Statistik Praktikum 2009
Aufgaben

Hajo Holzmann / Florian Ketterer

März 2009

1 Deskriptive Statistik

Exercise 1 (Simpson's Paradoxon): Wir betrachten den Datensatz `tax.dat`.

- Erzeugen Sie Tortendiagramme (Befehl `pie`), einmal für Einkommen und einmal für Steueraufkommen in den Einkommensgruppen in jedem Jahr (4 Plots insgesamt). Die Plots sollen auf einmal erscheinen (Option `mfrow=c(2,2)` bei `par`).
- Erzeugen Sie weiter Tortendiagramme, in denen jeweils der Anteil von (Netto)-Einkommen und Steuer für jede Einkommensgruppe und jedes Jahr abgebildet ist (eine Grafik: Option `mfrow=c(3,4)` bei `par`). Fügen Sie hier die Prozentangaben zu den Plots hinzu (Option `labels` bei `pie`).
- Verfahren Sie ebenso für Balkendiagramme (`barplot`). Erstelle zunächst ein Balkendiagramm, in dem in jeder Income-Bracket das gesamt Einkommen und Steueraufkommen beider Jahre nebeneinander abgebildet sind (Option `besides=TRUE`, als Argument für `barplot` eine Matrix, in denen in den Zeilen die Einkommen und Steueraufkommen der Jahre stehen). Erstellen Sie weiter ein Balkendiagramm, in dem jeweils der Steueranteil in beiden Jahren für jede Einkommensgruppe nebeneinander steht. Würden Sie die Darstellung über Tortendiagramme oder über Balkendiagramme bevorzugen?
- Erklären Sie Simpsons Paradoxon anhand der vorliegenden Daten.

Exercise 2 (Deskriptive Statistik): Wir betrachten in dieser Aufgabe den Datensatz `ZNS.dat` über das zentrale Nervensystem, und dabei die Variable AO.

(a) Plotte Histogramme (`hist`) für die AO Werte der gesunden Patienten beider Altergruppen, so dass die Zellgrößen in beiden Histogrammen gleich sind (übereinander: Option `mfrow=c(2,1)` bei `par`, für gleiche Zellen wähle Option `breaks` von `hist` geeignet). Unterscheidet sich die Verteilung in den beiden Altersgruppen? Macht es Sinn, die Gruppen getrennt zu behandeln, falls diese mit den kranken Patienten verglichen werden sollen?

(b) Wir vergleichen nun die AO Werte der gesunden Patienten (beide Gruppen zusammen) mit denen mit Demenz.

- Berechne Mittelwert, Standardabweichung, Median, Interquartilsabstand sowie Minimum und Maximum in beiden Gruppen (Befehle `summary`, `sd`, `IQR`). Stelle die Ergebnisse in einer Tabelle dar.
- Erzeuge Boxplots für beide Gruppen in einer Graphik (`boxplot`, damit die Boxplots der Daten `x` und `y` in eine Grafik aufgenommen werden, rufe die Funktion durch `boxplot(x,y)` auf).
- Erzeugen Sie einen Quantil-Quantil Plot (`qqplot`), um die Verteilung in beiden Gruppen zu vergleichen. Stellen Sie sicher, dass die Wertebereiche von x und y Achse gleich sind. Fügen Sie die Linie für $y = x$ zum Plot hinzu (`abline`).
- Geben Sie eine Entscheidungsregel, um zwischen den gesunden und dementen Patienten mit Hilfe ihres AO Wertes zu unterscheiden. Dabei sollen die Patienten oberhalb eines gewissen AO Wertes als dement und die unterhalb dieses Wertes als gesund klassifiziert werden. Die erste Entscheidungsregel soll die Summe der Prozente der korrekt klassifizierten Personen in jeder Gruppe maximieren. Die zweite soll die absolute Anzahl der korrekt klassizifierten Personen maximieren. (Durchführung: Sie müssen nur AO Werte, die in einer der beiden Gruppen vorkommen, als mögliche Schwellenwerte prüfen. Durchlaufen Sie diese mit einer `for` Schleife. Für die Überprüfung benötigen Sie eine `if` Abfrage). Welche Entscheidungsregel würden Sie wählen. Wieso?

(c) Plotte die empirischen Verteilungsfunktionen für jede der Gruppen in eine Grafik (die gesunden Patienten in eine Gruppe) (`plot.ecdf` plottet die empirische Verteilungsfunktion. Optionen für hinzufügen: `add=TRUE`. Weitere Optionen: `do.points=FALSE`). Benutze verschiedene Linientypen (Option `lty=i`, für i=2,3,4 in `plot.ecdf`) und füge eine Legende (`legend`) hinzu.

Exercise 3 (Analyse einer Verteilung): Betrachte den Datensatz `faithful` (direkt in R enthalten) und darin die Variable `waiting`.

(a) Erzeuge einen qq-Plot von `waiting` gegen die Normalverteilung (`qqnorm`). Füge eine Linie mit geeigneter Steigung und Achsenabschnitt hinzu (`abline`, was ist der Unterschied zu `qqline`). Sind die Daten normalverteilt? Führe auch den Shapiro-Wilk Test auf Normalverteilung durch (`shapiro.test`). Was ist das Ergebnis?

