein Modell mit Zielgrösse `sales` und den drei erklärenden Variablen `TV`, `radio` und `newspaper` an. Notieren Sie das Modell in mathematischer Notation.

```r
# Anpassung der Regression
fit.adv <- lm(sales ~ TV + radio + newspaper, data = adv)
summary(fit.adv)
```

```
Call:
lm(formula = sales ~ TV + radio + newspaper, data = adv)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9257 -0.6216  0.1613  0.6923  2.9996

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.352150   1.649721  -1.426    0.156
TV           0.046382   0.003162  14.669  < 2e-16 ***
radio        0.079783   0.011970   6.665 2.62e-10 ***
newspaper    0.021015   0.015354   1.369    0.173
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.135 on 196 degrees of freedom
Multiple R-squared:  0.6311,    Adjusted R-squared:  0.6254
F-statistic: 111.8 on 3 and 196 DF,  p-value: < 2.2e-16
```

Das Modell ist:

$$\text{sales}_i = \beta_0 + \beta_1 \cdot \text{TV}_i + \beta_2 \cdot \text{radio}_i + \beta_3 \cdot \text{newspaper}_i + E_i$$

wobei $i = 1, \ldots, 200$ und $E_i \sim \mathcal{N}(0, \sigma^2)$, unabhängig.

(b)

Wie lauten die geschätzten Koeffizienten? Geben Sie eine Interpretation zum Achsenabschnitt und dem Koeffizient für die Variable `radio`?

```
# Geschätzte Koeffizienten
coef(fit.adv)
```

```
(Intercept)          TV       radio   newspaper 
-2.35215035  0.04638172  0.07978281  0.02101510 
```

Die Schätzungen für die Koeffizienten sind: $\hat{\beta}_0 = -2.352$, $\hat{\beta}_1 = 0.0463, \hat{\beta}_2 = 0.0798$: und : $\hat{\beta}_3 = 0.0210$.

Interpretation:

- Wenn keine Werbung gemacht wird (TV = 0, radio = 0 und newspaper = 0), dann verkaufen sich im Mittel -2352 Artikel. Achtung, dieser Wert macht keinen Sinn. Es handelt es sich wieder einmal um eine Extrapolation.
- Wenn die Werbeausgaben für TV und Zeitung konstant bleiben und man die Werbeausgaben fürs Radio um 1000 US$ erhöht, dann steigen die Verkaufszahlen um 798 Artikel.

(c)

Prüfen Sie mit einem geeigneten statistischen Test, ob mindestens eine der erklärenden Variablen einen auf 5% signifikanten Einfluss auf die Verkaufszahlen hat?

```
F-statistic: 111.75 on 3 and 196 DF, p-value:  <2e-16
```

Aus dem `summary`-Output sieht man, dass der p-Wert für den globalen F-Test kleiner als 0.05 ist, somit hat mindestens eine der 3 erklärenden Variablen einen auf dem 5% Niveau signifikanten Einfluss auf die Verkaufszahlen.

(d)

Zu welchem Prozentanteil lässt sich die Verkaufszahl mit diesem Modell erklären?

Aus dem `summary()`-Output der Regression lässt sich das Bestimmtheitsmass R^2 ablesen:

```
summary(fit.adv)$r.squared
```

```
[1] 0.6310669
```

```
summary(fit.adv)$adj.r.squared
```

```
[1] 0.62542
```

Das Multiple R-Squared ist 63.11%. Korrigiert man für die Anzahl verwendeter erklärender Variablen ist der Anteil erklärter Varianz mit 62.54% fast identisch.

(e)

Berechnen Sie die 99% Vertrauensintervalle für den Anstiegs der Verkaufszahlen, wenn die Werbeausgaben für TV und Zeitung konstant bleiben und man die Werbeausgaben für radio um 1 US$ erhöht.

```
confint(fit.adv, level = 0.99)
```

```
                   0.5 %      99.5 %
(Intercept) -6.64331323 1.93901253
TV           0.03815706 0.05460639
radio        0.04864593 0.11091970
newspaper   -0.01892179 0.06095199
```

Aus der Zeile für `radio` entnehmen wir, dass das wahre β_2 mit 99% Wahrscheinlichkeit zwischen 0.0486 und 0.1109 liegt, d.h. wenn bei gleichen Werbeausgaben für TV und Newspaper die Werbeausgaben für Radio um 1 US$ erhöht werden, dann erhöhen sich die Verkaufszahlen mit 99% zwischen 48 und 111 Artikeln.

(f)

Berechnen Sie eine Vorhersage für die Anzahl verkaufter Artikel, wenn die Werbeausgaben für TV bei 150000 US$, für Radio bei 40000 US $ und für Zeitungen bei 100000 US $ liegen. Geben Sie zusätzlich noch ein 95% Prognose-Intervall an.

