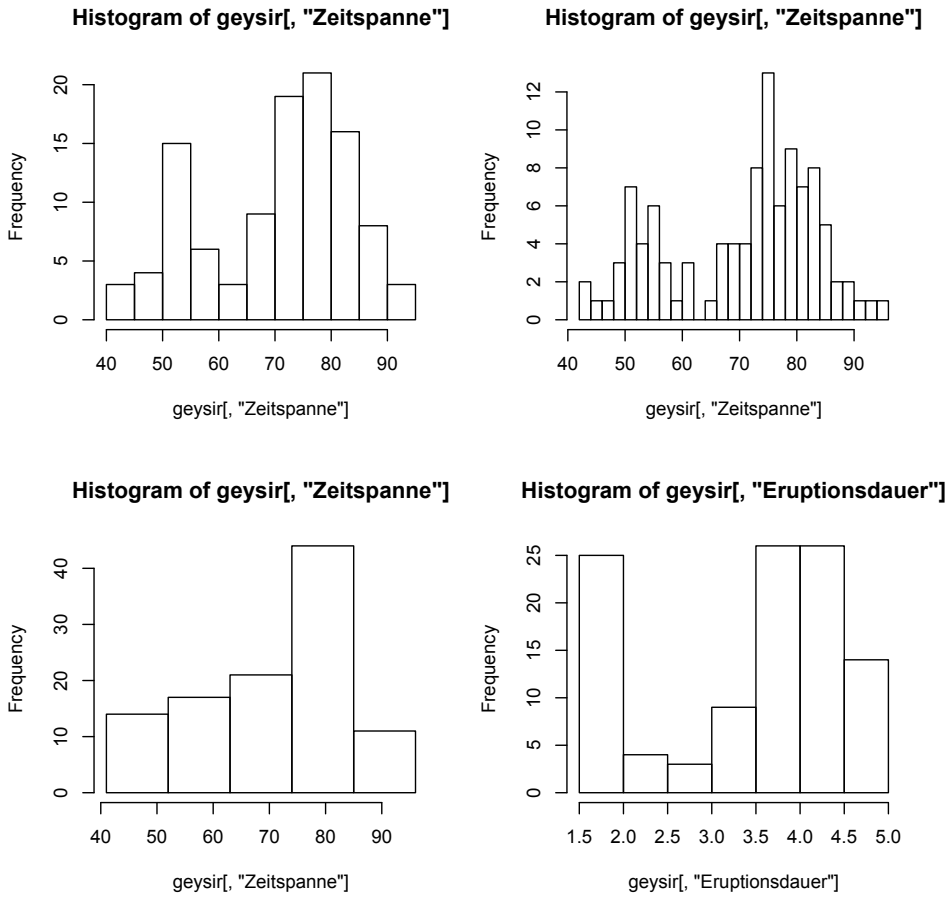
Stochastik

Musterlösungen zu Serie 2

Lösung 2.1

a)
```
# Datensatz einlesen
geysir <- read.table("./Daten/geysir.dat", header = TRUE)
par(mfrow = c(2, 2))  # 4 Grafiken im Grafikfenster
# Histogramme zeichnen
hist(geysir[, "Zeitspanne"])
hist(geysir[, "Zeitspanne"], breaks = 20)
hist(geysir[, "Zeitspanne"], breaks = seq(41, 96, by = 11))
hist(geysir[, "Eruptionsdauer"])
```



Die ersten drei Histogramme in der Abbildung zeigen die Intervalle zwischen zwei Ausbrüchen von Old Faithful. Auffallend ist, dass Zeitspannen um 55 Mi-

nuten aber auch zwischen 70 und 85 Minuten häufiger vorkommen als andere Intervalle. So eine Verteilung mit zwei Gipfeln heisst auch *bimodal*.

Werden die Klassenbreiten ungeschickt gewählt, entdeckt man diese Besonderheit der Geysirdaten nicht. Das ist im dritten Histogramm passiert. Das Beispiel illustriert, dass die richtige Wahl der Klassenbreiten- bzw. grenzen wohlüberlegt sein muss.

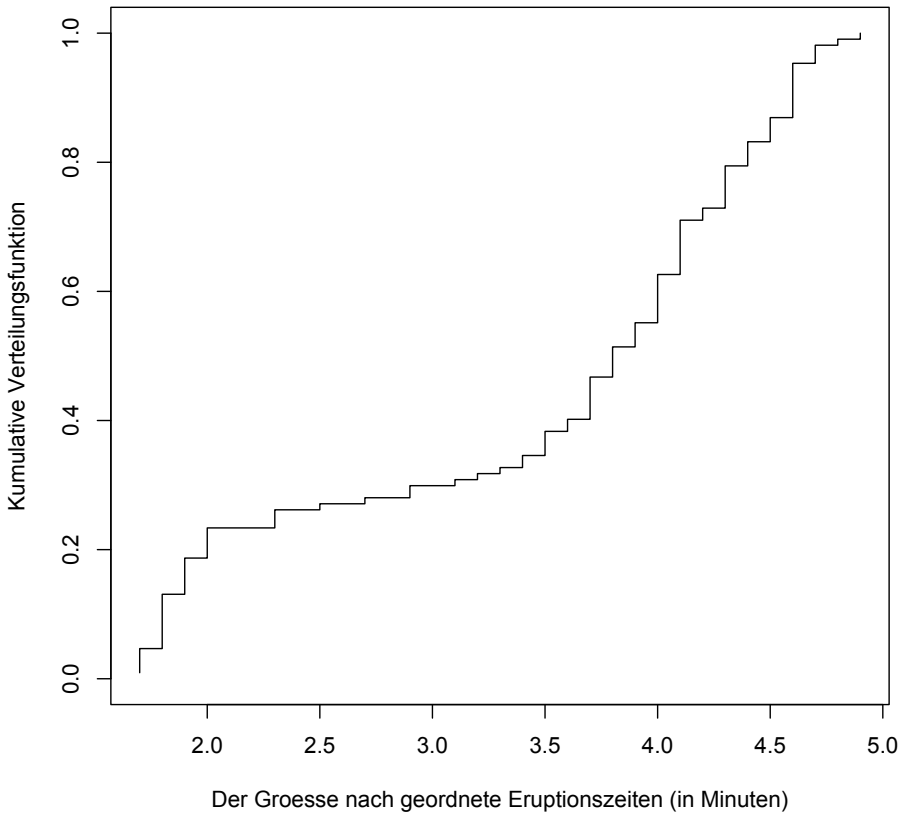
b) Das vierte Histogramm schliesslich zeigt die Häufigkeiten verschiedener Eruptionsdauern. Hier sind die beiden Gipfel sehr deutlich erkennbar: „Entweder ist der Ausbruch sofort wieder vorbei, oder er dauert mindestens dreieinhalb Minuten“. Ob die Dauer eines Ausbruchs aber etwas zu tun hat mit der Dauer des vorangegangenen Ruheintervalls (mit anderen Worten: ob die Gipfel des Histogramms aus Teilaufgabe **b**) den Gipfeln der Histogramme aus Teilaufgabe **a**) entsprechen), kann man aufgrund dieser Darstellungen nicht sagen.