(b) Plotte die empirische Verteilungsfunktion der `waiting` Daten (`plot.ecdf`). Füge punktweise Konfidenzintervalle zum Niveau $\alpha = 0.05$ and $\alpha = 0.1$, basierend auf dem zentralen Grenzwertsatz und Slutzky's Lemma, hinzu. Stelle sicher, dass die Bänder in $[0, 1]$ enthalten sind. Diese werden erzeugt über `stepfun` und dann geplottet mit `plot.stepfun`. Zum Erzeugen nutzte etwa

für die obere Grenze: `stepfun(sort(waiting),obere.grenze)`, wobei in der Variable `obere.grenze` die oberen Grenzen eingetragen sind (ein Wert mehr als in `waiting` für das linke und rechte Ende!).

(c) Erinnerung: $n\hat{F}_n(x)$ ist binomial verteilt $B(n, F(x))$. Für die Binomialverteilung gibt es geschlossene Ausdrücke für exakte obere und untere Grenzen eines Konfidenzintervalls. Für $\alpha = 0.05$, $n = 50$, $k = 10$, berechne die Pearson-Clopper Grenzen (Formel s. Handout). Überprüfe sie für diese Fälle wie folgt. Die obere Grenze p_u bei k wird derart berechnet, dass wenn $X \sim B(n, p)$, die Wahrscheinlichkeit $P(X \leq k \mid p = p_u)$ kleiner oder gleich $\alpha/2$ ist und so nah an $\alpha/2$ wie möglich ist. Gehe konkret wie folgt vor: schreibe eine Funktion (`function`) in Abhängigkeit von p, die $P(X \leq k \mid p = p_u) - \alpha/2$ (für obige Werte von k, n, α, nutze `pbinom`) zurückgibt. Von dieser Funktion finde die Nullstelle mit der Funktion `uniroot`. Die obere Grenze wird ähnlich berechnet.

(d) Nutze die Pearson-Clopper Grenzen, um punktweise Konfidenzbänder für die Verteilungsfunktion der `waiting` Daten zu berechnen.

(e) Plotte Kern-Dichte Schätzer für die Dichte der `waiting` Daten (`plot(density())`). Nutze dabei den Normalverteilungskern und verschiedene, datengestützt gewählten Bandbreiten (Package `library(KernSmooth)`, Option `bw` von `density`). Insbesondere nutzte normal reference (`bw.nrd`, `bw.nrd0`) cross-validation (`bw.bcv`, `ucv`), solve-the-equation und direct-plug-in (`bw.SJ` mit Optionen von `method`). Nimm eine sinnvolle Anzahl an Schätzungen in die jeweiligen Plots auf (nutze dabei `lines`). Basierend auf der Schätzung mit solve-the-equation Bandbreite, plotte punktweise Konfidenzbänder mit Hilfe der Normalverteilungsapproximation der Verteilung des Kern-Dichte Schätzers Hierzu werden die Schätzwerte des Dichteschätzers benötigt, die durch `density` (Wert `y`) erhalten werden.

(f) Plotte die empirische Verteilungsfunktion der waiting Daten mit *gleichmäßigen* Konfidenzbändern für $\alpha = 0.05$ und $\alpha = 0.1$. Nutze dazu die Tatsache, dass $\sup_t |\hat{F}_n(t) - F(t)|$ nicht von F abhängt falls F stetig ist. Simulieren Sie die Verteilung von $\sup_t |\hat{F}_n(t) - F(t)|$, indem Sie 100000 mal $n = \#\{\text{waiting}\}$ auf $(0, 1)$ gleichverteilte Zufallszahlen ziehen.

2 Statistische Tests

Exercise 4 (Tests für quantitative Daten): In dieser Aufgabe wollen wir verschiedene Test-Verfahren für quantitative Daten vergleichen.

(a) Betrachte den Datensatz `lead.dat`. Berechnet den Mittelwert in jeder Gruppe und erstelle QQ-Plots. Ist die Annahme normalverteilter Daten sinnvoll? Berechne dazu auch die Shapiro-Wilk Statistik. Berechne die Standardabweichungen in den Gruppen und teste (unter der Annahme nicht gepaarter Daten) auf gleiche Varianzen (`var.test`). Sollten die Daten im Folgenden als gepaarte Stichprobe behandelt werden? Verwende den t-Test um zu überprüfen, ob die Daten einen gleichen Erwartungswert haben (`t.test`, gegebenfalls Option `paired=TRUE`). Führe auch den Zwei-Stichproben Wilcoxon Test durch (`wilcox.test`, wieder gegebenfalls Option `paired=TRUE`). Was sind die Ergebnisse?

(b) In diesem Abschnitt wollen wir die Power Eigenschaften des t-Tests mit dem nichtparametrischen Wilcoxon Test vergleichen. Erzeuge dazu 100000 mal Stichproben der Größe $20, 50, 100$ von unabhängigen, normalverteilten Zufallszahlen mit $\mu_1 = 0.3$ and $\sigma_1 = 1$, und teste die Hypothese $H : \mu = 0$ (bzw. Symmetrie um 0 bei dem Wilcoxon Test). Schätze die Power der Verfahren bei einem Test zum Niveau $\alpha = 0.05$ (Anzahl der Ablehnungen durch n), und berechne 95% Konfidenzintervalle für die geschätzte Power (mit Hilfe der Pearson-Clopper Grenzen für die Binomialverteilung). Welcher Test hat eine höhere Power?