Wir erstellen einen Data.frame mit den "neuen" Werten für die erklärenden Variablen. Die Spaltennamen müssen mit denjenigen von `adv` übereinstimmen und auch auf die Einheiten ist zu achten.

```r
x0 <- data.frame(TV = 150, radio = 40, newspaper = 100)
```

Nun berechnen wir eine Vorhersage für `x0`.

```r
predict(fit.adv, newdata = x0, interval = "prediction", level = 0.95)
```

```
      fit      lwr      upr
1 9.89793 7.650372 12.14549
```

Für diese Werbeausgaben beträgt der mittlere Verkaufszahl bei 9897 Artikel. Mit 95% Wahrscheinlichkeit liegt der Verkaufszahl zwischen 7650 und 12145 Artikeln.

(g)

Prüfen Sie mit einem geeigneten statistischen Test, ob die Werbeausgaben für die Zeitungen einen auf dem 5% Niveau signifikanten Einfluss auf die Verkaufszahl des Artikels hat?

Hierfür benützen wir den t-Test mit Hypothese $H_0 : \beta_3 = 0$ gegen $H_A : \beta_3 \neq 0$.

```
               Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -2.35215035 1.64972054 -1.425787 1.555205e-01
TV           0.04638172 0.00316194 14.668755 2.294310e-33
radio        0.07978281 0.01197045  6.664979 2.620505e-10
newspaper    0.02101510 0.01535358  1.368743 1.726463e-01
```

Der p-Wert zu `newspaper` ist grösser als 0.05 (=Niveau). Somit kann die Null-Hypothese $H_0 : \beta_3 = 0$ nicht verworfen werden, d.h. wir müssen fast sicher davon ausgehen, dass die Werbung durch Zeitung **keinen** Einfluss auf die Verkaufszahlen hat.

(h)

Entfernen Sie die Variable `newspaper` aus dem Modell und visualisieren Sie die Situation mit einem 3D Plot. Verwenden Sie hierfür `scatter3d` aus dem R-Package `car`.

```
library(car)
scatter3d(sales ~ TV + radio, data = adv,  axis.scales = FALSE)
```

Aufgabe 2: Katheter

In dieser Aufgabe analysieren wir den Datensatz `catheter.dat`. Es handelt sich um Daten aus der Medizin. Die Variable `Groesse` ist die Grösse (in cm), `Gewicht` das Gewicht eines Patienten (in kg) und `y` die optimale Länge eines Katheters (in cm), der für die Untersuchung des Herzens eingesetzt wird. Man möchte gerne die Katheter-Länge aus den Patienten-Daten schätzen.

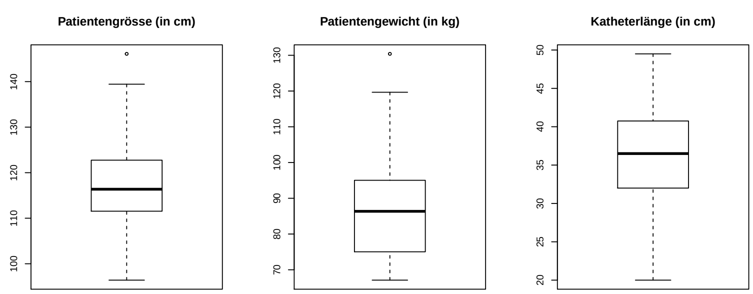
Einlesen der Daten

```
catheter <- read.table("data/catheter.dat", header=TRUE, sep = ",")
```

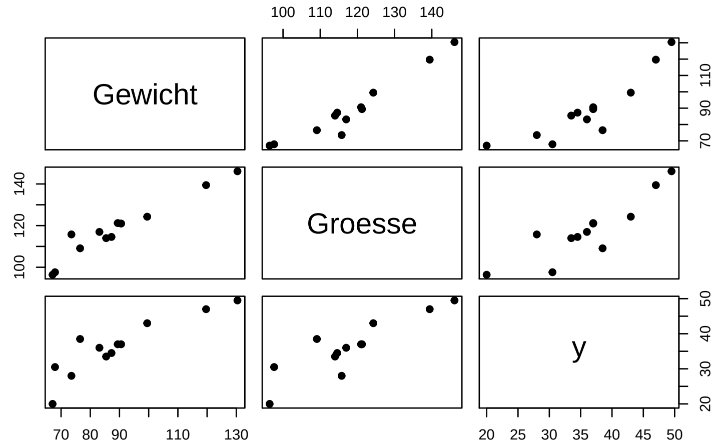
(a)

Untersuchen Sie den Datensatz mit Hilfe von Boxplots und zweidimensionalen Streudiagramme `y` gegen `Groesse`, `y` gegen `Gewicht` und `Gewicht` gegen `Groesse`. Was fällt Ihnen auf?

```
# Boxplots
par(mfrow = c(1,3))
boxplot(catheter$Groesse, main = "Patientengrösse (in cm)")
boxplot(catheter$Gewicht, main = "Patientengewicht (in kg)")
boxplot(catheter$y, main = "Katheterlänge (in cm)")
```




```
# Streudiagramme
pairs(catheter[,3:1], pch = 19)
```



Die Variable `Groesse` enthält Werte im Bereich von 96.4cm bis 146.1cm, was ziemlich klein ist im Vergleich zum Patientengewicht, das zwischen 67.1kg und 130.4kg liegt. Vermutlich beschreibt diese Variable nicht die Körpergrösse der Patienten, sondern eine andere charakteristische Körperlänge.

