Universität Leipzig Sommersemester 2024

Mathematische Statistik
Übungsblatt 4

Auf diesem Übungsblatt wollen wir die praktische Anwendung von linearen Modellen mit R anschauen. Sie können R hier herunterladen: https://cran.r-project.org/

Ich würde auch empfehlen RStudio zu installieren (es ist aber nicht notigwendig). Sie können es hier herunterladen (RStudio Desktop, die kostenlose Version): https://posit.co/downloads/

Wenn Sie keine oder nur wenig Erfahrung mit R haben, können Sie sich auch einen ChatGPT-Account anlegen, um sich bei der Übung helfen zu lassen: https://chatgpt.com/auth/login

Übung 13

In dieser Aufgabe wollen wir einige grundlegende Funktionen von R zu linearen Modellen betrachten.

a. Erzeugen Sie $n = 3000$ viele Kovariablen $X_i \in \mathbb{R}^4$ mit: $X_{i,1} = 1$, $X_{i,2} \sim \text{Exp}(2)$ und $X_{i,3}$ und $X_{i,4}$ sollen gemeinsam normalveteilt sein mit Erwartungswert 0, Varianz 1 bzw. 2 und Korrelation 0.5.

b. Erzeugen Sie nun Beobachtungen $Y_i = X_i^T \beta_0 + \varepsilon_i$ mit

$$\beta_0 = \begin{pmatrix} 0.5 & -0.3 & 0.8 & -1.2 \end{pmatrix}$$

und $\varepsilon_i \sim \mathcal{N}(0, 0.7)$.

c. Führen Sie nun eine lineare Regression von Y_i auf X_i durch. Wo finden sich p-Werte? Geben Sie die Konfidenzintervalle zum Niveau $1-\alpha = 0.99$ an. Sind Sie zufrieden mit der Schätzung?

d. Führen Sie nun eine weitere lineare Regression durch, diesmal jedoch nur von Y auf die ersten drei Kovariablen. Wie ändert sich das Ergebnis?

e. Um zu verstehen, wie die Standardfehler berechnet werden, berechnen wir die Fehler manuell mit der Formel für homoskedastische Daten:

```
X <- cbind(1,X1,X2,X3)
beta_hat <- coef(mod1)
epsilon_hat <- Y-X%*%beta_hat
```

```
sigma_epsilon_sq_hat <- sum(epsilon_hat^2)/(n-4)
sigma_sq_hat <- sigma_epsilon_sq_hat*solve(t(X)%*%X/n)

std_error <- sqrt(diag(sigma_sq_hat))/sqrt(n)
print(std_error)
```

Mit dem Befehl *cbind* werden die Kovariablen zur großen Matrix $\underline{X}_n$ verbunden (inklusive eines Intercepts, die *1* muss hier nicht explizit als Vektor angegeben werden, sondern R interpretiert dies automatisch richtig). Mit dem Befehl *coef* können wir die Werte von $\hat{\beta}_n$ aus dem Modell *mod1* erhalten. Der Operator *%*%* bezeichnet das Matrix Produkt in R (mit * wird elementweise Multiplikation bezeichnet, ebenso wird mit ^ das elementweise Potenzieren bezeichnet). Der Befehl *sum* summiert alle Einträge des eingegebenen Vektors. Um das Inverse einer Matrix zu berechnen, nutzen wir die Funktion *solve* und *t(M)* gibt das Transponierte der Matrix M zurück. Der Befehl *diag* extrahiert die Diagonale einer Matrix als Vektor.

Für die Schätzung von $\mathbb{E}(\varepsilon_1^2)$ nutzt R den erwartungstreuen Schätzer. Dieser ist gegeben durch

$$\frac{1}{n-p}\sum_{i=1}^{n}\left(Y_i - X_i'\hat{\beta}_n\right)^2,$$

wenn $X_i \in \mathbb{R}^p$ ist.

Übung 14

In dieser Aufgabe wollen wir uns mit der Berechnung von kausalen Effekten befassen. Dafür benötigen Sie den Datensatz *titanic_data.RData*, den Sie von Moodle herunterladen können. Am 14. April 1912 ist die RMS Titanic gesunken. Der Datensatz enthält folgende Variablen (jeweils Vektoren), die für 2201 Passagiere folgende Informationen enthalten:

- *survived*: 1 =Passagier hat überlebt, 0 =Passagier hat nicht überlebt
- *class*: 1 =Passagier fuhr erster Klasse, 0 =Passagier fuhr zweiter oder dritter Klasse
- *age*: 1 =Erwachsener, 0 =Kind
- *gender*: 1 =Mann, 0 =Frau

Wir interessieren für den kausalen Effekt, den die Klasse auf das Überleben hat.

a. Laden Sie den Datensatz in R. In RStudio geht dies durch Anklicken (im Fenster Files unten rechts). In der R-Konsole geht es z.B. so:

```
setwd("Path/to/folder")
load("titanic_data.RData")
```

Berechnen sie den kleinste Quadrate Schätzer in einer Regression der Variable *survived* auf *class*. Denken Sie, dass man das Ergebnis kausal interpretieren kann?

b. Diskutieren Sie, ob sich Ihre Antwort zu a. ändert, wenn noch die Variablen *age* und *gender* aufgenommen werden. Welche Regression würden Sie dann durchführen?

Hinweis: Vermutlich spielt die Variable *gender* bei Kindern eine geringere Rolle als bei Erwachsenen.

c. In der Regression in b. haben Sie vermutlich *age* als einzelne Kovariate in der Regression. Der zugehörige Parameter ist vermutlich positiv. Haben Erwachsene also eine höhere Überlebenswahrscheinlichkeit als Kinder? Warum ist diese Interpretation unzulässig?