(c) Abschließend wollen wir eine Studie auf Robustheit des t-Tests durchführen. Erzeuge dazu 100000 Stichproben der Größen $20, 50, 100$ mit unabhängigen, t-verteilten Zufallszahlen mit 4 Freiheitsgraden (`rt`), und teste die Hypothese $H_0 : \mu = 0$ mit dem t-Test und dem Wilcoxon Test zum nominalen Niveau $\alpha = 0.05$. Schätze das tatsächliche Niveau. Berechne 95% Konfidenzintervalle für die geschätzten Niveaus. Halten die Tests das Niveau ein?

Exercise 5 (Tests auf Korrelation und Unabhängigkeit): In dieser Aufgabe untersuchen wir Korrelation/ und Unabhängigkeit in bivariaten Daten.

(a) Betrachte den Datensatz `wuermer.csv` und darin die Variablen `Ind.Masse` (Gewicht) sowie `Pb` (Bleianreicherung). Wir wollen untersuchen, ob diese Variablen korreliert sind. Erstelle einen Scatterplot der Variablen (`plot`). Berechne den Pearsonschen und den Spearmanschen Korrelationskoeffizienten (`cor`, für Spearman Option `method="spearman"`). Teste jeweils, ob diese gleich 0 sind (`cor.test`, für Spearman Option `method="spearman"`) Was ist das Ergebnis?
Wir wollen nun die Analyse nach der Variable `Gruppe` (Fanggebiet der Fische) unterteilen. Erstellen Sie dazu Scatterplots der beiden Variablen in jeder Ausprägung von `Gruppe` (`coplot`), und führen Sie die beiden Tests auf Korrelation in den Gruppen aus. Wie sind die Ergebnisse zu interpretieren?

(b) Wir wollen nun in einer Simulation sehen, wie das Phänomen im ersten Teil ensteht. Simuliere jeweils $n = 200$ bivariate, normalverteilte Zufallszahlen (`mvrnorm` in der library `MASS`), jeweils mit Varianzen 1 und Kovarianz -0.2, und einmal mit Erwartungswertvektor $(0, 0)$ und einmal mit $(3, 3)$. Füge die beiden Vektoren hintereinander zu einem bivariaten Vektor der Länge 400. Schätze den Pearsonschen und den Spearmanschen Korrelationskoeffizienten und teste, ob diese gleich 0 sind. Was ist das Ergebnis, und was besagt es in Bezug auf den ersten Teil der Aufgabe?

(c) Betrachte den Datensatz `storch.csv` und darin die Variablen `Stoerche` (Anzahl der Störche) sowie `Geburtenrate` (Geburtenrate). Wir wollen untersuchen, ob diese Variablen korreliert sind. Erstelle einen Scatterplot der Variablen. Berechne den Pearsonschen und den Spearmanschen Korrelationskoeffizienten. Tests jeweils, ob diese gleich 0 sind. Was ist das Ergebnis?
Verfahre ebenso für die Koeffizienten aus `Geburtenrate` und `Menschen` (Anzahl der Menschen) sowie `Stoerche` und `Flaeche`.

(d) Wir wollen nun in einer Simulation sehen, wie das obige Phänomen ensteht. Simuliere einmal $n = 200$ und einmal $n = 10$ bivariate, normalverteilte Zufallszahlen, jeweils mit Varianzen 1 und Kovarianz -0.2, und einmal mit Erwartungswertvektor $(0, 0)$ und einmal mit $(5, 5)$. Füge die beiden Vektoren hintereinander zu einem bivariaten Vektor der Länge 210. Schätze den Pearsonschen und den Spearmanschen Korrelationskoeffizienten und teste, ob diese gleich 0 sind. Was ist das Ergebnis, und was besagt es in Bezug auf den dritten Teil der Aufgabe?

Exercise 6 (Standardtests für kategorielle Daten): In dieser Aufgabe betrachten wir verschiedene Tests für kategorielle Daten.

(a) Zwei verschiedene Patientengruppen wurden mit verschiedenen Medikamenten behandelt. In der ersten Gruppe (die Medikament 1 erhielt) wurden 20 Patienten gesund während 100 krank blieben. In der zweiten (die Medikament 2 erhielt), wurden 35 gesund während 110 krank blieben. Geben Sie die Daten in eine Kontingenz-Tafel eine, schätzen Sie die Werte für Heilungserfolg in den beiden verschiedenen Gruppen, und berechnen Sie den P-Wert des exakten Fisher Tests auf Homogenität (benutzen Sie sowohl die Hypergeometrische Verteilung als auch die Funktion `fisher.test`). Kommentieren Sie Ihr Ergebnis.

(b) Betrachten Sie den (künstlichen) Datensatz `dichotomous.txt`. Das Urin von 50 Patienten wurde auf Blut vor und nach der Behandlung mit einem Medikament (0 bedeutet kein Blut, während 1 Blut bedeutet) untersucht. Wir wollen testen, ob die Wahrscheinlichkeit für Blut im Urin durch die Behandlung sinkt. Warum können wir Fishers exakten Test in diesem Setting auf dieses Problem nicht anwenden? Was testet der Fisher Test in dieser Situation und was ist das Resultat? Erstellen Sie eine Kontingenztafel mit den Daten und schätzen Sie die Wahrscheinlichkeiten für Blut im Urin vor und nach der Behandlung.

Um auf ein Sinken dieser Wahrscheinlichkeit zu testen, nutzen Sie den Test von McNemar, in dem Sie einmal direkt die Binomialverteilung benutzen und einmal indem Sie die Funktion `mcnemar.test` benutzen. Kommentieren Sie Ihr Ergebnis.

