Applied Statistics for Data Science

Serie 3

Aufgabe 3.1

Der Geysir Old Faithful im Yellowstone National Park ist eine der bekanntesten heissen Quellen. Für die Zuschauer und den Nationalparkdienst ist die Zeitspanne zwischen zwei Ausbrüchen und die Eruptionsdauer von grossem Interesse. Auf Ilias sind die Messungen in der Datei **geysir.dat** vom 1.8.1978 - 8.8.1978 in 3 Spalten abgelegt: **Tag**, **Zeitspanne** und **Eruptionsdauer**.

a) Zeichnen Sie Histogramme von der Zeitspanne zwischen zwei Ausbrüchen:

```
# Datensatz einlesen
geysir <- read.table("./Daten/geysir.dat", header = TRUE)

# 4 Graphiken im Graphikfenster
par(mfrow = c(2,2))

hist(geysir[, "Zeitspanne"])
hist(geysir[, "Zeitspanne"], breaks = 20)
hist(geysir[, "Zeitspanne"], breaks = seq(41, 96, by = 11))
```

Was fällt auf? Was ist der Unterschied zwischen den drei Histogrammen?

Bemerkung: Wenn man die Anzahl Klassen mit **breaks = 20** vorgibt, so wird dies nur als „Vorschlag“ interpretiert und intern unter Umständen abgeändert.

b) Zeichnen Sie Histogramme (Anzahl Klassen variieren) von der Eruptionsdauer:

```
hist(geysir[, "Eruptionsdauer"])
```

Was fällt auf? Vergleichen Sie mit der ersten Teilaufgabe.

Aufgabe 3.2

In Aufgabe 2.4 hatten wir den Alterunterschied zwischen Ehemännern und Ehefrauen untersucht.

Nun wollen wir nun untersuchen, ob grosse Frauen auch grosse Männer heiraten.

a) Lesen Sie die Datei **`mannfrau.csv`** ein.

b) Erzeugen Sie das Streudiagramm aus **`groesse.mann`** und **`groesse.frau`** mit der Regressionsgerade. Sehen Sie gegebenfalls für die Befehle im Skript/Slides nach.

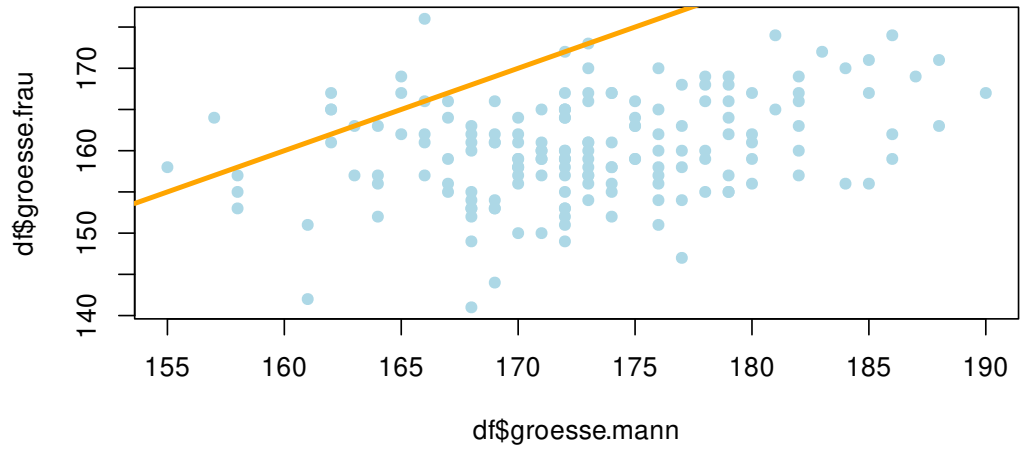
   Interpretieren Sie das Streudiagramm.

c) Bestimmen Sie die Koeffizienten der Regressionsgeraden

$$y = a + bx$$

   und interpretieren Sie diese Werte.

d) Im Streudiagramm in Abbildung unten ist die Gerade $y = x$, also **`groesse.mann`** gleich **`groesse.frau`** orange eingezeichnet.



