Lineare Modelle – SoSe 2025 **Übungsblatt 1**

Sabine Hoffmann, Christina Sauer, Cornelia Gruber

Aufgabe 1: Lineare Einfachregression

Der auf der Moodle-Homepage verfügbare Datensatz `RunningAgg.Rds` enthält Informationen zu einstündigen Laufeinheiten von 345 Hobbyläufern. Für die einzelnen Läufer sind die durchschnittliche Pace) (`pace`, Kehrwert der Geschwindigkeit) in min/km sowie die zugehörige durchschnittliche Herzfrequenz (`HR`) in *bpm* (beats per minute) gegeben.

(a) Lesen Sie den Datensatz in **R** ein und visualisieren Sie den Zusammenhang zwischen der Laufgeschwindigkeit (`pace`) und der Herzfrequenz (`HR`). Passen sie ein Lineares Modell für diesen Zusammenhang an und interpretieren Sie die Parameterschätzer $\hat{\beta}_0$ und $\hat{\beta}_1$.

Lösung:

- Einlesen der Daten und Überblick über Datensatz:

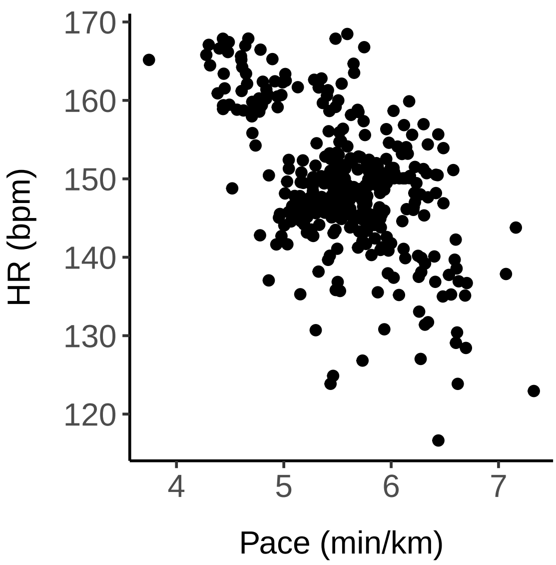
```
run <- readRDS("../Daten/RunningAgg.Rds")
head(run)

## # A tibble: 6 x 2
##    pace    HR
##   <dbl> <dbl>
## 1  5.44  124.
## 2  5.03  142.
## 3  5.28  146.
## 4  5.27  143.
## 5  5.00  144.
## 6  5.31  147.

str(run)

## tibble [345 x 2] (S3: tbl_df/tbl/data.frame)
##  $ pace: num [1:345] 5.44 5.03 5.28 5.27 5 ...
##  $ HR  : num [1:345] 124 142 146 143 144 ...

summary(run)

##       pace              HR
##  Min.   :3.744   Min.   :116.6
##  1st Qu.:5.254   1st Qu.:145.1
##  Median :5.546   Median :148.8
##  Mean   :5.564   Mean   :149.2
##  3rd Qu.:5.912   3rd Qu.:154.1
##  Max.   :7.329   Max.   :168.5
```

⇒ Der Datensatz enthält 345 Beobachtungen sowie die beiden interessierenden Variablen Pace (`pace`) und Herzfrequenz (`HR`).

⇒ Der Zusammenhang zwischen der Pace und der Herzfrequenz kann durch ein Streudiagramm visualisiert werden.

- Streudiagramm zwischen Pace und Herzfrequenz:

```
gg_pace <- ggplot(data = run, mapping = aes(x = pace, y = HR)) + geom_point() +
  xlab("Pace (min/km)") + ylab("HR (bpm)") + theme
gg_pace
```



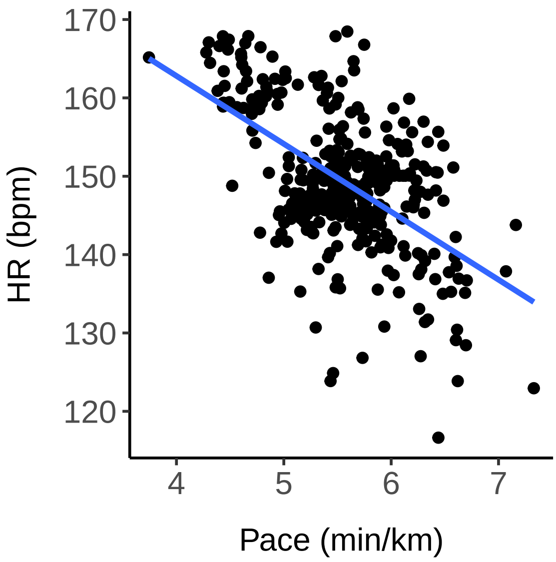
- Anpassung eines linearen Regressionsmodells zwischen Pace und Herzfrequenz mit folgenden Annahmen:
  - Strukturannahme: $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad$ für $i = 1, ..., n$
  - Verteilungsannahme: $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2) \quad$ für $i = 1, ..., n$

```
lm_pace <- lm(formula = HR ~ pace, data = run)
summary(lm_pace)

##
## Call:
## lm(formula = HR ~ pace, data = run)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.4970  -4.5080  -0.1373   5.0223  19.4594
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 197.4826     3.9167   50.42   <2e-16 ***
## pace         -8.6698     0.7004  -12.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.341 on 343 degrees of freedom
## Multiple R-squared:  0.3088, Adjusted R-squared:  0.3068
## F-statistic: 153.2 on 1 and 343 DF,  p-value: < 2.2e-16

beta_pace <- round(x = coef(lm_pace), digits = 2)
beta_pace

## (Intercept)        pace
##      197.48       -8.67
```