(c) Betrachten Sie den Datensatz `suicide.dat`. Wir wollen testen, ob männliche und weibliche Selbstmörder dazu tendieren, unterschiedliche Methoden zu verwenden, um ihre Ziele zu erreichen. Berechnen Sie die Häufigkeiten in den beiden Gruppen und zeichnen Sie Tortendiagramme für beide Gruppen. Wenden Sie nun den Chi-Quadrat-Test auf Homogenität an: einmal, indem Sie die Teststatistik direkt ausrechnen und einmal indem Sie die Funktion `chisq.test` verwenden. Kommentieren Sie Ihr Ergebnis.

3 Regression

Exercise 7 (Einfache lineare Regression): In dieser Aufgabe geht es um den Datensatz `hubble`. Dieser Datensatz ist in der Library `gamair` enthalten. Im Folgenden sollen vor allem die folgenden Fragen von Ihnen untersucht werden:

- Welcher Wert von β ist am besten mit den Daten vereinbar?
- Welcher Wertebereich von β ist mit den Daten vereinbar?
- Können bestimmte, beispielsweise aus theoretischen Überlegungen abgeleitete, Werte für β mit den Daten vereinbart werden?

(a) Laden Sie das Paket `gamair` mit dem `library`-Befehl und anschließend den Datensatz `hubble` mit `data`. Verschaffen Sie sich einen ersten Überblick über die Daten. Benutzen Sie hierzu unter anderem die Befehle `help` und `summary`.

(b) Zeichnen Sie ein Streudiagramm der Geschwindigkeiten `y` gegen die Distanzen `x` der einzelnen Galaxien.
Spricht dieses Streudiagramm eher gegen oder eher für das in (2) gegebene Hubble'sche Gesetz?

(c) Schätzen Sie nun β mit Hilfe der so genannten Kleinste-Quadrate Methode, das heißt, passen Sie ein einfaches lineares Modell (ohne Intercept) an die Daten an. Was für eine Einheit sollten β und somit auch das geschätzte $\widehat{\beta}$ vernünftigerweise haben?

(d) Geben Sie die in Teilaufgabe (c) angepasste Regressionsgerade an und zeichnen Sie diese in das Streudiagramm aus Teilaufgabe (b) ein.

(e) Identifizieren Sie mit Hilfe des Befehls `identify` im Streudiagramm diejenigen Punkte, welche vergleichsweise weit von der Regressionsgerade entfernt liegen. Dazu müssen Sie nach Eingabe des `identify`-Befehls mit der linken Maustaste im Schaubild in die Nähe des zu identifizierenden Punktes klicken. Der Identifikationsmodus wird beispielsweise durch anklicken des STOP-Buttons beendet. (Achtung: Solange der Identifikationsmodus aktiv ist, können Sie nicht in R fortfahren!)

(f) Überprüfen Sie, ob die Modellannahmen der einfachen linearen Regression erfüllt sind (Residuenanalyse!).

(g) Passen Sie nun nochmals ein Modell der Form (2) an die Daten an, wobei Sie diesmal die 3. und 15. Beobachtung weglassen.
Führt das Weglassen dieser Beobachtungen zu einer "Verbesserung" der Residuenplots?

(h) Berechnen Sie nun mit Hilfe der in den beiden Modelle aus den Teilaufgaben (c) und (g) geschätzten $\widehat{\beta}$ jeweils das ungefähre Alter des Universums (in Jahren).

(i) Einige Schöpfungswissenschaftler schätzen, basierend auf biblischen Überlegungen, das Alter der Erde auf ca. 6000 Jahre. Welchen Wert für β würde dieses Alter des Universums implizieren (Diese Überlegung ist natürlich nicht ganz richtig, da Schöpfungswissenschaftler nicht von der Urknalltheorie ausgehen)?

(j) Testen Sie die Nullhypothese "Das Universum ist 6000 Jahre alt" auf dem 1%−Signifikanzniveau, wenn Sie annehmen, dass die zufälligen Fehler in Ihrem Regressionsmodell normalverteilt sind.

(k) Geben Sie ein 95%−Konfidenzintervall für das Alter des Universums an.

Exercise 8 (Lineares Modell): Die folgende Tabelle enthält Daten über die Anzahl von Stunden, die 8 Studenten einer Vorlesung außerhalb der Vorlesungsstunden in einem Zeitraum von drei Wochen zum Lernen aufgewendet haben, sowie ihre Punktzahlen, die sie am Ende in der Prüfung erreicht haben.

Lernzeit in Stunden (x)	20	16	34	23	27	32	18	22
Punktzahl in der Prüfung (y)	64	61	84	70	88	92	72	77

Es sei das Modell der klassischen linearen Regression zugrunde gelegt, das heißt für die Fehler gelte die Normalverteilungsannahme.

(a) Geben Sie die Daten in R ein.

(b) Bestimmen Sie die Kleinste-Quadrate Regressionsgerade des linearen Modells $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ und tragen Sie diese in ein Streudiagramm der Daten ein.

(c) Testen Sie auf dem 1%−Signifikanzniveau die Nullhypothese, dass die Steigung der Regressionsgeraden Null ist und interpretieren Sie das Ergebnis.

(d) Prognostizieren Sie mit Hilfe der Regressionsgleichung aus Teilaufgabe a) das Prüfungsergebnis eines Studenten, der $x_9 = 30$ Stunden für das Studium des Vorlesungsmaterials verwendet hat.

