Aufgabenblatt

Die Aufgaben können in **2er Gruppen** bearbeitet werden. Für jeden Aufgabenblock sollen Sie ein R Skript erstellen, das Sie als „`Blocknr.Vorname1.Vorname2.R`“ speichern, also z.B. „`3.Bernd.Ute.R`“ für das R Skript zum Block 3. Bitte Kennzeichnen Sie in ihren Skripten deutlich den Anfang einer neuen Aufgabe und eines neuen Aufgabenteile.
Abgabe: **01.04.2015 bis 19:00 Uhr** per Email.

9 Lineare Modelle

Aufgabe 9.1. (6 Prozent)
Machen Sie sich die Schranken für r_{XY} plausibel (wann spricht man von starker Korrelation etc.), indem Sie 10 Samples x und y von jeweils 100 $N(0,1)$-verteilten Zufallsgrößen erstellen, und die empirischen Korrelationskoeffizienten von x und y sowie von y und $z = 2y + 1 + x$ berechnen.

Aufgabe 9.2. (8 Prozent)
Die 7 Teilnehmer eines Turnwettkampfes erhielten die folgenden Wertungen:

Teilnehmer	1	2	3	4	5	6	7
Reck	9.3	8.6	9.2	9.1	9.0	9.5	8.7
Barren	9.1	8.9	9.0	8.9	8.7	9.4	8.7

(a) Berechnen Sie die Ränge der `Reck`- und `Barren`-Vektoren. Welchen Rang erhalten gleiche Einträge?

(b) Bestimmen Sie den Spearman'schen Korrelationskoeffizienten beider Merkmale und interpretieren Sie das Ergebnis. Bestätigen Sie ihre Vermutung durch einen Plot.

Aufgabe 9.3. (6 Prozent)
Geben Sie stichpunktartig *jeweils* ein Beispiel für Schein- und verdeckte Korrelation an. Diese sollen nicht direkt aus dem Skript stammen.

Aufgabe 9.4. (6 Prozent)
Nutzen Sie die Ergebnisse des Fragebogens, um die Zielgröße *Gewicht* auf einen linearen Zusammenhang mit der Einflussgröße *Körpergröße* zu untersuchen. Inwiefern sind die Ergebnisse realistisch, welche Verzerrungen können durch das Umfragedesign aufgetreten sein?

Aufgabe 9.5. (12 Prozent)
In der in `R` bereits eingebauten Datentabelle `Orange` ist die Größe des Stammumfangs (in mm) von fünf Orangenbäumen zu jeweils sieben gleichen Zeitpunkten (Alter des Baumes in Tagen) angegeben.

(a) Erstellen Sie zwei 7-elementige Vektoren `alter` und `groesse`, wobei `alter` die Zeitpunkte der Messungen und `groesse` den mittleren Umfang der 5 Baumstämme zum jeweiligen Zeitpunkt enthält.

(b) Erstellen Sie ein Streudiagramm (x-Achse: `alter`, y-Achse: `groesse`) und ein lineares Regressionsmodell (mit `lm()`) `model` der Daten. Welchen Wert haben die KQS `alpha` und `beta`? Zeichnen Sie anschließend die Regressionsgerade rot in das Streudiagramm ein.

(c) Berechnen Sie SQT, SQE, SQR und das Bestimmtheitsmaß R^2. Prüfen Sie die Gleichungen $\mathrm{SQT} = \mathrm{SQE} + \mathrm{SQR}$ sowie $R^2 = r^2$.

Aufgabe 9.6. (12 Prozent)
Auf der Homepage finden Sie die Datei `Elektrolyse.txt`, in der Messungen zum Widerstand R von Natronlauge bei unterschiedlichen Konzentrationen c von NaOH aufgeführt sind. Dazu wurde bei einstellbarer Stromstärke I die Spannung U zwischen Anode und Kathode gemessen. Nach dem Ohm'schen Gesetz gilt die Beziehung $U = R \cdot I$ und wir können eine lineare Regression durchführen, um Werte für den Widerstand R bei unterschiedlichen Konzentrationen zu berechnen.