   Wie interpretieren Sie diese Gerade im Streudiagramm?

Aufgabe 3.3

Wir betrachten eine Studie, die 1979 in den Vereinigten Staaten durchgeführt wurde (National Longitudinal Study of Youth, NLSY79): von 2584 Amerikanern im Jahr 1981 wurde der Intelligenzquotient (gemäss AFQT - armed forces qualifying test score) gemessen; 2006 wurden dieselben Personen nach ihrem jährlichen Einkommen im Jahr 2005 und der Anzahl Jahre Schulbildung befragt. Uns interessiert hier natürlich, ob ein hoher IQ oder eine lange Schulbildung zu einem höheren Einkommen führen. In der auf Ilias abgelegten Datei **`income.dat`** finden Sie den Datensatz mit dem Einkommen, der Anzahl Jahre abgeschlossener Schulbildung und den ermittelten Intelligenzquotienten von 2584 Amerikanern.

a) Lesen Sie den Datensatz **income.dat** ein und generieren Sie Streudiagramme, in welchen das Einkommen versus Anzahl Jahre Schulbildung und Einkommen versus Intelligenzquotient aufgetragen sind.

b) Bestimmen Sie die Parameter a und b des linearen Modells $y = a + bx$, wobei y das Einkommen bezeichnet und x die Anzahl Jahre Schulbildung. Zeichnen Sie die Regressionsgerade mit der **R**-Funktion

```
plot(...,..., type="l")
```

Wie interpretieren Sie die Parameter a und b?

Aufgabe 3.4

In dieser Aufgabe betrachten wir 4 Datensätze, die von Anscombe konstruiert wurden. In jedem der Datensätze gibt es eine Zielvariable y und eine erklärende Variable x.

a) Die Datei ist **R** schon enthalten.

```
head(anscombe)

##   x1 x2 x3 x4   y1   y2    y3   y4
## 1 10 10 10  8 8.04 9.14  7.46 6.58
## 2  8  8  8  8 6.95 8.14  6.77 5.76
## 3 13 13 13  8 7.58 8.74 12.74 7.71
## 4  9  9  9  8 8.81 8.77  7.11 8.84
## 5 11 11 11  8 8.33 9.26  7.81 8.47
## 6 14 14 14  8 9.96 8.10  8.84 7.04
```

b) StellenSie jeden der 4 Datensätze als Streudiagramm dar, zeichnen Sie die Regressionsgerade ein und kommentieren Sie die Ergebnisse.

```
plot(anscombe$x1, anscombe$y1)
reg <- lm(anscombe$y1 ~ anscombe$x1)
abline(reg)
```

Mit **par(mfrow=c(2,2))** wird das Grafikfenster so eingeteilt, dass alle 4 Bilder nebeneinander passen.

c) Vergleichen Sie jeweils a und b, wobei $y = a + bx$.

```
lm(y1 ~ x1, data = anscombe) # oder

lm(anscombe$y1 ~ anscombe$x1)
```