(e) Bestimmen Sie ein 90%−Konfidenzintervall für die mittlere Punktzahl von Studenten, die $x_9 = 30$ Stunden gelernt haben.

(f) Bestimmen Sie ein 90%−Prädiktionsintervall (Prediction Interval oder auch Prognoseintervall) für die Punktzahl eines einzelnen Studenten, der $x_9 = 30$ Stunden für die Vorbereitung verwendet hat.

(g) Versuchen Sie die Teilaufgaben d) bis f) mit Hilfe der Funktion `predict()` in R zu lösen und vergleichen Sie die Ergebnisse. Schauen Sie sich hierzu zunächst die Hilfe zu `predict.lm` an.

Exercise 9 (Lineare Regression): In dieser Aufgabe geht es um die Datensätze `sperm.comp1` und `sperm.comp2`, aufgrund welcher Baker und Bellis schlossen, dass die Spermienzahl mit dem Anteil der Zeit, in der die Paare getrennt waren, und dem Gewicht der Frau ansteigt. Dieses Ergebnis sollen Sie im Folgenden (kritisch) überprüfen.

(a) Laden Sie zunächst das Paket `gamair` und dann den Datensatz `sperm.comp1`. Schauen Sie sich die Hilfe zu diesem Datensatz an und verschaffen Sie sich einen ersten Überblick über die Daten. Erstellen Sie dazu unter anderem eine so genannte Scatterplot-Matrix, das heißt Streudiagramme von allen Spalten des Datensatzes gegen alle anderen. In R erhalten Sie eine solche Scatterplot-Matrix mit Hilfe des Befehls `pairs`. Berücksichtigen Sie hierbei die erste Spalte des Datensatzes nicht, da es sich hierbei nur um die Kennzeichnung der einzelnen Paare handelt.

(b) Gemäß Baker und Bellis könnte das folgende Ausgangsmodell vernünftig sein:

$$y_i = \beta_0 + \beta_1\ t_i + \beta_2\ p_i + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2)\ iid,$$

wobei y_i die Spermienanzahl (`count`), t_i die Zeit seit der letzten Kopulation (`time.ipc`) und p_i der Anteil der seit der letzten Kopulation zusammen verbrachte Zeit (`prop.partner`) sind.
Schätzen Sie die Regressionskoeffizienten $\beta_0,\ \beta_1$ und β_2 mit Hilfe der Kleinste-Quadrate Methode und versuchen Sie diese zu interpretieren.

(c) Betrachten Sie den `summary`-Output des mit Hilfe des `lm`-Befehls angepassten Modells aus Teilaufgabe (b) und versuchen Sie alle dort angegebenen Größen "von Hand" (mit R aber ohne die Funktion `lm` oder ähnlichen Befehlen) zu berechnen.

(d) Führen Sie als nächstes eine Residuenanalyse (Plot der Residuen gegen die angepassten Werte (fitted values) und QQ-Plot (Normalplot) der Residuen gegen die Normalverteilung) für das in Teilaufgabe (b) angepasste Modell durch.

(e) Sie erhalten die in Teilaufgabe (d) erstellten Residuenplots neben zwei weiteren Residuenplots auch (einfacher), wenn Sie die Funktion `plot` auf ein `lm`-Objekt anwenden. Dabei werden für die jeweils drei "extremsten" Punkte in jedem der vier Plots die Beobachtungsnummer angezeigt. Gehört eine bestimmte Beobachtung in allen vier Plots jeweils zu diesen drei "extremsten" Beobachtungen? Schauen Sie sich eine solche Beobachtung gegebenenfalls näher an. Wo liegt diese im Streudiagramm von `count` gegen `prop.partner`?

(f) Untersuchen Sie, ob es besser ist, anstatt dem Anteil der seit dem letzten Geschlechtsverkehr zusammen verbrachten Zeit, den absoluten Wert der zusammen verbrachten Zeit (in Stunden) als erklärende Variable zu benutzen. Bestimmen Sie für diese neue erklärende Variable auch den so genannten Leverage der einzelnen Beobachtungen. Gibt es High Leverage Punkte, das heißt Beobachtungen, deren Leverage größer als der doppelte Mittelwert der Leverages von allen Beobachtungen ist?

(g) Betrachten Sie nun noch mal den `summary`-Output aus Teilaufgabe (c). Hat eine der beiden erklärenden Variablen `time.ipc` bzw. `prop.partner` einen

$p-$Wert der größer als 0.05 ist? Passen Sie gegebenenfalls ein neues Modell ohne diese Kovariable an die Daten an. Ist dieses neue Modell besser als das ursprünglich von Baker und Bellis vorgeschlagene Modell aus Teilaufgabe (b)?

(h) Für welches Modell entscheiden Sie sich, wenn Sie Ihre Modellwahl auf Basis der Informationskriterien von Akaike (AIC) bzw. Bayes (BIC) durchführen.

(i) Wie würden Sie abschließend auf Basis der Daten des Datensatzes `sperm.comp1` die These von Baker und Bellis, dass die Anzahl der produzierten Spermien in `time.ipc` ansteigt und in `prop.partner` fällt, beurteilen?

(j) Betrachten Sie jetzt den zweiten Datensatz `sperm.comp2` von Baker und Bellis. Laden Sie diesen zunächst und verschaffen Sie sich wieder einen ersten Überblick über die Daten. Was fällt Ihnen beim Betrachten dieses Datensatzes auf?

