MLE01 - Vertiefende statistische Verfahren

1. Übungsblatt SS 2024

Stefan Kolb, Joachim Waltl

Allgemeine Information

Alle Aufgaben sind mit R zu lössen. Die Berechnungen sollen nachvollziehbar und dokumentiert sein. Um die vollständige Punktezahl zu erreichen, müssen alle Ergebnisse und Fragen entsprechend interpretiert bzw. beantwortet werden. Code alleine ist nicht ausreichend! Die Abgabe erfolgt über Moodle entsprechend der Abgaberichtlinien als pdf und Rmd File. Bitte inkludieren Sie namentlich alle beteiligten Gruppenmitglieder sowohl im Bericht als auch im Source Code. Die jeweiligen Datensätze die für diese Übung relevant sind finden Sie ebenfalls in Moodle.

1 Lineare Regressionsanalyse [4P]

Für Menschen, die ihren Blutdruck senken wollen, ist eine häufig empfohlene Vorgehensweise, die Salzaufnahme zu senken. Sie möchten feststellen, ob es eine lineare Beziehung zwischen Salzaufnahme und Blutdruck gibt. Sie nehmen 52 Personen in die Stichprobe auf und messen deren diastolischen Blutdruck (in mmHg) und Natriumausscheidung (mmol/24h). [ref]

[2P] a: Importieren Sie den Datensatz `intersalt.csv`. Erstellen Sie zwei Regressionsmodelle für den diastolische Blutdruck (bp) in Abhängigkeit der Natriumausscheidung (na). Das erste Modell soll alle Datenpunkte verwenden. Für das zweite Modell sollen die vier Datenpunkte mit der geringsten Natriumausscheidung aus dem Datensatz entfernt werden.

```r
# Daten einlesen
intersalt <- read.csv("intersalt.csv", sep = ";", dec = ",")

# Überblick über die Daten
str(intersalt)
```

```
## 'data.frame':    52 obs. of  4 variables:
##  $ b      : num  0.512 0.226 0.316 0.042 0.086 0.265 0.384 0.501 0.352 0.443 ...
##  $ bp     : num  72 78.2 73.9 61.7 61.4 73.4 79.2 66.6 82.1 75 ...
##  $ na     : num  149.3 133 142.6 5.8 0.2 ...
##  $ country: chr  "Argentina" "Belgium" "Belgium" "Brazil" ...
```

```r
# Umwandlung der Variablen vom Typ Character in numerische Variablen
intersalt$b <- as.numeric(intersalt$b)
intersalt$bp <- as.numeric(intersalt$bp)
intersalt$na <- as.numeric(intersalt$na)

# Erstes Modell (alle Datenpunkte)
model_1 <- lm(bp ~ na, data = intersalt)
```

```r
# Datensatz nach Natriumausscheidung sortieren
intersalt_sorted <- intersalt[order(intersalt$na),]

# Enfernen der ersten vier Datenpunkte
intersalt_m2 <- intersalt_sorted[-c(1:4),]

# Zweites Modell
model_2 <- lm(bp ~ na, data = intersalt_m2)
```

Führen Sie für beide Modelle eine lineare Regressionsanalyse durch, die folgende Punkte umfasst:

i) Modellgleichung inklusive 95% Konfidenzintervall der Modellparameter
ii) Interpretation des Ergebnisses hinsichtlich Signifikanz und Modellgüte
iii) Grafische Darstellung der Regressionsgeraden inkl. Konfidenzintervall

Modell 1

```r
# Zusammenfassung und KI-Intervall für model_1
summary(model_1)
```

```
##
## Call:
## lm(formula = bp ~ na, data = intersalt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8625 -2.8906  0.0299  3.6470  9.4283
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 67.56245    2.14643  31.477   <2e-16 ***
## na           0.03768    0.01384   2.722   0.0089 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.511 on 50 degrees of freedom
## Multiple R-squared:  0.1291, Adjusted R-squared:  0.1117
## F-statistic: 7.411 on 1 and 50 DF,  p-value: 0.008901
```

```r
confint(model_1)
```

```
##                    2.5 %      97.5 %
## (Intercept) 63.251226427 71.87367736
## na           0.009878513  0.06547863
```

Die Modellgleichung für das erste Modell lautet: bp = 67.56 [63.25;71.87] + 0.038 [0.01;0.07] * na.

Intercept: Das 95% Konfidenzintervall für den Intercept liegt zwischen 63.25123 und 71.87368. Das bedeutet, dass der diastolische Blutdruck bei einer Natriumausscheidung von 0 mmol/24h mit einer 95%igen Sicherheit zwischen 63.25 und 71.87 mmHg liegt.

