Universität Duisburg-Essen
Lehrstuhl für Ökonometrie
Dr. Yannick Hoga
MSc. Martin Arnold

Wintersemester 2022/23

Zusatzaufgabe 1

Mit dieser Zusatzaufgabe können Sie bis zu 2.5 Zusatzpunkte für die Klausur erwerben. Die Zusatzpunkte werden nur bei *bestandener Klausur* angerechnet. Sie können die Aufgaben in einer Gruppe von bis zu drei Personen bearbeiten. Ihre Lösungen sind bis einschließlich Mittwoch, den 7.12.2022, unter Nennung von Namen und Matrikelnummer aller Beteiligten **als kommentiertes R-Skript per E-Mail einzureichen (martin.arnold@vwl.uni-due.de)**.

Aufgabe 1

Der Datensatz `stimmung.csv` enthält Beobachtungen des Wohlbefindens von Eltern (`stimmung`, Skala von 1 bis 100) sowie für die Schlaufdauer ihres Babys (`schlaf_baby`, Stunden pro Nacht) für 100 Tage.

(a) Lesen Sie `stimmung.csv` mit einer geeigneten Funktion aus dem Paket `tidyverse` als Objekt `stimmung` in R ein und erzeugen Sie eine statistische Zusammenfassung der Daten.

```
# install.packages("tidyverse")
stimmung <- read_csv("stimmung.csv")
summary(stimmung)
```

```
  schlaf_baby        stimmung
 Min.   : 3.250   Min.   : 9.00
 1st Qu.: 6.425   1st Qu.:29.00
 Median : 7.950   Median :38.00
 Mean   : 8.049   Mean   :36.29
 3rd Qu.: 9.635   3rd Qu.:43.00
 Max.   :12.070   Max.   :59.00
```

```
# tidyverse
stimmung %>%
  pivot_longer(everything(), names_to = "variable", values_to= "value") %>%
  group_by(variable) %>%
  summarise_at("value", .funs = c("min", "mean", "max", "sd"))
```

```
# A tibble: 2 x 5
  variable      min  mean   max    sd
  <chr>       <dbl> <dbl> <dbl> <dbl>
1 schlaf_baby  3.25  8.05  12.1  2.07
2 stimmung     9    36.3   59   10.0
```

(b) Erzeugen Sie eine statistische Zusammenfassung für die Schätzung des linearen Regressionsmodells

$$\text{stimmung}_i = \beta_0 + \beta_1 \text{schlaf\_baby}_i + u_i, \tag{1}$$

und nutzen Sie das Paket `ggplot2`, um den geschätzten Regressionszusammenhang gemeinsam mit den Daten grafisch darzustellen.

```r
stimmung_mod <- lm(stimmung ~ schlaf_baby, data = stimmung)
summary(stimmung_mod)
```

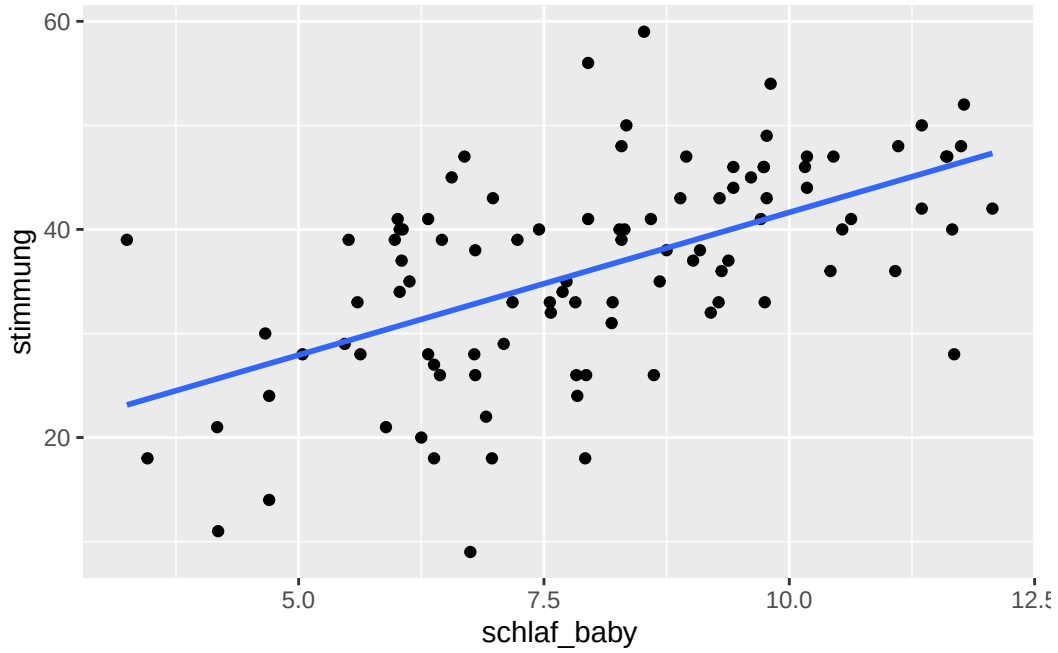
```
Call:
lm(formula = stimmung ~ schlaf_baby, data = stimmung)

Residuals:
     Min       1Q   Median       3Q      Max
-23.7275  -4.9567   0.0587   5.0049  21.4190

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.2183     3.3528   4.241 5.05e-05 ***
schlaf_baby   2.7421     0.4035   6.796 8.45e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.327 on 98 degrees of freedom
Multiple R-squared:  0.3203,    Adjusted R-squared:  0.3134
F-statistic: 46.18 on 1 and 98 DF,  p-value: 8.448e-10
```

```r
ggplot(stimmung, aes(x = schlaf_baby, y = stimmung)) +
  geom_point() +
  geom_smooth(method = "lm", se = F)
```