  - Interpretation von $\hat{\beta}_0$: Bei einer Pace von 0 min/km beträgt die durchschnittliche/geschätzte/erwartete Herzfrequenz 197.48 bpm.
  - Interpretation von $\hat{\beta}_1$: Erhöht sich die Pace um 1 min/km, so verringert sich die Herzfrequenz durchschnittlich um 8.67 bpm.

- Streudiagramm mit geschätzter Regressionsgerade:

```r
gg_pace + geom_smooth(method = "lm", se = FALSE)
```



(b) Warum könnte es sinnvoller sein, den Zusammenhang zwischen der Geschwindigkeit (in km/h) und der Herzfrequenz zu untersuchen? Passen Sie das Modell mit der neuen Einflussgröße erneut an. Vergleichen Sie die Anpassung der beiden Modelle.

Lösung:

Es könnte sinnvoller sein, den Zusammenhang zwischen der Geschwindigkeit und der Herzfrequenz zu modellieren, da mit der Pace als abhängiger Variable der Intercept $\hat{\beta}_0$ nicht interpretierbar ist. Mit der Geschwindigkeit als abhängiger Variable kann der Koeffizient dagegen als Ruheherzfrequenz interpretiert werden.

- Schätzung eines linearen Regressionsmodells zwischen der Geschwindigkeit und der Herzfrequenz:

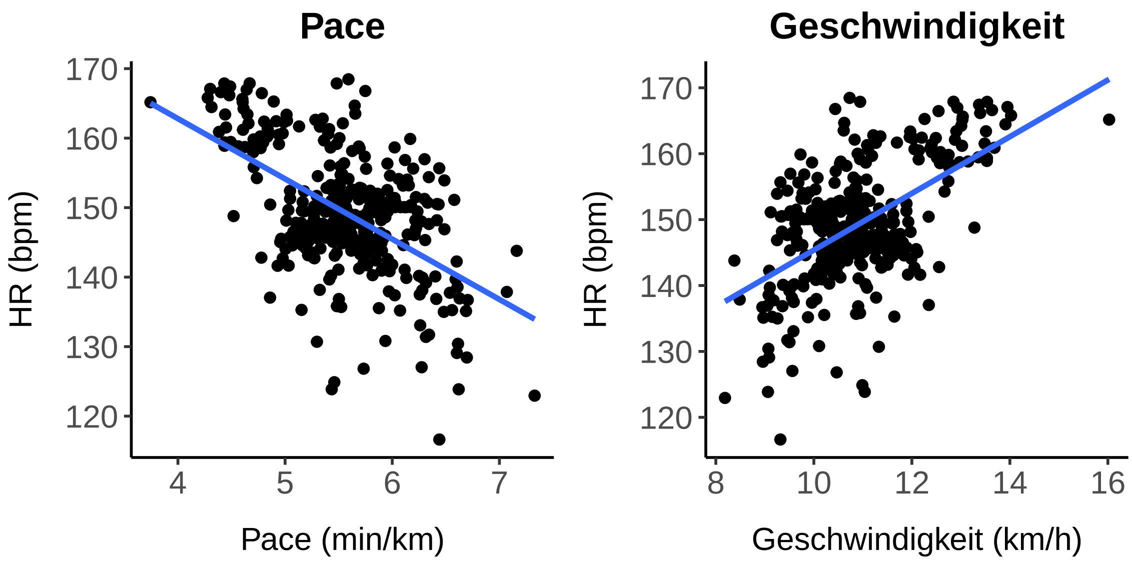
```r
run$speed <- 60 / run$pace
lm_speed <- lm(formula = HR ~ speed, data = run)
summary(lm_speed)

##
## Call:
## lm(formula = HR ~ speed, data = run)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.9834  -4.2382   0.0394   4.8816  19.9451
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  102.370      3.748   27.31   <2e-16 ***
## speed          4.301      0.342   12.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.305 on 343 degrees of freedom
## Multiple R-squared:  0.3156, Adjusted R-squared:  0.3136
## F-statistic: 158.1 on 1 and 343 DF,  p-value: < 2.2e-16

beta_speed <- round(x = coef(lm_speed), digits = 2)
```

```
beta_speed

## (Intercept)       speed
##      102.37        4.30
```

$\Rightarrow \ \hat{\beta}_0^{speed} = 102.37bpm, \ \hat{\beta}_1^{speed} = 4.30bpm$

- Graphische Visualisierung der beiden Regressionsmodelle:

```
# Pace und Herzfrequenz:
gg_original <- gg_pace + geom_smooth(method = "lm", se = FALSE) +
  ggtitle("Pace") + theme
# Geschwindigkeit und Herzfrequenz:
gg_transformed <- gg_original %+% run + aes(x=speed) +
  ggtitle("Geschwindigkeit") + xlab("Geschwindigkeit (km/h)") + theme
# Gemeinsame Visualisierung:
grid.arrange(gg_original, gg_transformed, nrow = 1)
```



- Modellanpassung:
  Die Anpassung an die Daten ist für zweites Modell geringfügig besser ($R^2{}_{pace} = 0.31$ vs. $R^2{}_{speed} = 0.32$).