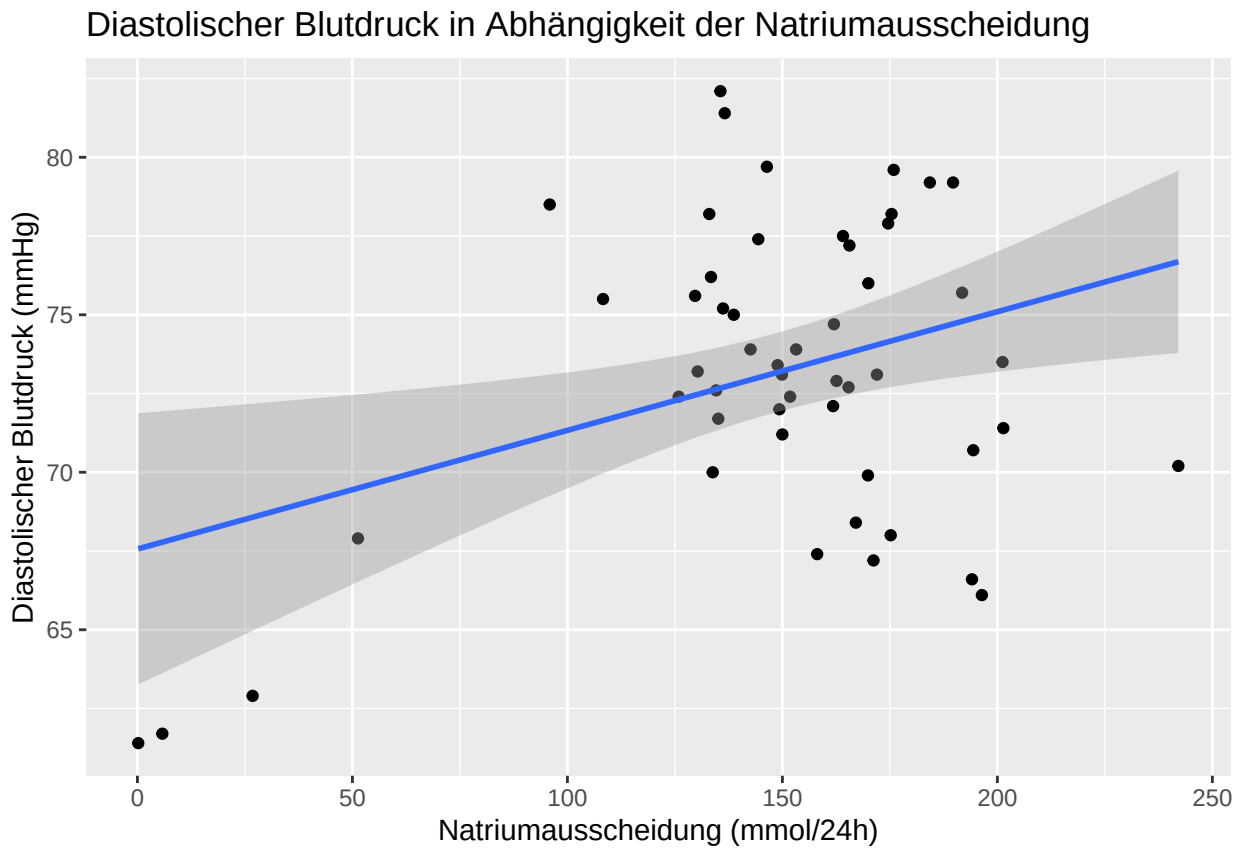
Steigung: Das 95% Konfidenzintervall für die Steigung bezüglich der Natriumausscheidung liegt zwischen 0.00988 und 0.06548. Wir können also mit 95%iger Sicherheit sagen, dass der Anstieg des diastolische Blutdruck zwischen 0.00988 und 0.06548 liegt, wenn die Natriumausscheidung um 1 mmol/24h steigt.

Signifikanz und Modellgüte: Die Signifikanz des p-Wertes für die Steigung ($p = 0.0089$) deutet auf einen statistisch signifikanten Zusammenhang zwischen der Natriumausscheidung und dem diastolischen Blutdruck hin. Der p-Wert für den Intercept dieses Modells ist sogar als hochsignifikant zu werten. Ein Wert von 0.1291 für das Bestimmtheitsmaß R^2 zeigt jedoch, dass das Modell nur etwa 12.91% der Variabilität im diastolischen Blutdruck mit der Natriumausscheidung erklären kann. Die Erkenntnis daraus ist, dass noch andere Faktor den Blutdruck beeinflussen.

i) Grafische Darstellung der Regressionsgeraden inkl. Konfidenzintervall

```
# Grafische Darstellung der Regressionsgeraden inkl. Konfidenzintervall für model_1
library(ggplot2)
ggplot(intersalt, aes(x = na, y = bp)) +
  geom_point() +
  geom_smooth(method = "lm", se = TRUE) +
  labs(title = "Diastolischer Blutdruck in Abhängigkeit der Natriumausscheidung",
       x = "Natriumausscheidung (mmol/24h)",
       y = "Diastolischer Blutdruck (mmHg)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



```r
# Zusammenfassung und KI-Intervall für model_2
summary(model_2)
```

```
##
## Call:
## lm(formula = bp ~ na, data = intersalt_m2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5966 -2.4042 -0.4884  2.8636  7.0977
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 81.06335    3.30938  24.495   <2e-16 ***
## na          -0.04470    0.02053  -2.177   0.0346 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.807 on 46 degrees of freedom
## Multiple R-squared:  0.09342,    Adjusted R-squared:  0.07371
## F-statistic:  4.74 on 1 and 46 DF,  p-value: 0.03464
```