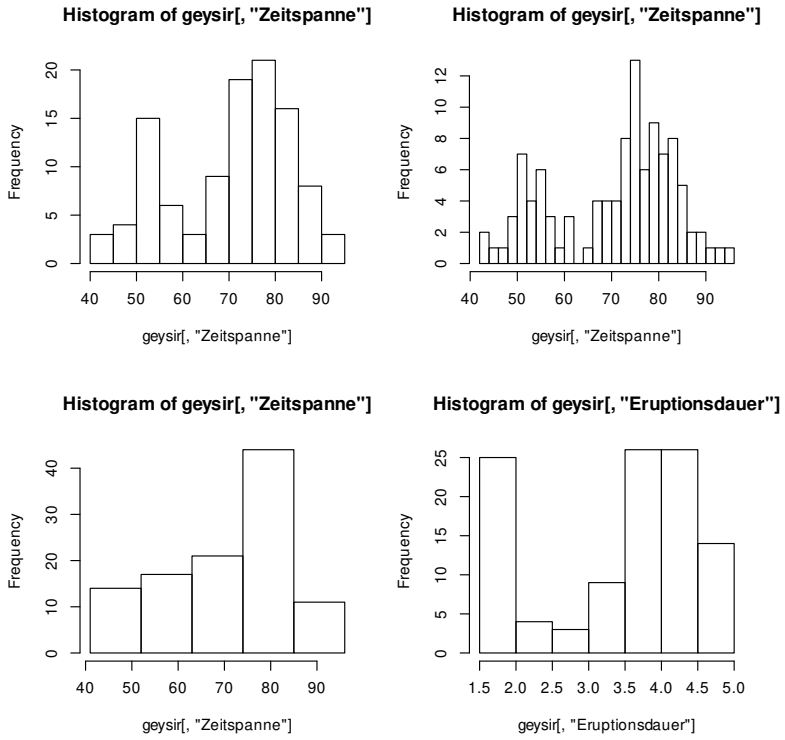
Applied Statistics for Data Science

Musterlösungen zu Serie 3

Lösung 3.1

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 Minuten 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.

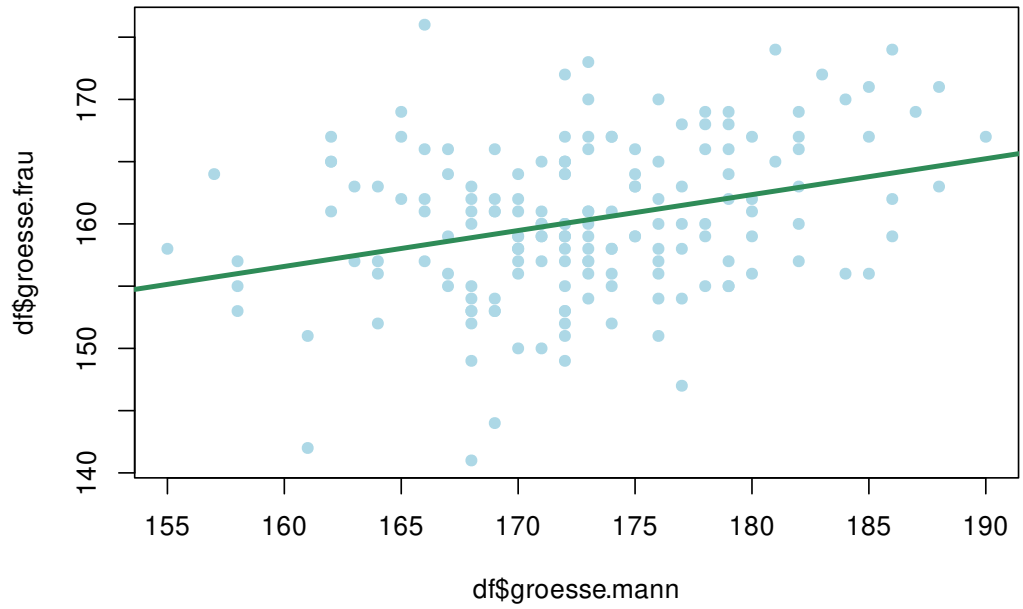
b) Das vierte Histogramm schliesslich zeigt die Häufigkeiten verschiedener Eruptionsdauern. Hier sind die beiden Gipfel sehr deutlich erkennbar: „Entweder ist der Ausbruch sofort wieder vorbei, oder er dauert mindestens dreieinhalb Minuten“.

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.

