MLE01 - Vertiefende statistische Verfahren

1. Übungsblatt SS 2024

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.

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

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.