(c) Leiten Sie die Umrechnungsformel für die Kleinste-Quadrate-Schätzer nach einer linearen Variablentransformation der folgenden Form her:

$$
\begin{aligned}
x_i \to t_i &= a_0 + a_1 x_i, \quad \text{mit } a_1 \neq 0 \\
y_i \to u_i &= b_0 + b_1 y_i, \quad \text{mit } b_1 \neq 0
\end{aligned}
$$

Lösung:

- Gegebene Variablentransformation (für $i = 1, ..., n$):

$$
\begin{aligned}
x_i \to t_i &= a_0 + a_1 x_i \\
y_i \to u_i &= b_0 + b_1 y_i
\end{aligned}
$$

$\Rightarrow$ Gesucht: Kleinste-Quadrate-Schätzer für Modell $u_i = \hat{\alpha}_0 + \hat{\alpha}_1 t_i + \epsilon_i$

- Kleinste-Quadrate-Schätzer für Modell $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$:

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\hat{Cov}(X,Y)}{\hat{Var}(X)}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

- Berechnung von $\hat{\alpha}_1$:

$$\hat{\alpha}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(t_i-\bar{t})(u_i-\bar{u})}{\frac{1}{n}\sum_{i=1}^{n}(t_i-\bar{t})^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(a_0+a_1x_i-a_0-a_1\bar{x})(b_0+b_1y_i-b_0-b_1\bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(a_0+a_1x_i-a_0-a_1\bar{x})^2}$$
$$= \frac{a_1b_1}{a_1{}^2}\cdot\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{b_1}{a_1}\hat{\beta}_1$$

- Berechnung von $\hat{\alpha}_0$:

$$\hat{\alpha}_0 = \bar{u} - \hat{\alpha}_1\bar{t} = b_0 + b_1\bar{y} - \frac{b_1}{a_1}\hat{\beta}_1(a_0+a_1\bar{x}) = b_0 + b_1\bar{y} - b_1\hat{\beta}_1\bar{x} - \frac{b_1}{a_1}\hat{\beta}_1 a_0$$
$$= b_0 + b_1(\bar{y}-\hat{\beta}_1\bar{x}) - \frac{b_1}{a_1}\hat{\beta}_0 a_0 = b_0 + b_1\hat{\beta}_0 - \frac{b_1}{a_1}a_0\hat{\beta}_1$$

(d) Wie ändern sich die Parameterschätzungen, wenn die Geschwindigkeit in Meilen pro Stunde ($1\ mi = 1.61\ km$) und die Herzfrequenz in Schläge pro Sekunde (bps) umgerechnet werden? Berechnen Sie hierzu die neuen Schätzungen anhand der Ergebnisse aus Teilaufgabe (b). Hätten hierfür auch die Ergebnisse aus Teilaufgabe (a) verwendet werden können?

Lösung:

- Umrechnungsfaktoren:
  - Geschwindigkeit: $a_1 = \frac{1}{1.61}, a_0 = 0$
  - Herzfrequenz: $b_1 = \frac{1}{60}, b_0 = 0$

- Parameterschätzer:

$$\hat{\alpha}_1 = \frac{\frac{1}{60}}{\frac{1}{1.61}}\cdot 4.3 = \frac{1.61}{60}\cdot 4.3 = 0.115$$
$$\hat{\alpha}_0 = \frac{1}{60}\cdot 102.37 = 1.706$$

$\Rightarrow$ Die Ergebnisse aus Teilaufgabe (a) hätten nicht verwendet werden können, da die Umrechnung von Pace in Geschwindigkeit keine lineare Transformation darstellt.

- Schätzung eines linearen Regressionsmodells zwischen Geschwindigkeit und Herzfrequenz auf transformierten Skalen:

```
# Berechnung der transformierten Variable:
run <- run %>% mutate(HRbps = HR / 60, speedMi = speed / 1.61)

# Schätzung eines linearen Regressionsmodells zwischen der Geschwindidkeit und
# der Herzfrequenz auf transformierten Skalen:
lm_trafo <- lm(formula = HRbps ~ speedMi, data = run)
summary(lm_trafo)

##
## Call:
## lm(formula = HRbps ~ speedMi, data = run)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.43306 -0.07064  0.00066  0.08136  0.33242
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.706160   0.062471   27.31   <2e-16 ***
## speedMi     0.115405   0.009177   12.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1218 on 343 degrees of freedom
## Multiple R-squared:  0.3156, Adjusted R-squared:  0.3136
## F-statistic: 158.1 on 1 and 343 DF,  p-value: < 2.2e-16

# Vergleich des Modelloutputs mit Teilaufgabe (b):
a1 <- 1 / 1.61
b1 <- 1 / 60
(b1 / a1) * coefficients(lm_speed)[2]

##     speed
## 0.1154045

b1 * coefficients(lm_speed)[1]

## (Intercept)
##     1.70616

coefficients(lm_trafo)

## (Intercept)     speedMi
##   1.7061602   0.1154045
```

(e) Betrachten Sie nun erneut das Modell aus Teilaufgabe (b). Zentrieren Sie die Einflussvariable `speed` und passen Sie das Modell mit der zentrierten Variable neu an. Interpretieren Sie die geschätzten Parameter.