Die beiden Variablen `Groesse` und `Gewicht` zeigen eine hohe Korrelation miteinander und auch mit der Zielgrösse `y`.

(b)

Berechnen Sie zwei einfache lineare Regressionen von `y` auf `Groesse` und `y` auf `Gewicht`. Geben Sie ausserdem jeweils die Schätzungen für die Koeffizienten und $\hat{\sigma}$ an.

```r
cath.fit1 <- lm(y ~ Groesse, catheter)
coef(cath.fit1)
```

```
(Intercept)     Groesse
-21.8911291   0.4922262
```

```r
summary(cath.fit1)$sigma # Schätzung für Sigma
```

```
[1] 4.008547
```

```r
cath.fit2 <- lm(y ~ Gewicht, catheter)
coef(cath.fit2)
```

```
(Intercept)     Gewicht
  2.9532420   0.3727923
```

```r
summary(cath.fit2)$sigma
```

```
[1] 3.79671
```

(c)

Testen Sie in beiden Modellen mit Hilfe des Regressions-Outputs die Hypothese H_0 : *Steigung* $\beta = 0$ gegen H_A : *Steigung* $\beta \neq 0$.

```
summary(cath.fit1)$coefficients
```

```
               Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -21.8911291 9.92649209 -2.205324 0.0519705788
Groesse       0.4922262 0.08352515  5.893149 0.0001524578
```

```
summary(cath.fit2)$coefficients
```

```
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 2.9532420 5.38015253 0.5489142 5.951073e-01
Gewicht     0.3727923 0.05904723 6.3134598 8.754745e-05
```

Die Nullhypothese kann bei beiden Modellen auf dem 5%-Niveau verworfen werden, da die beiden P-Werte kleiner als 0.05 sind. Somit scheinen das Gewicht und die Grösse einen Einfluss auf die Katheterlänge zu haben.

(d)

Wir führen nun eine multiple lineare Regression durch, d.h. passen Sie das Modell

$$Y_i = \beta_0 + \beta_1 \texttt{Groesse}_i + \beta_2 \texttt{Gewicht}_i + E_i$$

an die Daten an. Kommentieren Sie den globalen F-Test und t-Test für die einzelnen Koeffizienten. Vergleichen Sie die Ergebnisse mit denjenigen der einfachen linearen Regression.

```
cath.fit <- lm(y ~ Groesse + Gewicht, catheter)
summary(cath.fit)
```

```
Call:
lm(formula = y ~ Groesse + Gewicht, data = catheter)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.0497 -1.2753 -0.2595  1.9095  6.9933 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -5.6381    17.0599  -0.330    0.749
Groesse       0.1590     0.2984   0.533    0.607
Gewicht       0.2587     0.2227   1.161    0.275

Residual standard error: 3.94 on 9 degrees of freedom
Multiple R-squared:  0.8056,	Adjusted R-squared:  0.7624 
F-statistic: 18.65 on 2 and 9 DF,  p-value: 0.0006301
```

Der globale F-Test ist auf dem 5% Niveau signifikant (p-Wert < 0.05) und somit hat mindestens eine der beiden erklärenden Variablen einen signifikanten Einfluss auf die Katheterlänge. Die Nullhypothese $H_0\ \beta = 0$ wird jedoch für beide erklärende Variablen auf dem 5%-Niveau **nicht** verworfen werden, da die beiden P-Werte grösser als 0.05 sind. Im Gegensatz dazu zeigten bei der einfachen linearen Regression beide Variablen einen signifikanten Einfluss.

(e)

Vergleichen Sie die Schätzung für $\hat{\sigma}$ von der multiplen linearen Regression mit jenen von der einfachen linearen Regression.

```
summary(cath.fit1)$sigma
```

```
[1] 4.008547
```

```
summary(cath.fit2)$sigma
```

```
[1] 3.79671
```

```
summary(cath.fit)$sigma
```

```
[1] 3.940388
```

Bei der Hinzunahme einer Variable ins Modell wird die Fehlervarianz nicht besser. Die Ausdehnung des Modells von einer einfachen Regression zu einem Regressionsmodell mit zwei erklärenden Variablen scheint sich nicht zu lohnen. Das Gewicht und die Grösse erklären zusammen die Zielvariable nicht besser.

Schlussbemerkung

Was haben wir aus dieser Aufgabe gelernt?

- Falls die einzelnen Variablen gemäss t-Test nicht signifikant sind, heisst das noch lange nicht, dass all diese Variablen *insgesamt* nicht signifikant sind.
- In der Statistik sind mehr Informationen nicht immer besser. Erklärende Variablen, die nichts bzw. nicht genügend beitragen, bringen nur mehr Unsicherheiten ins Modell.