c) Die **kumulative Verteilungsfunktion** ist definiert als

$$F_n(x) = \frac{1}{n}\text{Anzahl}\{i \mid x_i \leq x\}.$$

Der Funktionswert von $F_n(x)$ springt also bei jeder Beobachtung $x_i \leq x$ um den Wert $1/n$. Mit R können wir den Funktionsgraphen der kumulativen Verteilungsfunktion folgendermassen zeichnen:

```
eruptionsdauern <- geysir[, "Eruptionsdauer"]
n <- length(eruptionsdauern)
plot(sort(eruptionsdauern), (1:n)/n,
    type = "s", ylim = c(0, 1),
    ylab = "Kumulative Verteilungsfunktion",
    xlab = "Der Groesse nach geordnete Eruptionszeiten (in Minuten)",
    main = "")
```



Aus der Graphik der kumulativen Verteilungsfunktion sehen wir, dass rund 20 % der Eruptionszeiten weniger als 2 Minuten dauern. Rund 60 % der Eruptionen dauern mindestens 3.7 Minuten. Aus der Graphik sehen wir ebenfalls, dass die Zeitspannen entweder unter 2 Minuten liegen, oder dann über 3.5 Minuten (bimodale Verteilung).

Lösung 2.2

a) Der ursprüngliche Datensatz hat für den Median und Mittelwert folgende Werte:

```
noten.1 <- c(4.2, 2.3, 5.6, 4.5, 4.8, 3.9, 5.9, 2.4, 5.9,
    6, 4, 3.7, 5, 5.2, 4.5, 3.6, 5, 6, 2.8, 3.3, 5.5, 4.2,
    4.9, 5.1)
median(noten.1)

## [1] 4.65

mean(noten.1)
```

```
## [1] 4.5125
```

Zuerst ordnen wir die Datenwerte der Grösse nach:

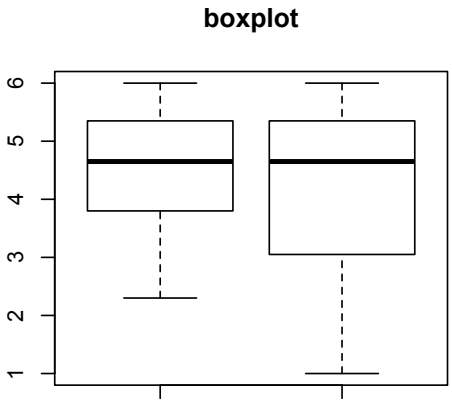
```
sort(noten.1)

##  [1] 2.3 2.4 2.8 3.3 3.6 3.7 3.9 4.0 4.2 4.2 4.5 4.5
## [13] 4.8 4.9 5.0 5.0 5.1 5.2 5.5 5.6 5.9 5.9 6.0 6.0
```

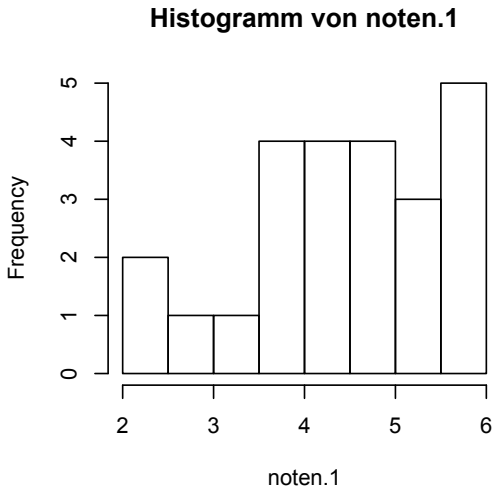
Da die Anzahl Noten gerade ist, wird der Median aus dem Mittelwert von $x_{(12)}$ und $x_{(13)}$ gebildet. Wenn wir also Noten kleiner als $x_{(12)}$ abändern, wird sich der Median nicht ändern. Dementsprechend ändern wir die Notenwert $x_{(9)}, x_{(10)}, x_{(11)}$ zu einer eins. Dies lässt den Median unverändert, lässt den Mittelwert aber maximal schrumpfen.