Lösung:

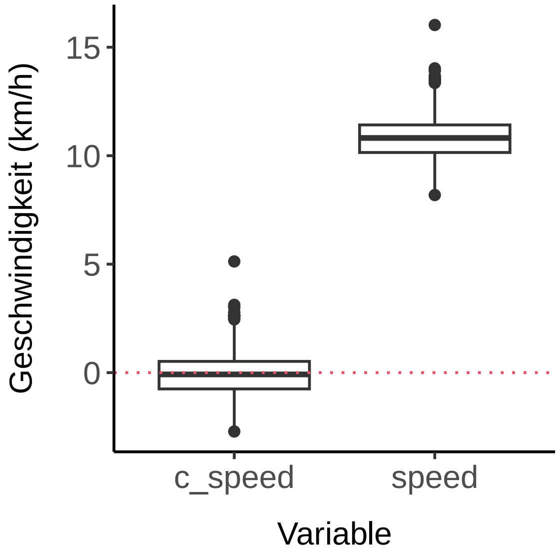
- Zentrieren der Einflussgröße:

  Die zentrierte Variable $\tilde{x}$ ergibt sich aus der ursprünglichen Variable durch $\tilde{x} = x - \bar{x}$:

```
mean(run$speed)

## [1] 10.89935

# Identisch zu run£speed - mean(run£speed), nur effizientere Berechnung
run <- run %>% mutate(c_speed = scale(x = speed, center = TRUE, scale = FALSE))
```

- Graphischer Vergleich zwischen der zentrierten und der originalen Variable:

```
run %>% gather(variable, value, speed, c_speed) %>%
  ggplot(aes(x = variable, y = value)) +
  geom_boxplot() + geom_hline(yintercept=0, col=2, lty=3) +
  xlab("Variable") + ylab("Geschwindigkeit (km/h)") + theme
```



- Schätzung eines linearen Regressionsmodells mit zentrierter Einflussgröße:

```
lm_c_speed <- lm(formula = HR ~ c_speed, data = run)
summary(lm_c_speed)

##
## Call:
## lm(formula = HR ~ c_speed, data = run)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.9834  -4.2382   0.0394   4.8816  19.9451
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 149.2454     0.3933  379.48   <2e-16 ***
## c_speed       4.3008     0.3420   12.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.305 on 343 degrees of freedom
## Multiple R-squared:  0.3156, Adjusted R-squared:  0.3136
## F-statistic: 158.1 on 1 and 343 DF,  p-value: < 2.2e-16
```

  - Interpretation von $\hat{\beta}_0$: Bei einer zentrierten Geschwindigkeit von 0 km/h beträgt die erwartete Herzfrequenz 149.25 bpm.
  - Interpretation von $\hat{\beta}_1$: Erhöht sich die (zentrierte) Geschwindidkeit um 1 km/h, so erhöht sich die Herzfrequenz durchschnittlich um 4.3 bpm.

  $\Rightarrow$ $\hat{\beta}_1$ ist identisch zum Modell ohne Zentrierung (Teilaufgabe (b)).

(f) Berechnen Sie die Parameterschätzer nun direkt aus den Parameterschätzern des Modells aus Teilaufgabe (c). Vergleichen Sie dazu die geschätzten Parameter aus Teilaufgabe (a).

Lösung:

- Bei der Zentrierung der Einflussvariable liegt die folgende lineare Variablentransformation vor:

$$u_i = b_0 + b_1 y_i = y_i \text{ mit } b_0 = 0 \text{ und } b_1 = 1$$
$$t_i = a_0 + a_1 x_i = \widetilde{x_i} \text{ mit } a_0 = -\bar{x} \text{ und } a_1 = 1$$

- Berechnung der Parameterschätzer gemäß Teilaufgabe (c):

$$\hat{\alpha}_1 = \frac{b_1}{a_1}\hat{\beta}_1 = \hat{\beta}_1 = 4.3$$

$$\hat{\alpha}_0 = b_0 + b_1\hat{\beta}_0 - \frac{b_1}{a_1}a_0\hat{\beta}_1 = 102.37 - (-10.9) \cdot 4.3 = 149.24$$

(g) Betrachten Sie nun das Regressionsmodell $\widetilde{HR}_i = \gamma_0 + \gamma_1\ \widetilde{speed}_i\ + \varepsilon_i$, wobei $\widetilde{HR}$ und $\widetilde{speed}$ jeweils die standardisierten Versionen der entsprechenden Variablen darstellen. Vergleichen Sie den Koeffizientenschätzer $\hat{\gamma}_1$ mit dem Bravais-Pearson-Korrelationskoeffizienten zwischen den beiden Variablen. Was fällt dabei auf?