```r
confint(model_2)
```

```
##                   2.5 %      97.5 %
## (Intercept) 74.40191390 87.72478658
## na          -0.08602363 -0.00337225
```

Die Modellgleichung für das zweite Modell lautet: bp = 81.06 [74.40;87.72] - 0.045 [-0.09;0.00] * na.

Intercept: Das 95% Konfidenzintervall für den Intercept liegt zwischen 74.40191 und 87.72479.

Steigung: Das 95Konfidenzintervall für die Steigung bezüglich der Natriumausscheidung liegt zwischen -0.08602 und -0.00337.

i) Interpretation des Ergebnisses hinsichtlich Signifikanz und Modellgüte

ii) Grafische Darstellung der Regressionsgeraden inkl. Konfidenzintervall

Vergleichen Sie beide Modelle. Was können Sie beobachten?

[2P] b: Lesen Sie den Artikel "The (Political) Science of Salt" und vergleichen Sie damit Ihre Beobachtungen. Gibt es Faktoren die in Ihren Modellen eventuell nicht berücksichtigt wurden? Wie lautet die Schlussfolgerung - führt eine Reduktion der Salzaufnahme zu einer Blutdrucksenkung?

2 Lineare Regressionsanalyse (kategorisch) [3P]

Der Datensatz `infant.csv` enthält Information über die unterschiedliche Kindersterblichkeit zwischen den Kontinenten. Die Variable `infant` enthält die Kindersterblichkeit in Tode pro 1000 Geburten. Unterscheidet sich die Kindersterblichkeit zwischen den Kontinenten?

[2P] a: Führen Sie eine Regressionsanalyse mit Europa als Referenz durch, welche die folgenden Punkte umfasst:

i) Modellgleichung inklusive 95% Konfidenzintervall der Modellparameter
ii) Interpretation des Ergebnisses hinsichtlich Signifikanz
iii) Beurteilung der Modellgüte und Residuenanalyse

[1P] b: Wie hoch ist die Kindersterblichkeit in Europa und wie hoch in Afrika (inkl. Unsicherheit)?

3 Regressionsanalyse [3P]

Die Daten `wtloss.xlsx` enthalten den Gewichtsverlauf eines adipösen Patienten im Zuge einer Diät. Sie als betreuender Mediziner und passionierter Freizeit Data Scientist möchten ein geeignetes Regressionsmodell erstellen, um den Verlauf der Diät besser steuern zu können. Das ideale Zielgewicht bezogen auf die Größe des Patienten wäre bei 80 kg. Importieren Sie den Datensatz mit Hilfe der `read_excel()` Funktion aus dem `library(readxl)` Paket.

[2P] a: Die Regressionsanalyse sollte folgende Punkte inkludieren:

i) Modellgleichung inklusive 95% Konfidenzintervall der Modellparameter
ii) Interpretation des Ergebnisses hinsichtlich Signifikanz
iii) Beurteilung der Modellgüte und Residuenanalyse
iv) Grafische Darstellung der Regressionsgeraden inkl. Konfidenz-und Vorhersageintervall

[1P] b: Welches Gewicht hat der Patient nach 30 Tagen bzw. nach 200 Tagen Diät?

4 Multiple Regressionsanalyse [3P]

Die Framingham-Herz-Studie war ein Wendepunkt bei der Identifizierung von Risikofaktoren für koronare Herzkrankheiten und ist eine der wichtigsten epidemiologischen Studien die je durchgeführt wurden. Ein großer Teil unseres heutigen Verständnisses von Herz-Kreislauf-Erkrankungen ist auf diese Studie zurückzuführen. Der Datensatz `Framingham.sav` enthält Varibalen hinsichtlich Demographie, Verhaltensweise, Krankengeschichte und Risikofaktoren. Finden Sie ein geeignetes Modell, dass den systolischen Blutdruck (`sysbp`) beschreibt. Vermeiden Sie nicht relevante bzw. redundante Variablen (z.B. "Incident" Variablen). Achten Sie auf Ausreißer und fehlende Daten (`NaN, NA's`).

[2P] a: Die Regressionsanalyse sollte folgende Punkte inkludieren:

i) Modellgleichung inklusive 95% Konfidenzintervall der Modellparameter
ii) Interpretation des Ergebnisses hinsichtlich Signifikanz
iii) Beurteilung der Modellgüte und Residuenanalyse

[1P] b: Welchen systolischen Blutdruck hat eine Person mit folgendem Profil:

Frau, 50 Jahre, High School, Raucher, 8 Zig/Tag, keine Blutdruck senkenden Medikamente, 220 mg/dl Serum Cholesterol, 85 mmHg diastolischer Blutdruck, BMI von 30, kein Diabetes, 90 bpm Herzrate und Glukoselevel von 90 mg/dl.