Arbeitsblatt 1

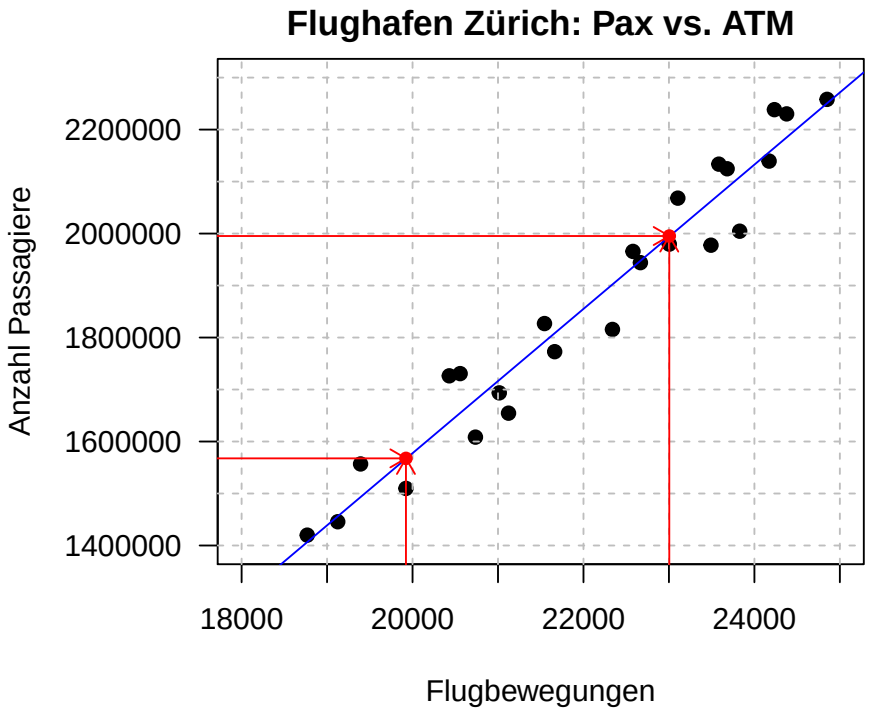
Einfache lineare Regression

Aufgabe 1 (Flughafen Zürich)

Im nachfolgenden Streudiagramm sind die monatlichen Anzahl Flugbewegungen (ATM) und Passagiere (Pax) für die Jahre 2010 und 2011 gegeneinander aufgezeichnet.

(a)

Zeichnen Sie in das Streudiagramm intuitiv ("nach Augenmass") eine Gerade ein, die "gut zu den Daten passt", d.h. die mutmassliche Regressionsgerade.



(b)

Als Schätzungen für die Regressionskoeffizienten erhält man für diese Datenpunkte mittels Kleinste-Quadrate:

$$\hat{\alpha} = -1197682.1, \quad \hat{\beta} = 138.8$$

Das Ziel ist es nun, die rechnerisch bestimmte Regressionsgerade in das Streudiagramm einzutragen. Lösen Sie dazu die folgenden Teilschritte:

- Bestimmen Sie den angepassten Wert $\hat{y}$ für $x = 19923$ Flugbewegungen (Januar 2009).

```
-1197682.1+138.8*19923
```

```
[1] 1567630
```

- Bestimmen Sie den angepassten Wert $\hat{y}$ nun auch für $x = 23004$ Flugbewegungen (September 2009).

```
-1197682.1+138.8*23004
```

```
[1] 1995273
```

- Verbinden Sie die beiden Punkte miteinander, um die Regressionsgerade zu erhalten. Vergleichen sie die Gerade mit der "von Auge" eingezeichneten Lösung.

(c)

Die Daten stehen Ihnen im Ordner `Daten` im File `flughafen.rda` auf Moodle zur Verfügung. Laden Sie die Daten in R, passen Sie die Regressionsgerade an und vergleichen Sie, ob Sie dieselben Koeffizieten wie in (b) erhalten.

```
load("data/flughafen.rda")
fit.zrh <- lm(Pax ~ ATM, data = zrh)
coef(fit.zrh)
```

```
  (Intercept)           ATM
-1197682.0740      138.7617
```

(d)

Welche Anzahl Passagiere gibt es laut Modell, falls keine Flugbewegungen durchgeführt werden (ATM = 0)? Halten Sie diese Anzahl für plausibel? Können Sie sich die Anzahl erklären?

Die Anzahl Passagiere bei keiner Flugbewegung entspricht dem geschätzten Achsenabschnitt:

```
coef(fit.zrh)[1]
```

```
(Intercept)
   -1197682
```

Der geschätzte Achsenabschnitt ist negativ und somit vollkommen unsinnig. Es zeigt auf, dass die Regressionsgerade nicht (weit) über den Bereich der beobachteten x-Werte (18768, 24850) hinaus gültig ist.