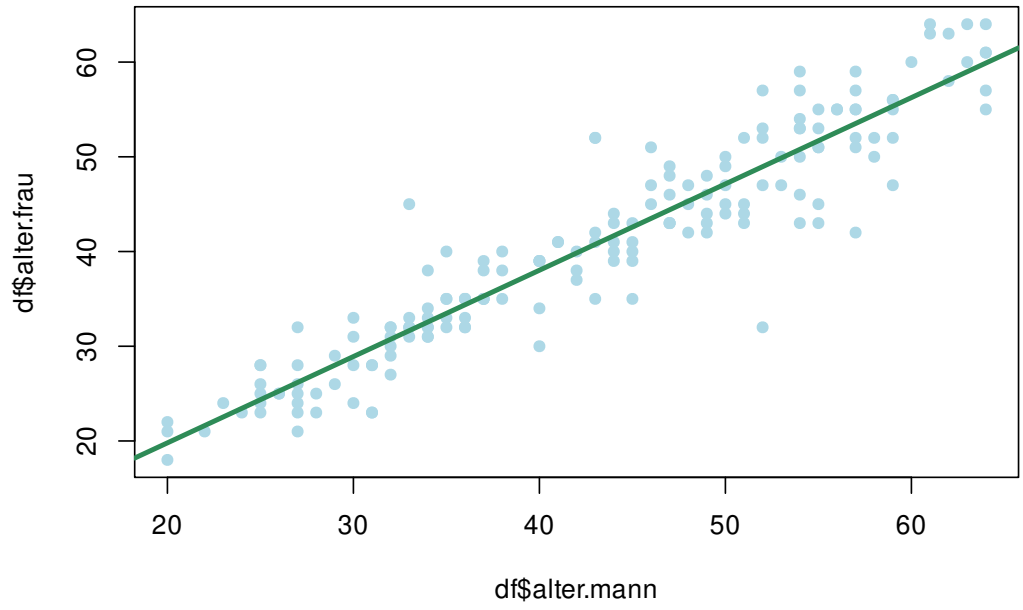
Lösung 3.2

a)
```
df <- read.csv("~/Dropbox/Statistics/Themen/Deskriptive_Statistik/Uebungen_de/Date
head(df)

##   alter.mann groesse.mann alter.frau groesse.frau
## 1         49          180         43          159
## 2         25          184         28          156
## 3         40          165         30          162
## 4         52          177         57          154
## 5         58          161         52          142
## 6         32          169         27          166
```

b)
```
plot(df$groesse.mann,
     df$groesse.frau,
     col = "lightblue",
     pch = 16)
abline(lm(df$groesse.frau~df$groesse.mann),
       lwd = 3,
       col = "seagreen")
```




```
plot(df$alter.mann,
     df$alter.frau,
     col = "lightblue",
     pch = 16)
abline(lm(df$alter.frau~df$alter.mann),
       lwd = 3,
       col = "seagreen")
```



Die Punkte sind recht zerstreut, aber leicht steigend. Das heisst es gibt eine leichte Tendenz, dass kleine Männer eher kleine Frauen heiraten und grosse Männer

eher grosse Frauen.

c) Die Regressionsgerade ist grün eingezeichnet. Geben Sie die Gleichung

$$y = a + bx$$

mit Hilfe des folgenden **R**-Outputs an und interpretieren Sie diese Werte.

```
lm(df$groesse.frau~df$groesse.mann)

##
## Call:
## lm(formula = df$groesse.frau ~ df$groesse.mann)
##
## Coefficients:
##     (Intercept)  df$groesse.mann
##        110.4440          0.2884
```

```
lm(df$alter.frau~df$alter.mann)

##
## Call:
## lm(formula = df$alter.frau ~ df$alter.mann)
##
## Coefficients:
##   (Intercept)  df$alter.mann
##        1.5740         0.9112
```

Der **Intercept** (y-Achsenabschnitt) ist $a = 110.44$ und die Steigung ist $b = 0.29$, also

$$y = 110.44 + 0.29x$$

Wäre der Ehemann 0 cm gross, so hätte er nach diesem Modell eine Ehefrau, die 110 cm gross ist. Dies ist natürliche eine absurde Interpretation.

Für jeden zusätzlichen Zentimeter Körpergrösse des Mannes, ist die Frau um 0.3 cm grösser.