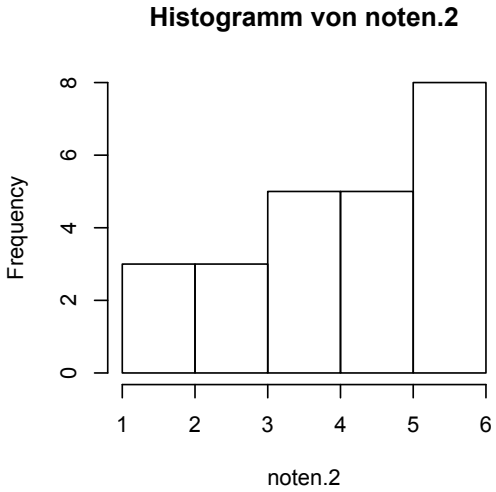
$$1, 2.3, 5.6, 4.5, 4.8, 3.9, 5.9, 2.4, 5.9, 6, 4, 3.7, 5, 5.2, 1, 3.6, 5, 6, 2.8, 3.3, 5.5, 1, 4.9, 5.1$$

```
noten.2 <- c(1, 2.3, 5.6, 4.5, 4.8, 3.9, 5.9, 2.4, 5.9,
    6, 4, 3.7, 5, 5.2, 1, 3.6, 5, 6, 2.8, 3.3, 5.5, 1, 4.9,
    5.1)
median(noten.2)

## [1] 4.65

mean(noten.2)

## [1] 4.1
```

b)
```
par(mfrow = c(2, 2))
boxplot(noten.1, noten.2, main = "boxplot")
hist(noten.1, main = "Histogramm von noten.1")
hist(noten.2, main = "Histogramm von noten.2")
```





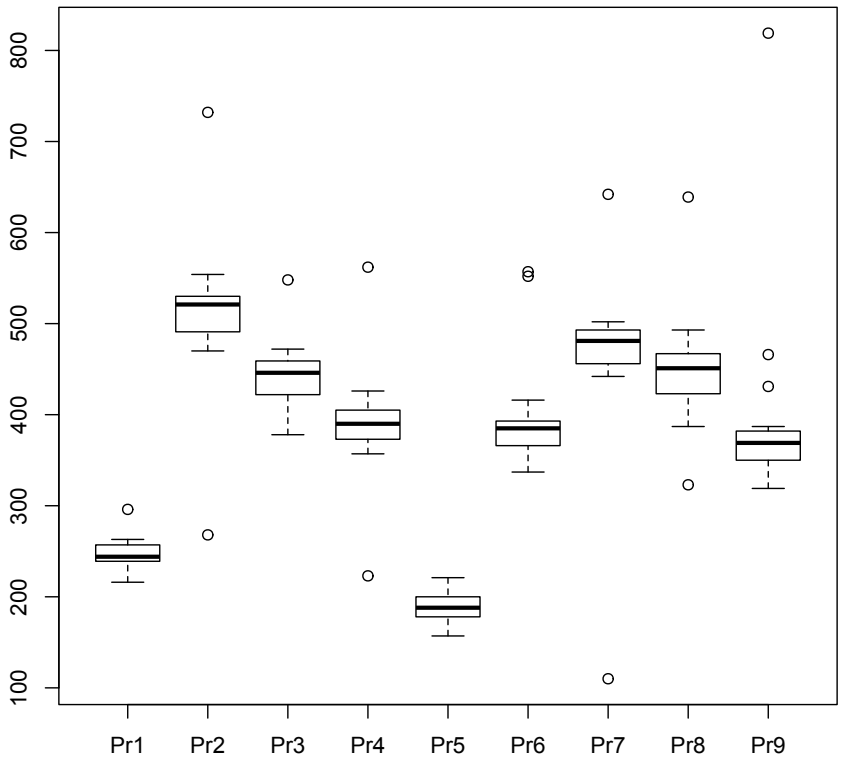


Lösung 2.3

```
schlamm.all <- read.table(file = "./Daten/klaerschlamm.dat",
    header = TRUE)
schlamm <- schlamm.all[, -1]   # Labor-Spalte entfernen
```

a) Aus den Boxplots erkennen wir, dass es vor allem bei den Proben 2, 4, 6, 7, 8 und 9 Ausreisser gibt. Das arithmetische Mittel und der Median unterscheiden wesentlich bei den Proben 2, 6, 7 und 9.

```
boxplot(schlamm)
```




```
summary(schlamm)
```