Übung 15

In dieser Aufgabe wollen wir LMMs an einen Datensatz anpassen. Dazu benötigen Sie das R-Package *lme4*. Dieses können Sie wie folgt installieren:

```
install.packages("lme4")
library(lme4)
```

Wir untersuchen einen Datensatz, in dem 18 Personen Schlafentzug ausgesetzt waren. Der Schlafentzug dauerte jeweils 10 Tage an und an jedem Tag wurde die Reaktionszeit gemessen, sodass insgesamt 180 Datenpunkte vorliegen. Diese Daten sind in der Variable *sleepstudy* gespeichert, die Sie nach dem Laden des Pakets *lme4* direkt anzeigen können, indem Sie *sleepstudy* in die Konsole schreiben. Sie können den Datensatz auch visuell anzeigen, z.B. durch:

```
library(lattice)
xyplot(Reaction ~ Days | Subject, sleepstudy,type="p",
       xlab = "Days of sleep deprivation",
       ylab = "Average reaction time (ms)", aspect = "xy")
```

Das Paket *lattice* müssen Sie evtl. auch vorher installieren.

a. Betrachten Sie den Datensatz und diskutieren Sie darüber, ob Sie eher ein lineares oder ein gemischtes lineares Model verwenden wollen. Was würden Sie als Fixed Effects nehmen und was als Random Effects?

b. Passen Sie ein lineares Modell an die Daten an, welches als Beobachtung die Reaktionszeit nimmt und als Kovariablen nur die Tage und einen Intercept enthält. Die Kovariablen und Beobachtungen sind diesmal nicht als einzelne Vektoren gegeben, sondern in einer Matrix mit benannten Spalten. Dies kann von *lm* direkt genutzt werden:

   ```
   LM <- lm(Reaction~Days, sleepstudy)
   ```

c. In dem Paket *lme4* ist die Funktion *lmer* enthalten, die gemischte lineare Modelle anpassen kann. Sie funktioniert sehr ähnlich zu *lm*: Fixed Effects werden genauso wie in *lm* angegeben. Random Effects können z.B. durch *+(1|Subject)* in der Modellgleichung hinzugefügt werden. Z.B. heißt *Reaction∼Days+(1|Subject)*, dass ein Modell

   $$Y_{ij} = \beta_0^{(1)} + \beta_0^{(2)} D_j + 1 \cdot \alpha_i + \varepsilon_{ij}$$

   angepasst wird, wobei i die Person und j den Tag bezeichnet. Die Koariable $D_j = j$ enthält den Tag. Berechnen Sie ein geeignetes gemischtes lineares Modell.

d. Verleichen Sie die Ergebnisse aus (b) und (c).

Übung 16

Diese Aufgabe erfordert Programmierung mit R, was nicht Teil des Kurses ist. Eine der ersten statistischen Fragestellungen tauchte bei der Bestimmung der Parameter der Umlaufbahn des Mondes auf. Wir wollen dieses historische Beispiel an einem allgemeinen physikalischen Beispiel betrachten: Wir haben Größen $Y \in \mathbb{R}$ und $X \in \mathbb{R}^p$, von denen wir wissen, dass sie durch die physikalische Gleichung $Y = \beta_0' X$ zusammenhängen, wobei der Vektor $\beta_0 \in \mathbb{R}^p$ unbekannt ist. Haben wir also p Messungen für (Y, X), können wir p Gleichungen aufstellen und diese nach β_0 auflösen. Für exakte Messungen ist dies eine gute Strategie. Im Beispiel mit dem Mond bestanden solche Messungen z.B. aus Daten, die in einer Sternwarte durch ein Teleskop abgelesen wurden. Man war sich früh bewusst, dass solche Methoden fehleranfällig sind, und man hat daher viel Arbeit darauf verwendet, sehr genau messen zu können, indem man z.B. bessere Teleskope gebaut hat. Die Erkenntnis, dass (bei ungenauen Messungen) mehr als p Messungen einen Mehrwert bieten, war ein Meilenstein in der Entwicklung der Statistik als Wissenschaft.
Ein erstes Verfahren zur Bestimmung von $\beta_0 \in \mathbb{R}^p$ funktionierte wie folgt: Angenommen die Anzahl der Datenpunkte ist ein Vielfaches von p, also $n = Kp$ für ein $K \in \mathbb{N}$. Für $r = 1, ..., p$ werden folgende Mittelwerte berechnet:

$$\overline{Y}_r := \frac{1}{K} \sum_{i=(r-1)K+1}^{rK} Y_i, \qquad \overline{X}_r := \frac{1}{K} \sum_{i=(r-1)K+1}^{rK} X_i.$$

Dann werden die p Gleichungen (für $r = 1, ..., p$)

$$\overline{Y}_r := \beta_0' \overline{X}_r$$

nach β_0 aufgelöst.
Wir wollen dieses Verfahren mit dem kleinste Quadrate Schätzer durch Simulationen vergleichen: Simulieren Sie 10.000 Datensätze für $p = 4$ und $n = 40$ wie folgt (alle Zufallsvariablen sind unabhängig voneinander):

$$Y = \beta_0' X + \varepsilon, \qquad \beta_0 = (-2.3, 1, -1, 0.5), \qquad \varepsilon \sim \mathcal{N}(0, 0.8)$$

sowie $X^{(1)} = 1$ und

$$X^{(2)}, X^{(3)}, X^{(4)} \sim \mathcal{N}(1, 2).$$

Berechnen Sie in jeder Simulation den Schätzer so wie oben beschrieben und den kleinste Quadrate Schätzer in der linearen Regression von Y auf X. Vergleichen Sie die beiden Schätzer auf Basis Ihrer Simulationen. Welchen Schätzer würden Sie bevorzugen?