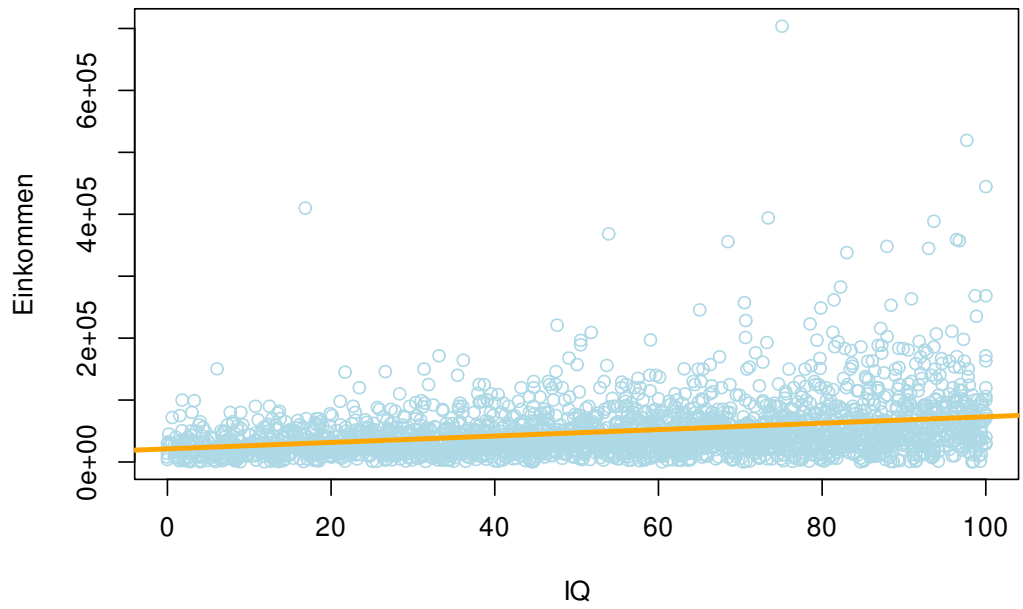
d) Die Punkte über der Geraden sind diejenigen, wo die Frau grösser ist als der Mann. Die sind sehr wenige Punkte. Diese Beobachtung ist aber nicht sonderlich überraschend, da die Männer generell grösser sind als die Frauen.

Dies hat nichts damit zu tun, ob grosse Frauen auch grosse Männer heiraten oder nicht.

Lösung 3.3

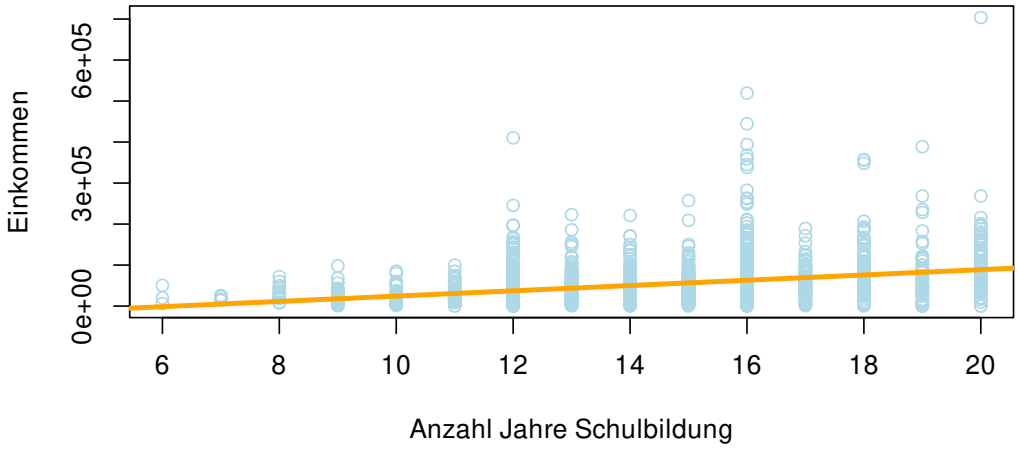
a)
```
income <- read.table(file="./Daten/income.dat", header=TRUE)
head(income)
```

```
##      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",
     col = "lightblue"

)

abline(lm(einkommen ~ iq),
       col = "orange",
       lwd = 3)
```




```
plot(anzahl_jahre_schule,
     einkommen,
     type="p",
```

```
     xlab = "Anzahl Jahre Schulbildung",
     ylab="Einkommen",
     col = "lightblue"

)

abline(lm(einkommen ~ anzahl_jahre_schule),
       col = "orange",
       lwd = 3)
```