Lösung:

- Standardisierung der Variablen:

  Eine Variable $\tilde{x}$ ist standardisiert, wenn $\bar{\tilde{x}} = 0$ und $S_{\tilde{x}} = 1$ gelten:
  $\Rightarrow\ \tilde{x}_i = \frac{x_i - \bar{x}}{\sqrt{s_x^2}}$ mit $S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

- Überblick über standardisierte Variablen:

```
run <- run %>% mutate(s_speed = scale(speed), s_HR = scale(HR))
summary(select(run, s_speed, s_HR))

##      s_speed.V1            s_HR.V1
##  Min.   :-2.355281   Min.   :-3.698395
##  1st Qu.:-0.652302   1st Qu.:-0.472048
##  Median :-0.070721   Median :-0.044847
##  Mean   : 0.000000   Mean   : 0.000000
##  3rd Qu.: 0.451214   3rd Qu.: 0.552478
##  Max.   : 4.450954   Max.   : 2.180000
```

- Zusammenhang mit Bravais-Pearson-Korrelation:

$$\hat{\gamma}_1 = \frac{\text{Cov}(\widetilde{HR}, \widetilde{speed})}{\text{Var}(\widetilde{speed})} = \frac{\text{Cov}(\widetilde{HR}, \widetilde{speed})}{\sqrt{\text{Var}(\widetilde{HR})Var(\widetilde{speed})}} = \rho(\widetilde{HR}, \widetilde{speed}) = \rho(HR, speed)$$

  $\Rightarrow$ $\hat{\gamma}_1$ entspricht dem Bravais-Pearson-Korrelationskoeffizienten zwischen $\widetilde{HR}$ und $\widetilde{speed}$ bzw. zwischen HR und $speed$.

- Schätzung eines linearen Regressionsmodells mit standardisierten Variablen:

```
lm_s_speed <- lm(formula = s_HR ~ s_speed, data = run)
summary(lm_s_speed)

##
## Call:
## lm(formula = s_HR ~ s_speed, data = run)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.94694 -0.48068  0.00447  0.55366  2.26210
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -1.042e-14  4.461e-02    0.00        1    
## s_speed      5.617e-01  4.467e-02   12.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.8285 on 343 degrees of freedom
## Multiple R-squared:  0.3156, Adjusted R-squared:  0.3136 
## F-statistic: 158.1 on 1 and 343 DF,  p-value: < 2.2e-16
```

- Überprüfung der Gleichheit zwischen Regressionskoeffizient und Bravais-Pearson-Korrelation:

```
cor(x = run$HR, y = run$speed, method = "pearson")

## [1] 0.561747

cor(x = run$s_HR, y = run$s_speed, method = "pearson")

##          [,1]
## [1,] 0.561747

coef(lm_s_speed)["s_speed"]

##  s_speed 
## 0.561747

all.equal(cor(run$HR, run$speed), coef(lm_s_speed)["s_speed"],
          check.attributes = FALSE)

## [1] TRUE
```

Aufgabe 2: Eigenschaften der Koeffizientenschätzer

(a) Betrachten Sie das einfache lineare Regressionsmodell (siehe Vorlesung, Folie 19)

$$Y_i \quad = \quad \beta_0 + \beta_1 x_i + \epsilon_i;\; i = 1, \ldots, n$$

mit den Annahmen

$$\begin{aligned} E(\epsilon_i) &= 0;\; i = 1, \ldots, n \\ \mathrm{Var}(\epsilon_i) &= \sigma^2;\; i = 1, \ldots, n \\ \{\epsilon_i \mid i = 1, \ldots, n\} & \quad \text{stochastisch unabhängig} \\ \epsilon_i &\sim N(0, \sigma^2);\; i = 1, \ldots, n. \end{aligned}$$

Leiten Sie die Varianzen $\mathrm{Var}(\hat{\beta}_0)$ und $\mathrm{Var}(\hat{\beta}_1)$ der Koeffizientenschätzer $\hat{\beta}_0$ und $\hat{\beta}_1$ her (Gleichungen (1.13) und (1.14) aus der Vorlesung Folie 26).