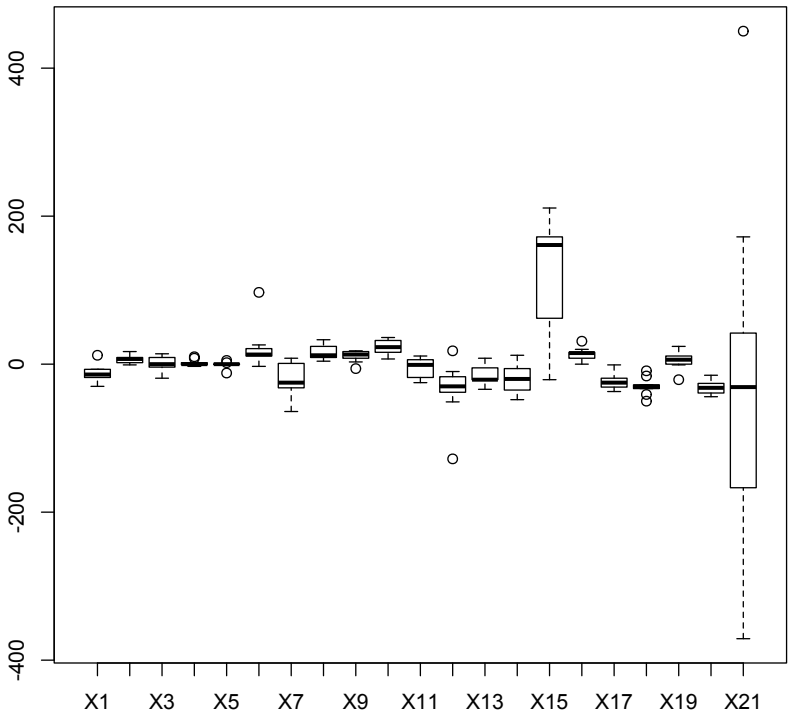
```
##       Pr1             Pr2             Pr3
##  Min.   :216.0   Min.   :268.0   Min.   :378.0
##  1st Qu.:239.0   1st Qu.:491.0   1st Qu.:422.0
##  Median :244.0   Median :521.0   Median :446.0
##  Mean   :246.1   Mean   :511.4   Mean   :443.4
##  3rd Qu.:257.0   3rd Qu.:530.0   3rd Qu.:459.0
##  Max.   :296.0   Max.   :732.0   Max.   :548.0
##       Pr4             Pr5             Pr6
##  Min.   :223.0   Min.   :157.0   Min.   :337.0
##  1st Qu.:373.0   1st Qu.:178.0   1st Qu.:366.0
##  Median :390.0   Median :188.0   Median :385.0
##  Mean   :389.2   Mean   :188.2   Mean   :394.9
##  3rd Qu.:405.0   3rd Qu.:200.0   3rd Qu.:393.0
##  Max.   :562.0   Max.   :221.0   Max.   :557.0
##       Pr7             Pr8           Pr9
##  Min.   :110.0   Min.   :323   Min.   :319.0
```

```
##  1st Qu.:456.0   1st Qu.:423   1st Qu.:350.0
##  Median :481.0   Median :451   Median :369.0
##  Mean   :465.5   Mean   :450   Mean   :388.9
##  3rd Qu.:493.0   3rd Qu.:467   3rd Qu.:382.0
##  Max.   :642.0   Max.   :639   Max.   :819.0
```

Bei den Proben 1 und 5 ist es plausibel, dass die Konzentration unter 400 mg/kg liegt, während wir bei Probe 2, 3, 7 und 8 dazu tendieren, den Grenzwert 400 mg/kg als überschritten zu betrachten. Die übrigen Proben, Probe 4, 6 und 9 sind eher Grenzfälle. Die Konzentrationen scheinen zwar unter 400 mg/kg zu liegen, die drei Proben weisen jedoch jeweils extreme Ausreisser über dem Grenzwert auf.

b) Als erstes stechen die Messungen der Labors 15 und 21 ins Auge. Beide haben sowohl eine grosse Standardabweichung als auch systematische Fehler. Die Labors 6 und 12 haben beide Ausreisser zu verzeichnen. Die Labors 1, 7, 12, 13, 14, 17, 18, 20 und 21 geben systematisch zu kleine Werte an, während die Labors 6, 8, 10 und 15 zu grosse Werte erhalten. Die Labors 2, 3, 4, 5 und 19 scheinen zuverlässige Untersuchungen durchzuführen. Sowohl systematische wie auch Zufallsfehler scheinen sich hier in Grenzen zu halten.

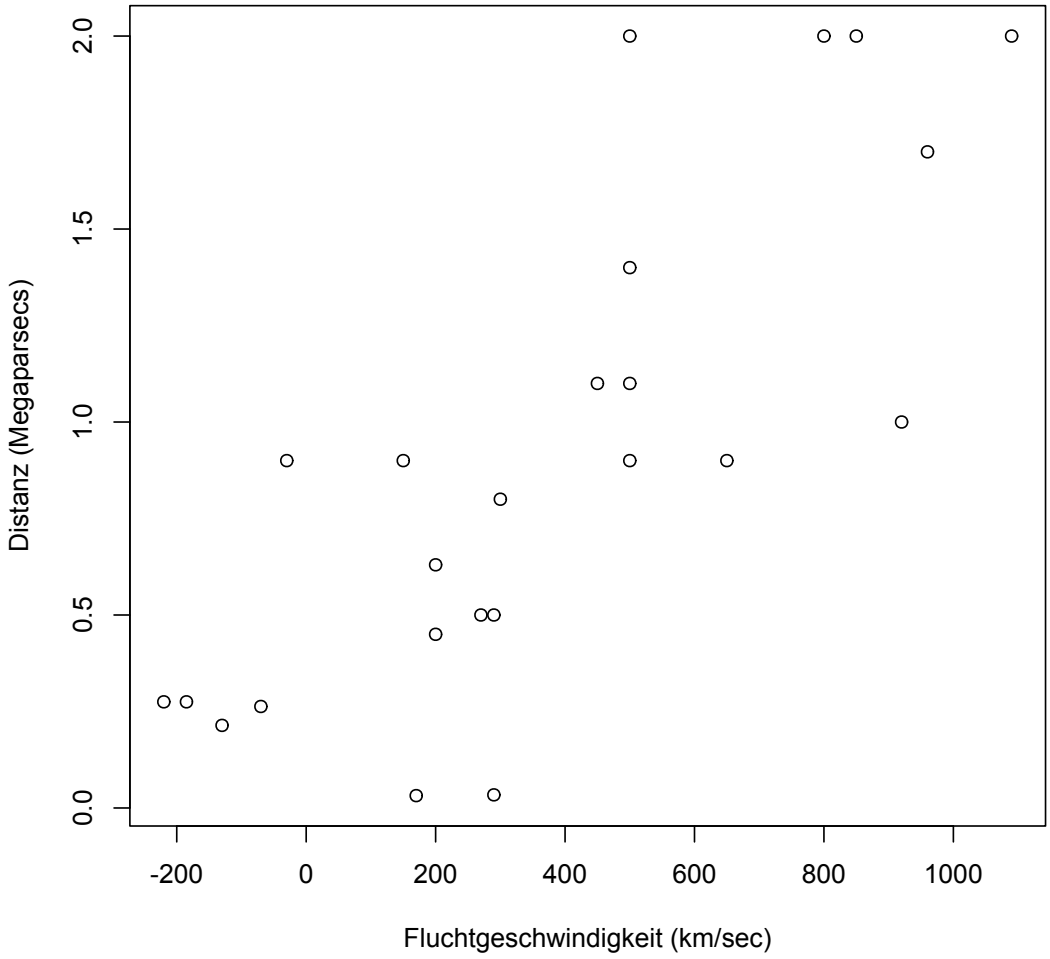
```r
# Fuer jede Spalte Median berechnen
med <- apply(schlamm, 2, median)
# Median von jeder *Spalte* abziehen
schlamm.centered <- scale(schlamm, scale = FALSE, center = med)
# Boxplot zeichnen. Dazu zuerst data-frame transponieren
boxplot(data.frame(t(schlamm.centered)))
```