Aufgabe 2 (Antike Uhren)

McClave und Benson haben Daten über das Alter (in Jahre) und den Preis (in US$) von antiken Uhren an Auktionen zusammengetragen. Sie stehen Ihnen im Ordner `Daten` im File `AntikeUhren.dat` auf Moodle zur Verfügung.

Einlesen der Daten

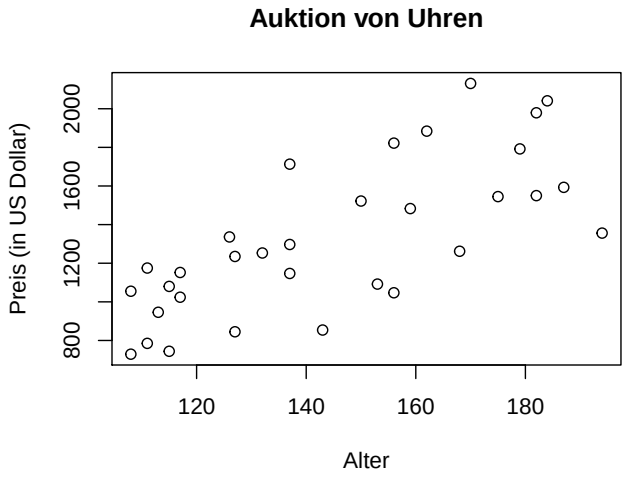
```
au <- read.table("data/AntikeUhren.dat", header=TRUE)
```

(a)

Stellen Sie die Daten in einem Streudiagramm Preis (y-Achse) gegen Alter (x-Achse) dar und beschreiben Sie den funktionalen Zusammenhang in Worten.

Es gibt 2 Möglichkeiten, ein Streudiagramm zu erstellen

```
plot(x=au$Alter, y=au$Preis, xlab = "Alter", ylab = "Preis (in US Dollar)",
     main = "Auktion von Uhren")
```




```
# Alternativ:
# plot(Preis ~ Alter, data=au)
```

Die Daten streuen sehr stark. Trotzdem ist ein linearer Trend sichtbar: je älter die Uhr, desto teurer ist sie.

(b)

Passen Sie eine Gerade an die Datenpunkte an. Geben Sie die geschätzten Koeffizientenwerte an. Wie lautet die angepasste Geradengleichung?

```
au.fit <- lm(Preis ~ Alter, data = au)
```

Die beiden geschätzten Koeffizientenwerte sind:

```
coef(au.fit)
```

```
(Intercept)       Alter
 -191.65757    10.47909
```

Die angepasste Geradengleichung ist $\widehat{\texttt{Preis}} = -191.66 + 10.48 \cdot \texttt{Alter}$.

(c)

Welche Auswirkung hat eine um ein Jahr ältere Uhr auf den erwarteten Auktionspreis?

Der Preis einer um ein Jahr älteren Uhr entspricht der geschätzten Steigung der Geraden:

```
coef(au.fit)[2]
```

```
   Alter
10.47909
```

Der Auktionspreis nimmt also um 10.48 US Dollar zu.

(d)

Wie gross ist der Standardfehler der Residuen? Was gibt er an?

```
# Standardfehler
summary(au.fit)$sigma
```

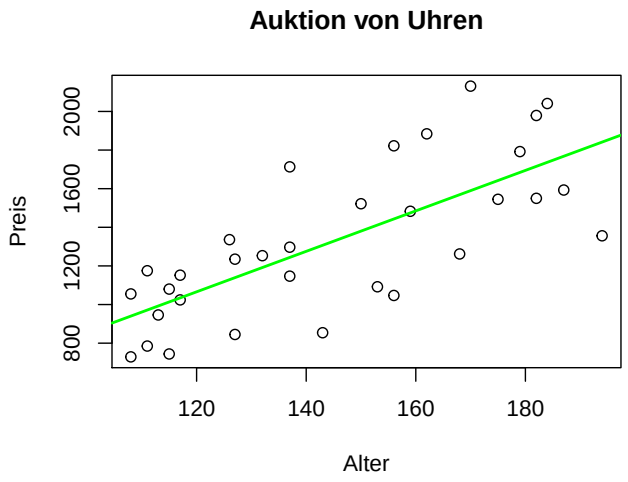
```
[1] 273.0284
```

Der Standardfehler der Residuen gibt an, wie stark die Beobachtungen um die Regressionsgerade streuen. Falls die Modellannahmen erfüllt sind, so können wir davon ausgehen, dass sich rund 95% der Punkte in einem Intervall von $\pm 2 \cdot 273$ US Dollar um die Regressionsgerade befinden.

(f)

Zeichnen Sie die Gerade in das Streudiagramm von Teilaufgabe (a) ein. Kommentieren Sie die Lösung.

```
plot(au$Alter, au$Preis, xlab = "Alter", ylab = "Preis",
     main = "Auktion von Uhren")
abline(au.fit, col = "green", lwd=2)
```




```
## Alternativ:
# abline(a = coef(au.fit)[1], b = coef(au.fit)[2])
```

Die Gerade liegt schön in den Daten, wie man es erwarten würde.