Lösung:

- Varianz des Steigungskoeffizienten:

$$\begin{aligned} \mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}\left(\frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2}\right) &= \frac{\frac{1}{n^2}}{(S_x^2)^2} \cdot \mathrm{Var}\Big(\sum_{i=1}^n y_i (x_i - \bar{x})\Big) \\ = \frac{1}{n^2 (S_x^2)^2} \sum_{i=1}^n (x_i - \bar{x})^2 \cdot \mathrm{Var}(y_i) &= \frac{\sigma^2}{n^2 (S_x^2)^2} \cdot n S_x^2 = \frac{\sigma^2}{n S_x^2} \end{aligned}$$

$$
\begin{aligned}
&\text{mit } \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\\
&=\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \bar{x} y_i - \sum_{i=1}^{n} x_i \bar{y} + \sum_{i=1}^{n} \bar{x}\bar{y}\\
&=\sum_{i=1}^{n} y_i(x_i-\bar{x}) - \bar{y}\sum_{i=1}^{n} x_i + n\bar{y}\bar{x}\\
&=\sum_{i=1}^{n} y_i(x_i-\bar{x}) - \bar{y}\sum_{i=1}^{n} x_i + n\bar{y}\frac{\sum_{i=1}^{n} x_i}{n}\\
&=\sum_{i=1}^{n} y_i(x_i-\bar{x}) - \bar{y}\sum_{i=1}^{n} x_i + \bar{y}\sum_{i=1}^{n} x_i\\
&=\sum_{i=1}^{n} y_i(x_i-\bar{x})\\
&\text{und } \mathrm{Var}(y_i)=\sigma^2
\end{aligned}
$$

- Varianz des Achsenabschnitts:

$$
\mathrm{Cov}(y_i,\hat{\beta}_1)=\mathrm{Cov}\Big(y_i,\frac{1}{nS_x^2}\cdot\sum_{k=1}^{n} y_k(x_k-\bar{x})\Big)=\frac{1}{nS_x^2}\sum_{k=1}^{n}(x_k-\bar{x})\mathrm{Cov}(y_i,y_k)
$$

$$
\begin{aligned}
\mathrm{Cov}(\bar{y},\hat{\beta}_1)&=\mathrm{Cov}(\frac{1}{n}\sum_{i=1}^{n} y_i,\hat{\beta}_1)=\frac{1}{n}\sum_{i=1}^{n}\mathrm{Cov}(y_i,\hat{\beta}_1)\\
&=\frac{1}{n^2S_x^2}\sum_{i=1}^{n}\sum_{k=1}^{n}(x_k-\bar{x})\mathrm{Cov}(y_i,y_k)=\frac{\sigma^2}{n^2S_x^2}\sum_{i=1}^{n}(x_i-\bar{x})=0
\end{aligned}
$$

$$
\text{mit } \mathrm{Cov}(y_i,y_k)=\begin{cases}0 \text{ für } i\neq k\\ \sigma^2 \text{ für } i=k\end{cases}
$$

$\Rightarrow$

$$
\begin{aligned}
\mathrm{Var}(\hat{\beta}_0)&=\mathrm{Var}(\bar{y}-\hat{\beta}_1\bar{x})=Var(\bar{y})+\bar{x}^2\mathrm{Var}(\hat{\beta}_1)-2\bar{x}\mathrm{Cov}(\bar{y},\hat{\beta}_1)\\
&=\mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n} y_i)+\bar{x}^2\mathrm{Var}(\hat{\beta}_1)=\frac{\sigma^2}{n}+\bar{x}^2\cdot\frac{\sigma^2}{nS_x^2}=\frac{\sigma^2}{n}\Big(1+\frac{\bar{x}^2}{S_x^2}\Big)
\end{aligned}
$$

(b) Veranschaulichen Sie anhand einer Simulation in **R**, dass die Koeffizientenschätzer des in Teilaufgabe (b) beschriebenen einfachen linearen Regressionsmodells folgende Verteilungen besitzen:

$$
\begin{aligned}
\hat{\beta}_0 &\sim N(\beta_0,\mathrm{Var}(\hat{\beta}_0))\\
\hat{\beta}_1 &\sim N(\beta_1,\mathrm{Var}(\hat{\beta}_1)).
\end{aligned}
$$

Verwenden Sie hierzu den folgenden Regressionszusammenhang zur Durchführung von $N=$ 10.000 Simulationen:

$$
Y_i=-2+3.5\cdot x_i+\epsilon_i \quad \text{mit } \epsilon_i\sim N(0,10).
$$

Lösung:

- Vorbereitung der Simulationen:

```r
# Setzen der Zufallszahlen:
set.seed(3456)
# Definition der Modellparameter:
N <- 100000
n <- 100
beta.0 <- -2
beta.1 <- 3.5
sigma.sq <- 100
# Zufällige Ziehung von N Beobachtungen der Einflussgröße aus einer
# Gleichverteilung:
x_unif <-  runif(n = N, min = 0, max = n)
# Alternative: z.B. Ziehung aus einer Exponentialverteilung
# Zufällige Ziehung des Fehlerterms epsilon aus einer Normalverteilung:
epsilon <- rnorm(N, mean = 0, sd = sqrt(sigma.sq))
# Berechnung der Werte der Zielgröße über Regressionsmodell:
y_vals <- beta.0 + beta.1 * x_unif + epsilon
# Abspeichern der Informationen in einem Datensatz:
predictor <- x_unif
response <- y_vals
data_sim <- data.frame(predictor, response)
```

- Berechnung der wahren Varianzen mit Formeln aus Teilaufgabe (a):

```r
var_b0_true <- (sigma.sq / n) * (1 + (mean(predictor)^2) / var(predictor))
var_b0_true

## [1] 3.997113

var_b1_true <- sigma.sq / (n * var(predictor))
var_b1_true

## [1] 0.001199093
```