Lösung 2.4 1b, 2c, 3a

Lösung 2.5

a)
```
recession.velocity <- c(170, 290, -130, -70, -185, -220, 200, 290,
    270, 200, 300, -30, 650, 150, 500, 920, 450, 500, 500, 960, 500,
    850, 800, 1090)
distance <- c(0.032, 0.034, 0.214, 0.263, 0.275, 0.275, 0.45, 0.5,
    0.5, 0.63, 0.8, 0.9, 0.9, 0.9, 0.9, 1, 1.1, 1.1, 1.4, 1.7, 2,
    2, 2, 2)
plot(recession.velocity, distance, ylab = "Distanz (Megaparsecs)",
    xlab = "Fluchtgeschwindigkeit (km/sec)")
```



b) Die Parameter β_0 und β_1 lassen sich mit der Methode der kleinsten Quadrate schätzen, wobei wir folgende Formeln für die Parameterschätzungen erhalten hatten:

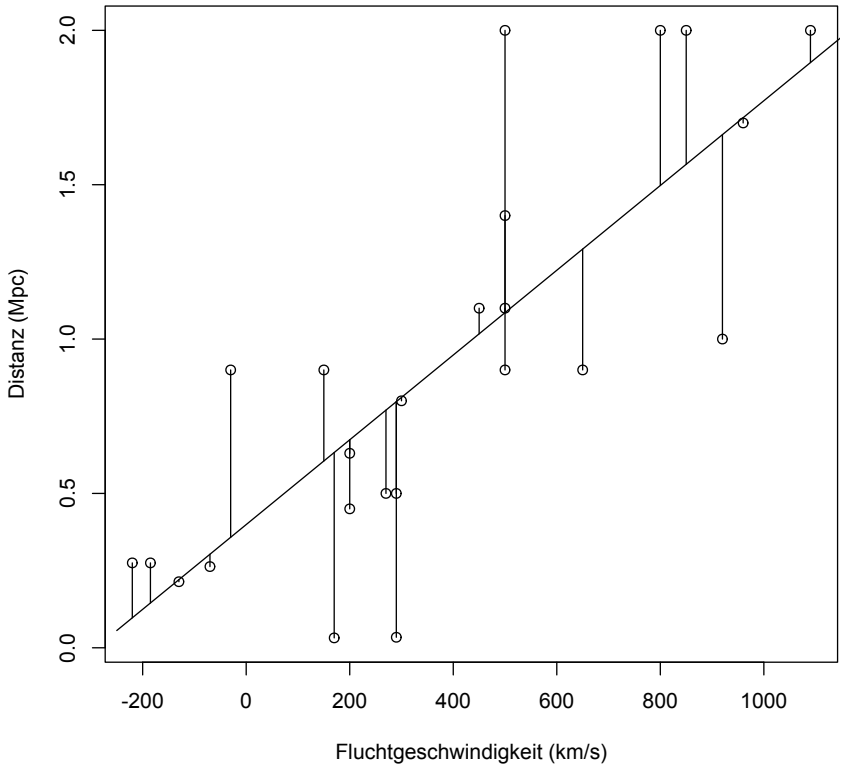
$$\hat{\beta_1} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x}\,.$$

```
beta_1 <- sum((distance - mean(distance)) * (recession.velocity -
    mean(recession.velocity)))/(sum((recession.velocity -
    mean(recession.velocity))^2))
beta_1
```

```
## [1] 0.001372936

beta_0 <- mean(distance) - beta_1 * mean(recession.velocity)
beta_0

## [1] 0.3990982
```

```
plot(recession.velocity, distance, ylab = "Distanz (Mpc)",
    xlab = "Fluchtgeschwindigkeit (km/s)")
lines(-250:1200, beta_0 + beta_1 * (-250:1200),
    type = "l", new = TRUE)
segments(recession.velocity, beta_0 + beta_1 *
    (recession.velocity), recession.velocity,
    distance)
```



c)

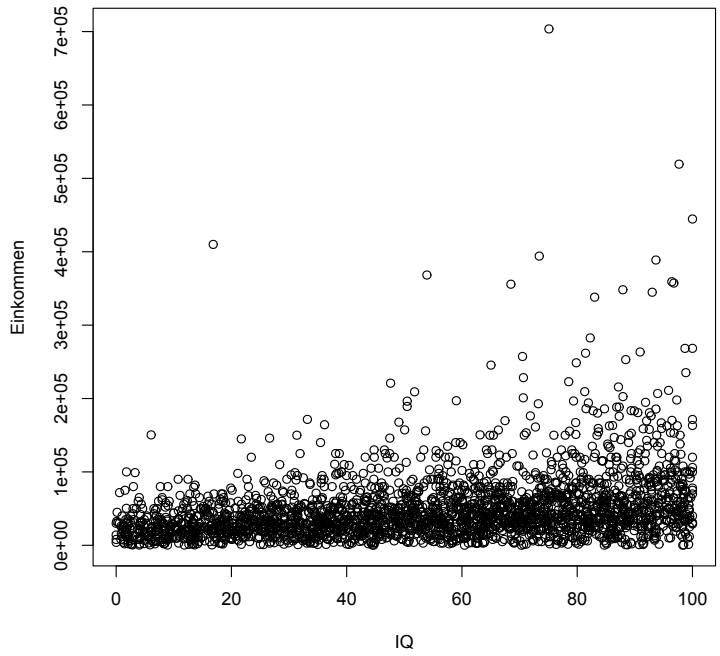
```
lm(distance ~ recession.velocity)

##
## Call:
```

```
## lm(formula = distance ~ recession.velocity)
##
## Coefficients:
##         (Intercept)  recession.velocity
##            0.399098            0.001373
```