(c) Interpretieren Sie das in (b) geschätzte Modell. Beurteilen Sie die Anpassung des Modells an die Daten.

Lösung:

- $\widehat{\beta}_0$: Falls das Baby in der vorangegangenen Nacht gar nicht geschlafen hat, beträgt die geschätzte Stimmung der Eltern 14.22 Punkte.

- $\widehat{\beta}_1$: Pro zusätzliche Stunde Schlaf des Babys in der vorangegangenen Nacht verbessert sich die Stimmung der Eltern um geschätzte 2.74 Punkte.
- Der Anpassungskoeffizient R^2 beträgt etwa 32%, d.h. die Anpassung der Schätzung an die Daten ist mäßig.

(d) Nutzen Sie R-Funktionen, um folgende aus der Übung bekannte Eigenschaften der KQ-Schätzung im allgemeinen Modell $y_i = \beta_0 + \beta_1 X_i + u_i,\ i = 1, \ldots, n$ für die Schätzung `stimmung_mod` zu validieren:

1. $\sum_{i=1}^n \widehat{u}_i = 0$
2. $\sum_{i=1}^n \widehat{u}_i X_i = 0$
3. $\frac{1}{n}\sum_{i=1}^n \widehat{y}_i = \frac{1}{n}\sum_{i=1}^n y_i$
4. Die geschätzte Regressionsgerade verläuft durch den Punkt $(\overline{X}, \overline{Y})$.
   Hinweis: nutzen Sie `predict()`.

```
# 1.
res <- residuals(stimmung_mod)
sum(res)
```

```
[1] 2.309264e-14
```

```
# 2.
sum(stimmung$schlaf_baby * res)
```

```
[1] 4.121148e-13
```

```
# 3.
mean(stimmung$stimmung)
```

```
[1] 36.29
```

```
mean(stimmung$stimmung) == mean(fitted(stimmung_mod))
```

```
[1] TRUE
```

```
# 4.
predict(
  stimmung_mod,
  new = data.frame("schlaf_baby" = mean(stimmung$schlaf_baby))
) == mean(stimmung$stimmung)
```

```
    1
FALSE
```

(e) Wiederholen Sie die Teilaufgaben (b) und (c) für das Modell

$$\text{stimmung}_i = \beta_1 \text{schlaf\_baby}_i + u_i,\ i = 1, \ldots, 100. \tag{2}$$

```
# (b)
stimmung_mod <- lm(stimmung ~ schlaf_baby - 1, data = stimmung)
summary(stimmung_mod)
```