- Durchführung der Simulationen:

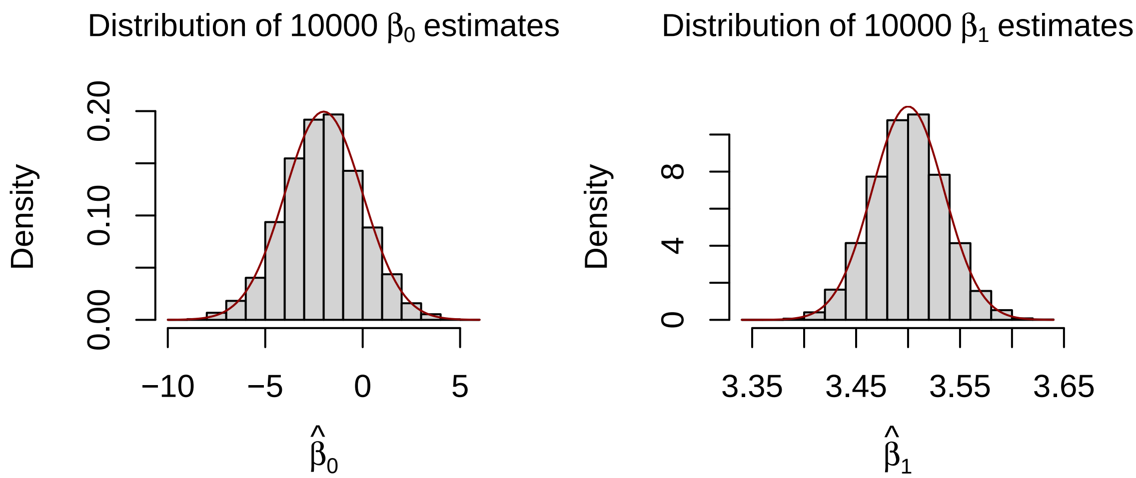
  Um die Verteilung des Schätzers abbilden zu können, reicht es nicht aus, das Modell nur ein einzelnes Mal zu schätzen. Stattdessen wird das Modell im Folgenden 10000 mal gefittet, damit Kenngrößen (Mittelwert und Varianz) der Verteilung sinnvoll angegeben werden können.

```r
reps <- 10000
# Matrix der Ergebnisse:
fit <- matrix(ncol = 2, nrow = reps)
# for-Schleife über die Wiederholungen:
for (i in 1:reps){
  sample <- data_sim[sample(1:N, n), ]
  fit[i, ] <- lm(response ~ predictor, data = sample)$coefficients
}
# Erhaltene Varianzschätzungen:
var(fit[, 1])

## [1] 3.98428

var(fit[, 2])

## [1] 0.001208074
```

$\Rightarrow$ Die geschätzten Varianzen der Koeffizientenschätzer sind sehr ähnlich zu den theoretisch hergeleiteten Varianzen.

- Graphischer Vergleich der erhaltenen Verteilungen für beide Koeffizienten mit theoretischen Normalverteilungen:

```
par(mfrow = c(1, 2))
# Achsenabschnitt:
hist(x = fit[, 1], cex.main = 1,
     main = bquote(Distribution  ~ of ~ 10000 ~ beta[0] ~ estimates),
     xlab = bquote(hat(beta)[0]), freq = FALSE)
curve(dnorm(x = x,  mean = -2, sd = sqrt(var_b0_true)), add = TRUE,
      col = "darkred")
# Steigungsparameter:
hist(x = fit[, 2], cex.main = 1,
     main = bquote(Distribution  ~ of ~ 10000 ~ beta[1] ~ estimates),
     xlab = bquote(hat(beta)[1]), freq = FALSE)
curve(dnorm(x = x,  mean = 3.5, sd = sqrt(var_b1_true)), add = TRUE,
      col = "darkred")
```



Aufgabe 3: Quadratsummenzerlegung

In einer Erfassung verschiedener Merkmale von Baseball-Spielern der Major League Baseball (MLB) wurden unter anderem die Körpergröße (in cm) sowie das Gewicht (in kg) der Spieler erhoben. Für 10 zufällig ausgewählte Spieler liegen folgende Werte vor:

Spieler i	1	2	3	4	5	6	7	8	9	10
Größe x_i	198	188	196	190	180	183	196	196	193	183
Gewicht y_i	104	84	107	95	76	79	109	94	113	93

Hinweis: Für die Teilaufgaben (b)-(d) kann **R** als Taschenrechner verwendet werden.

(a) Zeichnen Sie die Information in ein Streudiagramm. Welcher Zusammenhang ist zwischen Körpergröße und Gewicht erkennbar?

