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

Wintersemester 2022/23

Übungsblatt 3 — Einfaches Regressionsmodell

R-Lösungen

Aufgabe 4 – Regressionsbeispiel: Bremsweg

Bei Verkehrsunfällen spielt für die Klärung der Schuldfrage oft die Beziehung zwischen Geschwindigkeit und Bremsweg eine wichtige Rolle. Die folgende Tabelle enthält diesbezügliche Informationen für einen PKW, die im Rahmen eines Sicherheitstrainings gemessen wurden:

Geschwindigkeit (in km/h)	20	30	50	80	100
Bremsweg (in m)	5	10	25	60	10

(f) Lösen Sie die Teilaufgaben (a)–(e) mit `R`.

(a) Wir definieren zunächst Vektoren für die beobachteten Daten.

```
Geschwindigkeit <- c(20, 30, 50, 80, 100)
Bremsweg <- c(5, 10, 25, 60, 10)
```

Anschließend schätzen wir das Modell mit `lm()`. Das Modell wird als Formel an `formula` übergeben.

```
# Modell schätzen
mod <- lm(formula = Bremsweg ~ Geschwindigkeit)
mod
```

```
Call:
lm(formula = Bremsweg ~ Geschwindigkeit)

Coefficients:
    (Intercept)  Geschwindigkeit
         6.0177           0.2854
```

`summary()` liefert eine detailierte statistische Zusammenfassung der Schätzung.

```
s <- summary(mod)
s
```

```
Call:
lm(formula = Bremsweg ~ Geschwindigkeit)

Residuals:
      1       2       3       4       5
 -6.726  -4.580   4.712  31.150 -24.558
```

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.0177    22.2513   0.270    0.804
Geschwindigkeit  0.2854     0.3501   0.815    0.475

Residual standard error: 23.54 on 3 degrees of freedom
Multiple R-squared:  0.1814,    Adjusted R-squared:  -0.09152
F-statistic: 0.6646 on 1 and 3 DF,  p-value: 0.4746
```

(b) Die Residuen sind Bestandteil von `mod` und können mit `residuals()` ausgelesen werden. Wir können diese jedoch auch aus dem output von `summary()` erhalten. Ähnlich ist es für die geschätzen ("gefitteten") Werte.

```
res <- residuals(mod)
res
```

```
         1          2          3          4          5
 -6.725664  -4.579646   4.712389  31.150442 -24.557522
```

```
# res <- s$residuals  # alternativ

# Gefittete Werte
fitted(mod)
```

```
       1        2        3        4        5
11.72566 14.57965 20.28761 28.84956 34.55752
```

Tatsächlich ist die Summe der Residuen (approximativ) 0. *Beachte die wissenschaftliche Notation!*

```
sum(res)
```

```
[1] 0
```

(c) Mit `predict()` erhalten wir anhand eines geschätzten Modells Schätzungen der abhängigen Variable für beliebige Regressor-Werte. *Regressor-Werte müssen in einem* `data.frame` *übergeben werden! Die Variablennamen müssen übereinstimmen!*

```
neu <- data.frame("Geschwindigkeit" = 60)
neu
```

```
  Geschwindigkeit
1              60
```

```
predict(mod, newdata = neu)
```

```
       1
23.14159
```

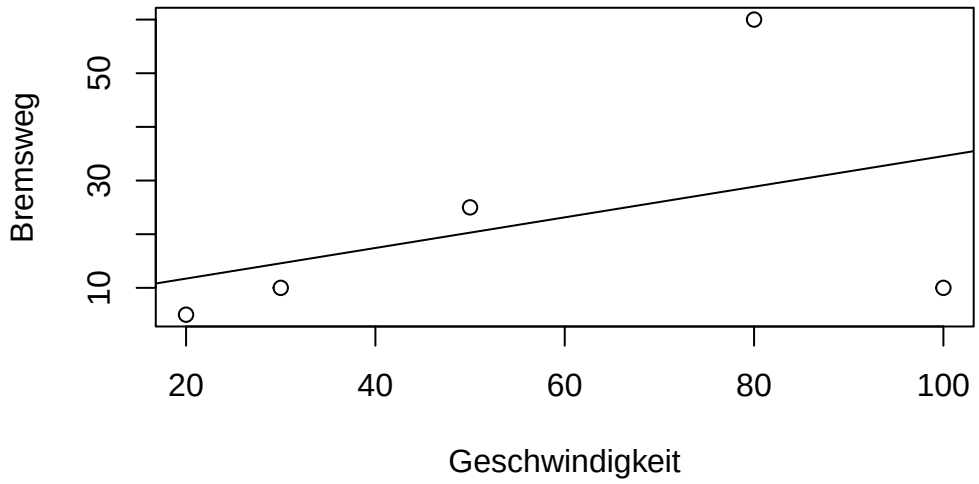
(d) R^2 kann aus dem Output von `summary(mod)` ausgelesen werden.