(k) Passen Sie zunächst das folgende Modell, welches alle erklärende Variablen (bis auf die Anzahl `n` der Kopulationen und die Paarkennzeichnung `pair`) enthält, an die Daten an:

$$\begin{aligned}\texttt{count}_i =& \beta_0 + \beta_1\, \texttt{f.age}_i + \beta_2\, \texttt{f.weight}_i + \beta_3\, \texttt{f.height}_i + \beta_4\, \texttt{m.age}_i \\ &+ \beta_5\, \texttt{m.weight}_i + \beta_6\, \texttt{m.height}_i + \beta_7\, \texttt{m.vol}_i + \epsilon_i\end{aligned}$$

Führen Sie auch wieder eine Residuenanalyse durch. Fällt eine Beobachtung besonders auf?

(l) Betrachten Sie den `summary`-Output des in Teilaufgabe (k) angepassten Modells. Lassen Sie nun sukzessive jeweils diejenige erklärende Variable im Modell weg, die den größten $p-$Wert hat. Fahren Sie so lange damit fort, bis die $p-$Werte aller verbliebenen Kovariablen kleiner als 0.05 sind. Mit welchem Modell endet dieses Vorgehen?

(m) Führen Sie die Modellwahlstrategien Backward Elimination bzw. Forward Selection jeweils für das AIC und BIC mit Hilfe des `step`-Befehls durch. Hierzu müssen Sie erst alle Beobachtungen für die nicht alle Variablen zur Verfügung mit `na.omit` entfernen. Welche Modelle liefern Ihnen jeweils diese Verfahren für die Variablenselektion.

(n) Wiederholen Sie die Teilaufgabe (l), wenn Sie die auffällige Beobachtung in den Residuenplots aus Teilaufgabe (k) weglassen. Welche erklärenden Variablen sind jetzt in Ihrem finalen Modell enthalten?

Exercise 10 (Logistische Regression): In dieser Aufgabe sollen Sie die Daten aus der Studie von Smith aus dem Jahr 1967 näher untersuchen. Es wäre wünschenswert, wenn man anhand des CK-Wertes eines Patienten eine Wahrscheinlichkeit angeben könnte, mit der dieser einen Herzinfarkt hat.

(a) Zeichnen Sie zunächst die so genannte Responsefunktion (Antwortfunktion) der logistischen Regression, das heißt die Umkehrfunktion der Logit-Linkfunktion

$$F(p) = \log\left(\frac{p}{1-p}\right)$$

im Bereich -6 bis 6.

(b) Erstellen Sie mit einem beliebigen Editor (zum Beispiel Excel) eine Text-Datei, welche die Daten der Studie von Smith enthält und lesen Sie diese Datei anschließend mit dem Befehl `read.table` in R ein.

(c) Erstellen Sie mit Hilfe des Befehls `mosaicplot` einen so genannten Mosaikplot um zu überprüfen, ob sich in den verschiedenen CK-Gruppen die Wahrscheinlichkeiten für einen Herzinfarkt unterscheiden. Liefert Ihnen die Funktion `prop.test`, mit der Sie die Hypothese, dass für alle CK-Levels die Wahrscheinlichkeiten für das Vorliegen eines Herzinfarkts gleich groß sind, testen können, das gleiche Ergebnis? Testen Sie jetzt auch noch mit der Funktion `pairwise.prop.test` für welche CK-Gruppen sich die Wahrscheinlichkeiten signifikant unterscheiden.

(d) Bestimmen Sie für die verschiedenen CK-Werte jeweils die (empirische) Wahrscheinlichkeit, dass ein Patient einen Herzinfarkt erlitten hat und plotten Sie diese Wahrscheinlichkeiten gegen die entsprechenden CK-Werte. Macht ein lineares Modell hier Sinn?

(e) Passen Sie mit Hilfe des `glm`-Befehls ein geeignetes Modell an die Daten an und bestimmen Sie die geschätzten Parameter des Modells. Zeichnen Sie anschließend die angepasste Regressionskurve (beispielsweise mit dem Befehl `lines`) in das Streudiagramm aus Teilaufgabe (d) ein.

(f) Bei wie vielen der in den einzelnen Gruppen für die verschiedenen CK-Werte beobachteten Patienten würden Sie aufgrund Ihres in Teilaufgabe (e) angepassten Modells einen Herzinfarkt erwarten. Vergleichen Sie diese Werte auch mit den tatsächlich beobachteten Zahlen.

(g) Bewerten Sie die Güte des angepassten Modells (Devianz und Residuenanalyse). Was fällt Ihnen bei den Residuenplots besonders auf?

(h) Passen Sie nun ein logistisches Modell mit Termen bis zum Grad 3 in CK an die Daten an. Vergleichen Sie dieses Modell mit dem vorherigen Modell.

(i) Testen Sie die Nullhypothese, dass das Modell aus Teilaufgabe (e) die Daten ausreichnend gut beschreibt, gegen die Alternative, dass das Modell aus Teilaufgabe (h) benötigt wird.

Exercise 11 (Verallgemeinerte lineare Modelle): In dieser Aufgabe geht es um die Aids-Daten aus Belgien, welche man bei Venables und Ripley findet. Im Folgenden sollen Sie untersuchen, ob diese Daten einen Anhaltspunkt für eine Verlangsamung des Anstiegs der Neuerkrankungen geben.

(a) Geben Sie die Daten in R ein.

(b) Zeichnen Sie ein Streudiagramm der Neuerkrankungszahl gegen die Zeit. Wählen Sie hierbei den Bereich $[0, 280]$ für die Y-Achse.