Lösung:

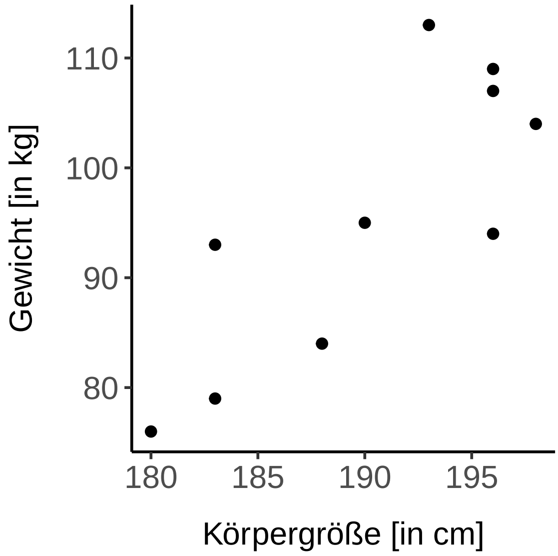
- Erstellung des Datensatzes in **R**:

```
data <- data.frame("Groesse" = c(198, 188, 196, 190, 180, 183, 196, 196, 193,
                                 183),
                   "Gewicht" = c(104, 84, 107, 95, 76, 79, 109, 94, 113, 93))
data
```

```
##    Groesse Gewicht
## 1      198     104
## 2      188      84
## 3      196     107
## 4      190      95
## 5      180      76
## 6      183      79
## 7      196     109
## 8      196      94
## 9      193     113
## 10     183      93
```

- Graphische Visualisierung des Zusammenhangs in einem Streudiagramm:

```
# Graphische Visualisierung des Zusammenhangs in einem Streudiagramm:
ggplot(data = data, mapping = aes(x = Groesse, y = Gewicht)) + geom_point() +
  xlab("Körpergröße [in cm]") + ylab("Gewicht [in kg]") + theme
```



⇒ Das Streudiagramm zeigt einen klar positiven Zusammenhang zwischen Körpergröße und Gewicht.

(b) Berechnen Sie die Gesamtstreuung (SST) der Variable `Gewicht`.

Lösung:

- Komponenten der Streuungszerlegung:
  - SST (Sum of Squares Total): Gesamtstreuung der Zufallsvariable Y
  - SSE (Sum of Squares Error): Streuung der Residuen
  - SSM (Sum of Squares Model): Streuung, die das Modell erklärt
- Streuungszerlegung:

$$SST = SSE + SSM \quad \Leftrightarrow \quad \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

- Berechnung der Gesamtstreuung der Variable `Gewicht`:

```
SST <- sum((data$Gewicht - mean(data$Gewicht))^2)
SST

## [1] 1486.4
```

(c) Bestimmen Sie per Hand die Koeffizientenschätzer $\hat{\beta}_0$ und $\hat{\beta}_1$ für ein lineares Regressionsmodell, das das Gewicht in Abhängigkeit der Körpergröße modelliert.

Lösung:

- Bestimmung des Steigungsparameters $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\hat{Cov}(X,Y)}{\hat{Var}(X)}$$

```
beta1 <- cov(data$Gewicht, data$Groesse) / var(data$Groesse)
beta1

## [1] 1.603769
```

- Bestimmung des Achsenabschnitts $\hat{\beta}_0$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

```
beta0 <- mean(data$Gewicht) - beta1 * mean(data$Groesse)
beta0

## [1] -209.7972
```

(d) Berechnen Sie mit Hilfe der in Teilaufgabe (c) bestimmten Parameter die Streuung, die durch das Modell erklärt wird (SSM). Bestimmen Sie anschließend das Bestimmtheitsmaß R^2 und interpretieren Sie dieses.

Lösung:

- Berechnung der Streuung, die das Modell erklärt:

```
pred <- beta0 + beta1 * data$Groesse
SSM <- sum((pred - mean(data$Gewicht))^2)
SSM

## [1] 982.7894
```

- Bestimmung des Bestimmtheitsmaßes $R^2 = \frac{SSM}{SST} = 1 - \frac{SSE}{SST}$:

```
R2 <- SSM / SST
R2

## [1] 0.6611877
```

$\Rightarrow$ Ca. 66% der Streuung des Gewichtes können durch das geschätzte lineare Regressionsmodell erklärt werden.

(e) Überprüfen Sie Ihre Berechnungen, indem Sie die Schätzung des Regressionsmodells nun in **R** mit der Funktion `lm` durchführen.

Lösung:

- Überprüfung der Berechnungen durch Funktion `lm`:

```
model <- lm(formula = Gewicht ~ Groesse, data = data)
summary(model)

##
## Call:
## lm(formula = Gewicht ~ Groesse, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.541  -4.457  -1.400   3.958  13.270
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -209.7972    77.2826  -2.715  0.02647 *
## Groesse        1.6038     0.4059   3.951  0.00423 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.934 on 8 degrees of freedom
## Multiple R-squared:  0.6612, Adjusted R-squared:  0.6188
## F-statistic: 15.61 on 1 and 8 DF,  p-value: 0.004229
```

$\Rightarrow$ Die beiden Koeffizientenschätzer sowie das Bestimmtheitsmaß stimmen mit den über die Formeln berechneten Werten überein.

- Graphische Visualisierung des Zusammenhangs mit eingezeichneter Regressionsgerade:

```
ggplot(data = data, mapping = aes(x = Groesse, y = Gewicht)) + geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  xlab("Körpergröße [in cm]") + ylab("Gewicht [in kg]") + theme
```