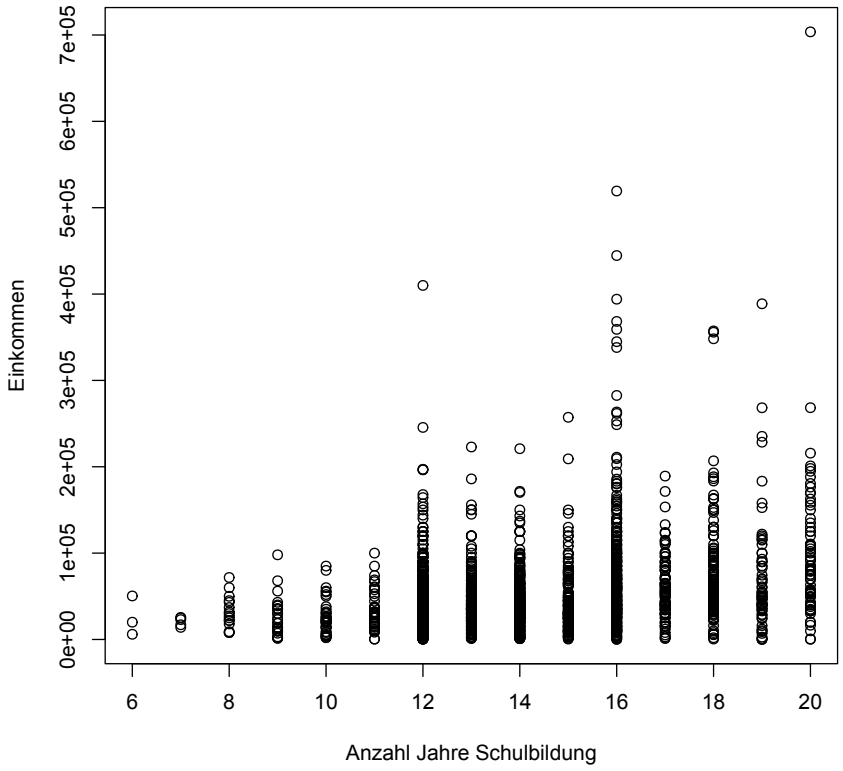
Lösung 2.6

a)
```
income <- read.table(file = "./Daten/income.dat", header = TRUE)
head(income)

##      AFQT Educ Income2005
## 1   6.841   12       5500
## 2 99.393   16      65000
## 3 47.412   12      19000
## 4 44.022   14      36000
## 5 59.683   14      65000
## 6 72.313   16       8000

iq <- income[, 1]
anzahl.jahre.schule <- income[, 2]
einkommen <- income[, 3]
```

```
plot(iq, einkommen, type = "p", xlab = "IQ", ylab = "Einkommen")
```



```r
plot(anzahl.jahre.schule, einkommen, type = "p",
    xlab = "Anzahl Jahre Schulbildung", ylab = "Einkommen")
```