(c) Linearisieren Sie das Modell (3). Welche Linkfunktion erhalten Sie?

(d) Passen Sie nun mit Hilfe des `glm`-Befehls ein geeignetes Modell an die Daten an.

(e) Führen Sie eine Residuenanalyse für Ihr in Teilaufgabe (d) angepasstes Modell durch. Sind Sie mit den Residuenplots zufrieden? Überlegen Sie sich, wie Sie weiter vorgehen könnten.

(f) Ergänzen Sie Ihr Modell noch um einen quadratischen Term in der Zeit und passen Sie auch dieses Modell an die gegebenen Daten an. Beschreibt dieses Modell die Daten besser? (Begründen Sie Ihre Antwort kurz!)

(g) Zeichnen Sie nochmals das Streudiagramm aus Teilaufgabe (b) und ergänzen Sie die beiden angepassten Regressionskurven der Modelle aus den Teilaufgaben (d) und (f).

(h) Testen Sie die Hypothese: "Das einfachere (Teil-) Modell beschreibt im Vergleich zum komplizierteren (Ober-) Modell die Daten ausreichend gut". Verwenden Sie hierzu einen verallgemeinerten Likelihood-Quotiententest, welchen Sie in R bekommen, indem Sie im `anova`-Befehl die Option `test="Chisq"` wählen.

4 ANOVA

Exercise 12 (Einfaktorielle Varianzanalyse): Wir betrachten den Datensatz `InsectSprays`, der in R enthalten ist.

- Erstelle Boxplots der Variable `count` in jeder Ausprägung des Faktors `spray`, so dass alle in einer Grafik enthalten sind. Berechne Mittelwert und Standardabweichung von `count` in jeder Ausprägung von `spray` (nutze `tapply`), und erstelle entsprechende separate QQ-Plots.

- Führe eine Varianzanalyse durch (`aov` und dann `summary`). Erstelle QQ-plots der Residuen (`residuals` für das ANOVA Objekt) sowie einen Plot der Fitted Values (`fitted.values`) gegen die Residuen. Sind die Voraussetzungen der Varianzanalyse erfüllt? Führe auch den Bartlett Test (`bartlett.test`) auf Gleichheit der Varianzen durch. Was sind die Ergebnisse?

- Als Alternative zur Varianzanalyse führe den Kruskal-Wallace Test durch (`kruskal.test`) Vergleiche das Ergebnis mit dem der Varianzanalyse.

- Um herauszufinden, in welchen Merkmalen Unterschiede vorliegen, können paarweise Tests durchgeführt werden. Dazu nutzte den paarweisen t-Test (`pairwise.t.test`) sowie den paarweisen Wilcoxon Test (`pairwise.wilcox.test`), jeweils mit der Bonferroni Korrektur (Option `method="bonferroni"`). Erstelle weiter Konfidenzintervalle für die Lageunterschiede mit der Tukey Methode (`TukeyHSD`). Vergleiche die Ergebnisse. Werden die Erwartungen aus der deskriptiven Analyse erfüllt?

Exercise 13 (Mehrfaktorielle Varianzanalyse): In dieser Aufgabe betrachten wir verschiedene Datensätze, die zu mehrfaktorieller Varianzanalyse führen. Die Datensätze sind in der Library `faraway` enthalten.

- Betrachte den Datensatz `pvc`, die Zielgröße ist `psize`, die Faktoren sind `operator` und `resin`. Erstelle Boxplots von `psize`, jeweils separat für jeden der beiden Faktoren. Erstelle weiter Interaktionsplots (`interaction.plot`). Welches faktorielle Design liegt vor? Führe eine Varianzanalyse durch. Ist der Interaktionseffekt signifikant? Führe eine Residuenanalyse durch. Reduziere das Modell um den Interaktionseffekt, und führe danach abermals eine Varianzanalyse und eine Residuenanalyse durch. Was sind die Ergebnisse?

- Betrachte den Datensatz `oatvar`, die Zielgröße ist `yield`, `variety` ist ein Faktor und `block` eine Blockvariable. Erstelle Boxplots von `yield`, jeweils separat für Faktor und Block Variable. Welches faktorielles Design liegt vor? Führe eine Varianzanalyse durch. Führe eine Residuenanalyse durch.

- Betrachte den Datensatz `abrasion`, die Zielgröße ist `wear`, `material` ist ein Faktor und `run`, `position` sind Blockvariable. Erstelle Boxplots von `wear`, jeweils separat für Faktor und Block Variable. Welches faktorielles Design liegt vor? Führe eine Varianzanalyse durch. Führe eine Residuenanalyse durch.

5 Maximum Likelihood

Exercise 14 (Mischungen von Normalverteilungen): In dieser Aufgabe betrachten wir Zweikomponenten Mischungen von Normalverteilungen

(a) Mischungen werden häufig zur Modellierung von multimodalen Daten benutzt. Wir betrachten Zweikomponenten Mischungen. Diese müssen jedoch nicht bimodal sein. Plotte dazu die Dichte einer Zweikomponentenmischung von zwei Normalverteilungen mit $p = 0.5$, $\mu_1 = 0$, $\sigma_1 = 2$, $\sigma_2 = 1$ und $\mu_2 = 1, 2, 3, 4$ in einen Plot mit verschiedenen Linientypen (`dnorm` gibt die Dichte der Normalverteilung). Gib eine sinnvolle Schätzung, für welches μ_2 die Mischung bimodal wird (mit Intervallschachtelung).