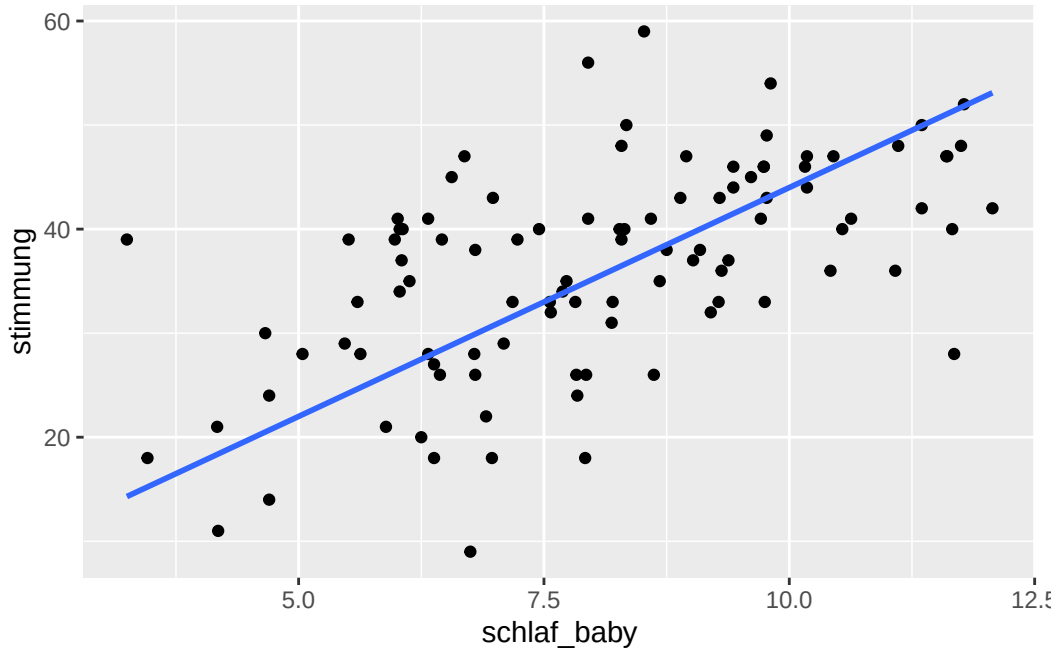
```
Call:
lm(formula = stimmung ~ schlaf_baby - 1, data = stimmung)

Residuals:
    Min      1Q  Median      3Q     Max
-23.387  -4.925   0.184   6.315  24.701

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
schlaf_baby   4.3996     0.1085   40.56   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.013 on 99 degrees of freedom
Multiple R-squared:  0.9432,    Adjusted R-squared:  0.9427
F-statistic:  1645 on 1 and 99 DF,  p-value: < 2.2e-16
```

```r
ggplot(stimmung, aes(x = schlaf_baby, y = stimmung)) +
  geom_point() +
  geom_smooth(method = "lm", se = F, formula = formula("y ~ x - 1"))
```



Lösung:

- $\widehat{\beta}_1$: Pro zusätzliche Stunde Schlaf des Babys in der vorangegangenen Nacht verbessert sich die Stimmung der Eltern um geschätzte 4.40 Punkte.
- Der Anpassungskoeffizient R^2 beträgt etwa 94.32%, d.h. die Anpassung der Schätzung an die Daten ist sehr gut und deutlich besser als im Modell mit Konstante.

(f) Implementieren Sie den Schätzter für β_1 im einfachen Regressionsmodell $y_i = \beta_1 X_i + u_i,\ i = 1, \ldots, n$ als R-Funktion `KQOK()`,

```
KQOK <- function(X, Y) {
  ...
}
```

und überprüfen Sie Ihre Funktion anhand eines Vergleich mit dem Ergebnis in (e).

Lösung:

Der KQ-Schätzer ist $\widehat{\beta}_0 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$.

```
# Funktion definieren
KQOK <- function(X, Y) {
  sum(X * Y) / sum(X^2)
}

# Funktion für Teilaufgabe (e) testen und Schätzung für beta_0 vergleichen
KQOK(X = stimmung$schlaf_baby, Y = stimmung$stimmung)
```

```
[1] 4.399558
```

```
stimmung_mod$coefficients[1]
```

```
schlaf_baby
   4.399558
```

```
# Die Schätzungen stimmen überein.
```

Hinweise: Die Paketsammlung `tidyverse` *kann mit* `install.packages('tidyverse')` *installiert werden. Mit* `library(tidyverse)` *werden sämtliche für die Zusatzaufgabe nötige R-Pakete geladen. Wir empfehlen den DataCamp-Kurs Introduction to Regression in R.*