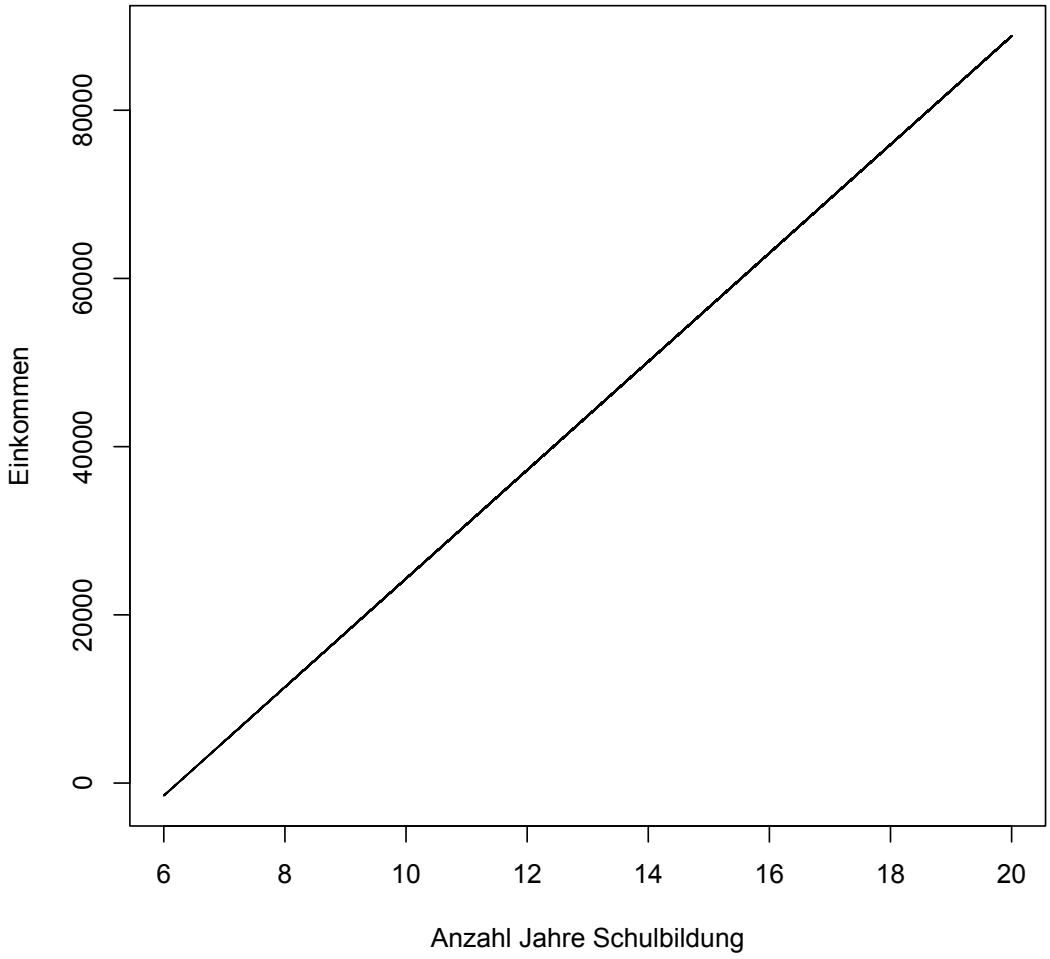


b) Mit R ermitteln wir für a und b

```r
lm(einkommen ~ anzahl.jahre.schule)

##
## Call:
## lm(formula = einkommen ~ anzahl.jahre.schule)
##
## Coefficients:
##         (Intercept)  anzahl.jahre.schule
##               -40200                 6451
```

Somit lautet die Gleichung für die Regressionsgerade $y = -40200 + 6451 \cdot x$.

```r
plot(anzahl.jahre.schule, -40200 + 6451 *
    anzahl.jahre.schule, type = "l",
    xlab = "Anzahl Jahre Schulbildung",
    ylab = "Einkommen")
```



Wir finden also die Werte $a = -40'200$ und $b = 6451$ für den Fall von Einkommen gegen Anzahl Jahre Schulbildung (und $a = 21'182$ und $b = 518.68$ für den betrachteten Fall Einkommen gegen Intelligenzquotient). Mit jedem zusätzlichen Jahr Schulbildung geht also eine jährliche Einkommenszunahme von 6451 USD einher. Nun ist allerdings Vorsicht geboten: jemand ohne Schulbildung würde ein Einkommen von $-40'200$ USD haben. Dies macht natürlich keinen Sinn. Wann immer man in Bereiche extrapoliert, wo keine Datenpunkte vorhanden waren, ist Vorsicht bei der Interpretation geboten.

c) Für die **empirische Korrelation** erhalten wir dann

```
cor(anzahl.jahre.schule, einkommen)

## [1] 0.3456474
```

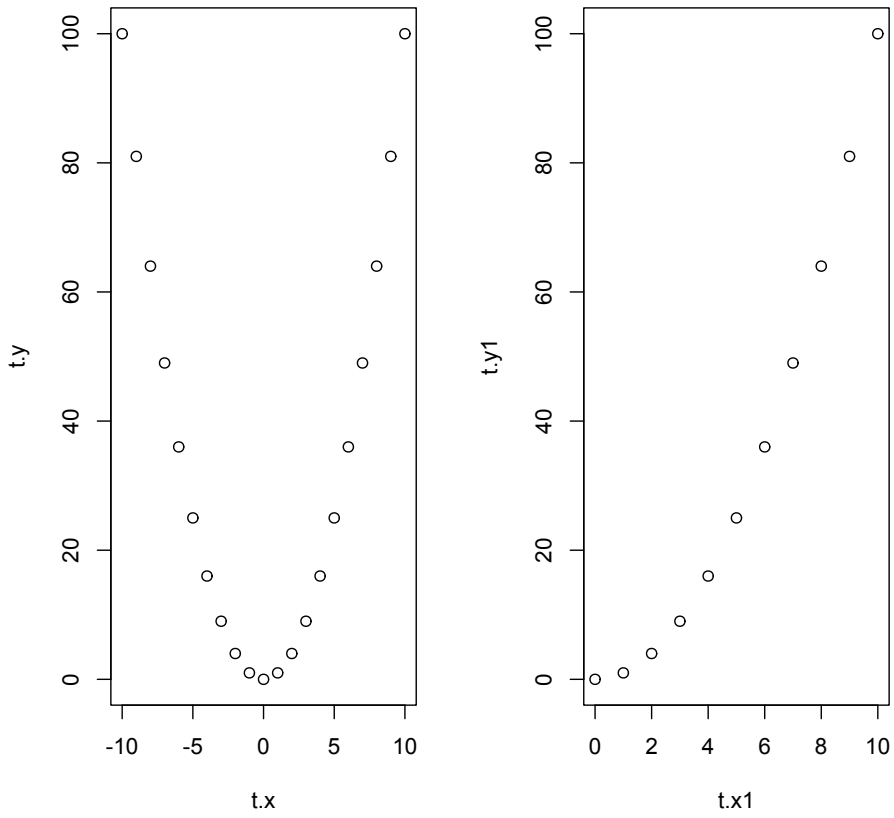
Da der Korrelationskoeffizient relativ klein ist, scheint ein Modell beruhend auf einem linearen Zusammenhang zwischen Einkommen und Anzahl Jahre Schulbildung nicht angebracht zu sein.