(b) Betrachte den Datensatz `faithful` und die Variable `waiting` (vgl. Aufgabe 3). Plotte eine Dichteschätzung der Daten. Teile den Datensatz in zwei Gruppen, einmal für Werte < 67, und einmal für ≥ 67. Berechne Mittelwert und Standardabweichung in jeder Gruppe. Plotte qqplots in jeder Gruppe gegen die Normalverteilung. Ist es sinnvoll, komponentenweise eine Normalverteilung anzunehmen?
Passe nun eine Zweikomponentenmischung mit Maximum Likelihood an. Schreibe dazu eine Funktion, die für gegebene Parameterwerte die log-Likelihood berechnet. Maximiere diese Funktion numerisch mit der Funktion `nlm` Transformiere dabei die Parameter geeignet, so dass ihr Wertebereich unbeschränkt wird (benutze `log` and `logit`). Als Startwerte für die numerische Maximierung benutze obige Schätzwerte in den Gruppen. Mache dasselbe unter der Annahme gleicher Varianzen. Führe den Likelihood Quotiententest auf Gleichheit der Varianzen durch. Sind die Vorraussetzungen dafür erfüllt? Füge die geschätzten parametrischen Dichten (mit und ohne gleiche Varianzen) zu dem Plot des Dichteschätzers hinzu.

(c) Wir wollen in obiger Situation nochmals die Verteilung des Likelihood Quotienten Tests auf gleiche Varianzen simulieren. Erzeuge dazu 10000 Stichproben der Länge 200 einer Zweikomponenten Mischung mit $p = 0.5$, $\mu_1 = 0$, $\mu_2 = 3$, $\sigma_1 = \sigma_2 = 1$. Schätze jeweils die Parameter, einmal unter der Annahme gleicher Varianzen und einmal für ungleiche Varianzen, und berechne jeweils die LQT Statistik. Plotte deren empirische Verteilungsfunktion, und füge die Verteilungsfunktion der χ^2_1-Verteilung zu dem Plot hinzu.

Exercise 15 (Markov Ketten): In dieser Aufgabe betrachten wir die Schätzung der Übergangswahrscheinlichkeiten einer Markov Kette.

(a) Betrachte eine Markov Kette mit drei Zuständen und folgender Übergangsmatrix

$$\begin{pmatrix} 0.2 & 0.6 & 0.2 \\ 0.4 & 0.3 & 0.3 \\ 0.2 & 0.3 & 0.5 \end{pmatrix}.$$

Berechne numerisch die stationäre Verteilung (mit `solve`). Simuliere 1000 Stichproben dieser Markov Kette der Länge 250, und berechne jeweils die Anzahl der Übergänge $f_{i,j}$ von Zustand i nach Zustand j, $i, j = 1, 2, 3$. Berechne den ML Schätzer in jeder Stichprobe, berechne deren Mittelwerte, Standardabweichungen, und plotte qq Plots gegen die Normalverteilung sowie Dichte Plots.

(b) Betrachte die Statistiken $\xi_{i,j} = (f_{i,j} - f_i p_{i,j})/f_i^{1/2}$, wobei $f_i = f_{i,1} + f_{i,2} + f_{i,3}$, $i, j = 1, 2, 3$. Diese konvergieren gegen eine multivariate Normalverteilung mit Erwartungswertvektor 0 und Kovarianzmatrix $(\lambda_{ij,kl})$, wobei

$$\lambda_{ij,kl} = \delta_{i,k}(\delta_{j,l} p_{i,j} - p_{i,j} p_{i,l}). \tag{1}$$

Interpretiere die Kovarianzstruktur (1). Schätze diese Kovarianzmatrix (`cov` auch mehrdimensional) und vergleiche mit dem theoretischen Wert.

(c) Für die Schätzung jeder Zeile der Übergansmatrix entspricht die asymptotische Verteilung der Schätzer genau der der Multinomialverteilung (mit drei Zuständen), und die Schätzer für die verschiedenen Spalten sind asymptotisch unabhängig. Daher kann als Anpassungstest die gewöhnliche χ^2 Statistik verwendet werden,

$$Y_i = \sum_j \frac{(f_{i,j} - f_i p_{i,j})^2}{f_i p_{i,j}}, \quad , i = 1, 2, 3.$$

Diese haben asymptotisch unter der Hypothese, dass die $p_{i,j}$ die wahren Übergänge sind, eine χ^2 Verteilung mit zwei Freiheitsgraden, und sind asymptotisch unabhängig. Berechne diese Statistiken für $i = 1, 2, 3$ für die oben simulierten Daten und die wahren Übergangswahrscheinlichkeiten (benutzte `chisq.test`), und vergleiche deren empirische Vetrielungsfunktion mit der theoretischen Verteilungsfunktion. Mache das gleiche für $Y = Y_1 + Y_2 + Y_3$, welche asymptotisch χ^2 verteilt ist mit 6 Freiheitsgraden.

(d) Konstruiere eine χ^2 Statistik, welche bei zwei Markov Ketten mit drei Zuständen vergleicht, ob die Übergangswahrscheinlichkeiten gleich sind. Wie viele Freiheitsgrade hat die asymptotische χ^2 Verteilung? Simuliere diese für zwei Stichproben (wieder 1000 mal Länge 250) aus der obigen Übergangsmatrix (es kann einfach wieder `chisq.test` verwendet werden).