Aufgabe 3 (Gotthard Strassentunnel)

Wir betrachten in dieser Aufgabe einen Datensatz, welcher über die Jahre 2004-2016 die Anzahl Tage mit Stau vor dem Gotthard Strassentunnel Nordportal beschreibt. Lesen sie diese Daten in R (`gotthard.rda`).

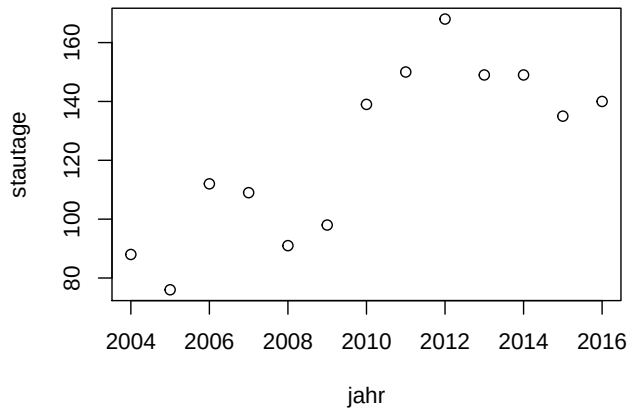
Einlesen der Daten

```
load("data/gotthard.rda")
```

(a)

Stellen Sie die Daten in einem Streudiagramm dar. Gibt es einen Trend?

```
plot(stautage ~ jahr, data = gotthard)
```



Die Daten streuen sehr stark. Es ist ein positiver Trend sichtbar bis etwa 2012, danach ist sie aber eher konstant, beziehungsweise sogar leicht abnehmend.

(b)

Passen Sie eine Gerade an die Datenpunkte an. Geben Sie die geschätzten Werte der beiden Koeffizienten an.

```
gs.fit <- lm(stautage ~ jahr, data = gotthard)
```

Die beiden geschätzten Koeffizientenwerte sind:

```
coef(gs.fit)
```

```
 (Intercept)         jahr
-11815.13187      5.93956
```

(c)

Wie viele Stautage werden vom Modell für 2016 geschätzt? Was ist das Residuum für diesen Datenpunkt?

```
fitted(gs.fit)[gotthard$jahr == 2016]
```

```
     13
159.022
```

```
resid(gs.fit)[gotthard$jahr == 2016]
```

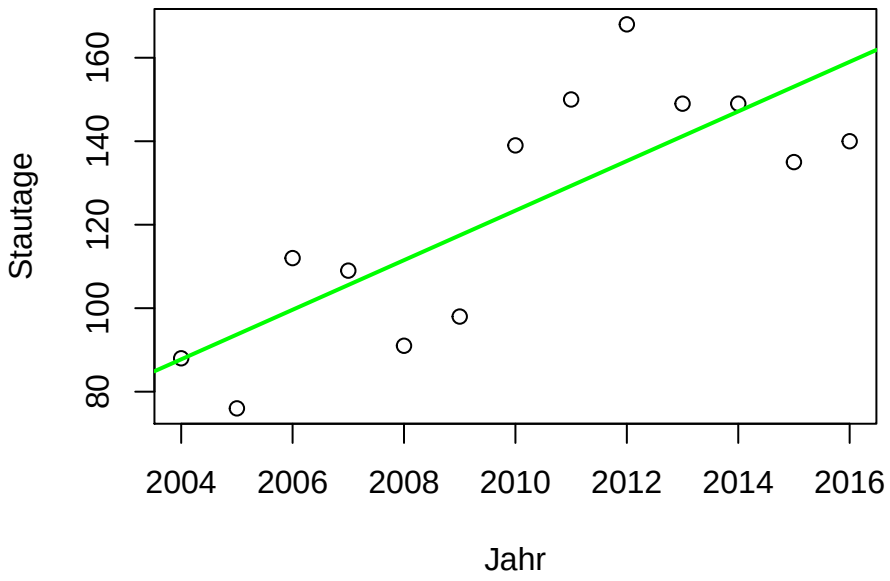
```
       13
-19.02198
```

Mit dem Modell erhalten wir eine Schätzung von 160 Stautagen für 2016. Das Residuum ist -19 Tage, d.h. das Modell überschätzt hier die Anzahl Tage.

(d)

Zeichnen Sie die Gerade in das Streudiagramm von Teilaufgabe (a) ein. Halten Sie das lineare Regressionsmodell für plausibel?

```
plot(gotthard$jahr, gotthard$stautage, xlab = "Jahr", ylab = "Stautage")
abline(gs.fit, col = "green", lwd=2)
```



Das lineare Regressionsmodell ist hier eher ungeeignet. Bis 2012 scheint die Anzahl Stautage zwar linear anzusteigen, danach ist sie aber eher konstant, beziehungsweise sogar leicht abnehmend. Wie wir testen, ob die lineare Beziehung plausibel ist, werden wir in Woche 3 kennenlernen.