Lösung 2.7

a) Erzeugen der Vektoren:

```
t.x <- (-10):10
t.x1 <- 0:10
t.y <- t.x^2
t.y1 <- t.x1^2
```

b)
```
par(mfrow = c(1, 2))   # zwei Grafiken im Grafikfenster
plot(t.x, t.y)
plot(t.x1, t.y1)
```



c)
```
cor(t.x, t.y)

## [1] 0
```

```r
cor(t.x1, t.y1)

## [1] 0.9631427
```

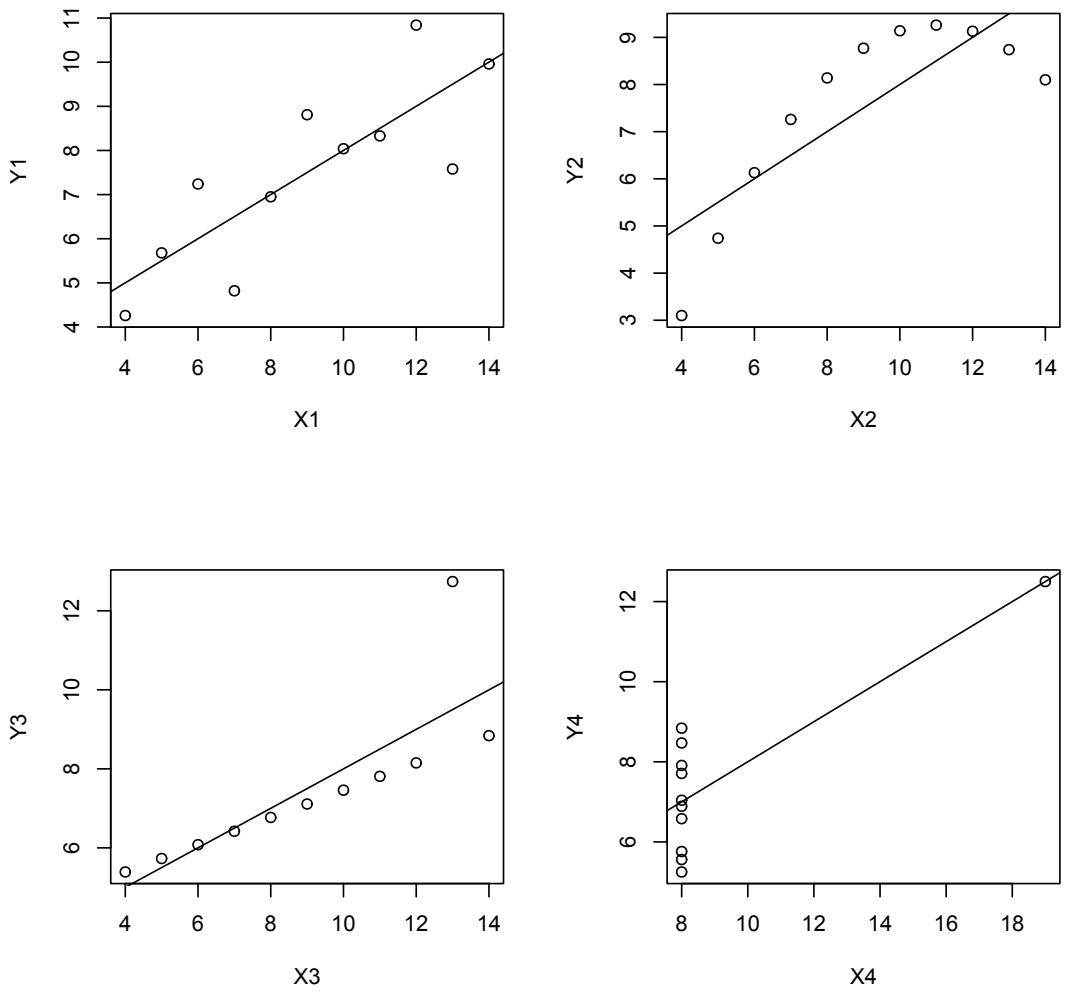
Die Korrelation zwischen `t.x` und `t.y` ist 0, weil die Daten symmetrisch zur y-Achse liegen.

Im zweiten Fall ist die Korrelation hoch (0.96), obwohl die Daten keine lineare Beziehung aufweisen. Der Grund dafür ist, dass x und y monoton steigen.

Lösung 2.8

a) Betrachtet man die vier Streudiagramme, so sieht man, dass nur im ersten Fall eine lineare Regression korrekt ist. Im zweiten Fall ist die Beziehung zwischen X und Y nicht linear, sondern quadratisch. Im dritten Fall gibt es einen Ausreisser, welcher die geschätzten Parameter stark beeinflusst. Im vierten Fall wird die Regressionsgerade durch einen einzigen Punkt bestimmt.

```r
data(anscombe)
reg <- lm(anscombe$y1 ~ anscombe$x1)
reg2 <- lm(anscombe$y2 ~ anscombe$x2)
reg3 <- lm(anscombe$y3 ~ anscombe$x3)
reg4 <- lm(anscombe$y4 ~ anscombe$x4)
par(mfrow = c(2, 2))
plot(anscombe$x1, anscombe$y1, ylab = "Y1", xlab = "X1")
abline(reg)
plot(anscombe$x2, anscombe$y2, ylab = "Y2", xlab = "X2")
abline(reg2)
plot(anscombe$x3, anscombe$y3, ylab = "Y3", xlab = "X3")
abline(reg3)
plot(anscombe$x4, anscombe$y4, ylab = "Y4", xlab = "X4")
abline(reg4)
```



b) Bei allen vier Modellen sind die Schätzungen des Achsenabschnitts β_0 und der Steigung β_1 fast identisch:

	Modell 1	Modell 2	Modell 3	Modell 4
Achsenabschnitt ($\widehat{\beta_0}$)	3.000	3.001	3.002	3.002
Steigung ($\widehat{\beta_1}$)	0.500	0.500	0.500	0.500

Fazit: Es genügt **nicht**, nur $\widehat{\beta_0}$ und $\widehat{\beta_1}$ anzuschauen. In allen Modellen sind diese Schätzungen fast gleich, aber die Datensätze sehen ganz unterschiedlich aus. Eine (graphische) Überprüfung der Modellannahmen ist also unumgänglich.