In beiden Fällen ist die Regressionsgerade sehr flach und die Punkte streuen ziemlich um die Regressionsgerade.

b) Mit **R** ermitteln wir für a und b

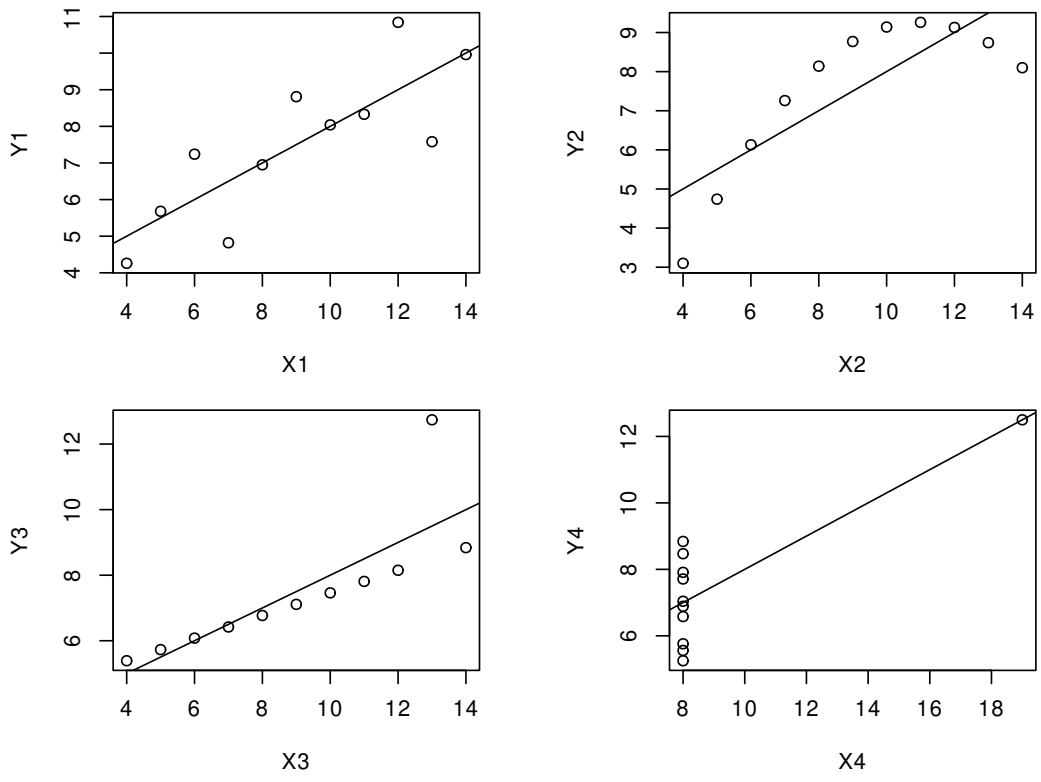
```
lm(einkommen ~ anzahl_jahre_schule)

##
## Call:
## lm(formula = einkommen ~ anzahl_jahre_schule)
##
## Coefficients:
##         (Intercept)  anzahl_jahre_schule
##               -40200                 6451
```

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.

Lösung 3.4

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.

```
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 (a)	3.000	3.001	3.002	3.002
Steigung (β_1)	0.500	0.500	0.500	0.500

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