```
s$r.squared
```

```
[1] 0.1813614
```

(e) Ein Punkteplot mit geschätzter Regressionsgerade ist schnell erstellt. Beachte, dass `abline()` *nur in eine bereits erzeugte Grafik* eingezeichnet werden kann!

```
plot(x = Geschwindigkeit, y = Bremsweg)
abline(mod)
```



Aufgabe 5 – Regressionsanalyse mit `CASchools`

(a) Laden Sie den Datensatz `CASchools` (verfügbar in Moodle). Nutzen sie hierzu den Befehl `load()`.

Wir setzen zunächst das Arbeitsverzeichnis und laden anschließend den Datensatz `CASchools.Rdata`. Dateien mit der Endung `.Rdata` enthalten *R-Objekte* (`CASchools.Rdata` enthält den `data.frame` `CASchools`) und der Funktion `load()` eingelesen werden.

```
## setwd("Z:/")
load("CASchools.Rdata")
```

(b) Erzeugen Sie eine statistische Zusammenfassung des Datensatzes mit `summary()`.

Wir nutzen `summary()` für die statistische Zusammenfassung aller Variablen in `CASchools`.

```
summary(CASchools)
```

```
      math            STR           income          computer
 Min.   :605.4   Min.   :14.00   Min.   : 5.335   Min.   :   0.0
 1st Qu.:639.4   1st Qu.:18.58   1st Qu.:10.639   1st Qu.:  46.0
 Median :652.5   Median :19.72   Median :13.728   Median : 117.5
 Mean   :653.3   Mean   :19.64   Mean   :15.317   Mean   : 303.4
 3rd Qu.:665.9   3rd Qu.:20.87   3rd Qu.:17.629   3rd Qu.: 375.2
 Max.   :709.5   Max.   :25.80   Max.   :55.328   Max.   :3324.0
  expenditure
 Min.   :3926
 1st Qu.:4906
 Median :5215
 Mean   :5312
 3rd Qu.:5601
 Max.   :7712
```

Stichproben-Standardabweichung und -varianz können mit `sd()` bzw. `var()` für die Variablen in `CASchools` einzeln berechnet werden. Beachte, dass `$` genutzt wird, um Variablen in `CASchools` zu referenzieren.

```r
# Empirische Standardabweichungen für math und STR
sd(CASchools$math)
sd(CASchools$STR)
```

```
[1] 18.7542
[1] 1.891812
```

```r
# Empirische Varianzen für math und STR
var(CASchools$math)
var(CASchools$STR)
```

```
[1] 351.7201
[1] 3.578952
```

(c) Regressieren Sie die im Mathetest erreichte Punktzahl (`math`) auf die Klassengröße (`STR`). Lassen Sie sich mit `summary()` eine statistische Zusammenfassung des Modells anzeigen.

Wir übergeben den Datensatz sowie das zu schätzende Modell als Formel an `lm()` und erzeugen mit `summary()` eine statistische Zusammenfassung der Schätzung.

```r
# Schätzen
mod_CASchools <- lm(math ~ STR, data = CASchools)
# Geschätzte Koeffizienten 'schnell' erhalten
mod_CASchools
```

```
Call:
lm(formula = math ~ STR, data = CASchools)

Coefficients:
(Intercept)          STR
    691.417       -1.939
```

```r
# Statistische Zusammenfassung des Modells erzeugen
summary(mod_CASchools)
```

```
Call:
lm(formula = math ~ STR, data = CASchools)

Residuals:
    Min      1Q  Median      3Q     Max
-44.615 -13.374  -0.828  12.728  52.711

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 691.4174     9.3825  73.692  < 2e-16 ***
STR          -1.9386     0.4755  -4.077 5.47e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.41 on 418 degrees of freedom
Multiple R-squared:  0.03824,   Adjusted R-squared:  0.03594
F-statistic: 16.62 on 1 and 418 DF,  p-value: 5.467e-05
```

(d) Plotten Sie die Daten zusammen mit der Regressionsgeraden. Interpretieren Sie *kurz* die Ergebnisse.
Hinweis: `CASchools` ist ein `data.frame`. Variablen können mit `$` ausgelesen werden.
Beispiel: `CASchools$math`

```
plot(x = CASchools$STR,                    # Daten
     y = CASchools$math,
     pch = 19,                             # ausgemahlte Kreise
     col = "blue",                         # Farbe
     cex = .5,                             # Durchmesser der Kreise
     main = "math vs. STR",                # Titel
     xlab = "Klassengroesse (STR)",        # Beschriftung X
     ylab = "Mathe-Punktzahl (math)"       # Beschriftung Y
    )
abline(mod_CASchools,
   col = "red",
   lwd = 2                                 # Linienbreite
   )
```