(a) Führen Sie dies für die Konzentrationen $c = 3$, $c = 6$ und $c = 40$ durch, wobei Sie Regressionsgeraden durch den Ursprung berechnen (Aufruf `lm(y~x+0)`).

(b) Plotten Sie die entsprechenden Messwerte und die erhaltenen Regressionsgeraden in gemeinsame Grafiken.

(c) Erstellen Sie die 3 Residualplots und interpretieren Sie diese.

(d) Wie lauten die jeweiligen KQ-Schätzungen für den Widerstand R? Plotten Sie diese Werte gegen die jeweilige Konzentration c.

Aufgabe 9.7. (12 Prozent)
Auf der Homepage finden Sie die Datei `luftuntersuchung.csv`, in der die Schwefeldioxidkonzentration- Y und Temperaturmessungen X einer Luftmessstation in München aufgeführt werden. Es wurden dabei an 14 Tagen die gemessenen Tagesmittelwerte gebildet (SO_2-Werte wurden logarithmiert). Die Spalte `we` gibt an, ob am Wochenende gemessen wurde (`we = 1`) oder nicht (`we = 0`).

(a) Lesen Sie die Daten ein und erstellen Sie ein Streudiagramm von X gegen Y. Färben Sie die Messpunkte, die am Wochenende entstanden sind, dabei rot.

(b) Passen Sie die Daten an ein multiples Regressionsmodell $Y = \theta_1 + \theta_2 \cdot X + \theta_3 \cdot \mathrm{WE}$ an und bestimmen Sie den KQS für θ.

(c) Interpretieren Sie die Werte für θ_2 und θ_3. Gehen beide Parameter signifikant in die Modellanpassung ein?

Aufgabe 9.8. (18 Prozent + 4* Prozent)
In New York wurden in den Monaten Mai bis September im Jahr 1973 täglich Messungen zur Luftqualität vorgenommen. Untersucht wurden dabei der Ozonwert `Ozone`, Sonneneinstrahlung `Solar.R`, Windgeschwindigkeit `Wind` und die Temperatur `Temp` und das Datum wurde in der Reihenfolge Monat, Tag notiert. Diese Daten sind in dem eingebauten Datensatz `airquality` enthalten. Stellen Sie ein Regressionsmodell für die Ozonwerte auf:

(a) Untersuchen Sie mit Hilfe von `cor` und Streudiagrammen die Abhängigkeit der Ozonbelastung von jeweils einer der übrigen meteorologischen Variablen. Durch welche Variablen wird die Ozonkonzentration in der Luft gut über einen linearen Zusammenhang erklärt? Stellen Sie eine Vermutung auf. (Beachten Sie, dass nicht alle Messwerte vorliegen, und schauen Sie in der Hilfe nach, wie mit `NA`-Werten umgegangen werden kann.)

(b) Erstellen Sie mit der Funktion `step` ein geeignetes Regressionsmodell `ozon.lm` für die Ozonbelastung in Abhängigkeit der anderen Einflussgrößen. Stimmt dieses Modell mit dem von Ihnen vermuteten überein?

(c*) Führen Sie die folgenden `R`-Befehle aus: `par(mfrow=c(2,2)); plot(ozon.lm)`. Welche Grafiken erhalten Sie als Ausgabe und was beschreiben sie?

Aufgabe 9.9. (20 Prozent)
Lesen Sie den auf der Homepage verlinkten Datensatz `Patentdaten` ein. Darin sind Daten im Zusammenhang mit Patentanträgen beim Europäischen Patentamt aus der Biotechnologie-Pharmazie sowie der Computertechnologie erfasst. Untersuchen Sie die Wirkung der verschiedenen erfassten Einflussgrößen auf die Wahrscheinlichkeit der Zielgröße *Einspruch gegen das Patent ja / nein* in einem binären Regressionsmodell, indem Sie wie folgt vorgehen:

(a) Untersuchen Sie zunä chst den Datensatz auf untypische Daten, indem Sie sich die `summary` anzeigen lassen, und identifizieren Sie (bspw. mittels Boxplots) zwei Einflussgrößen mit deutlichen Ausreißern.

(b) Schätzen Sie mit `glm()` den Einfluss der Größen *jahr, azit, ansp, uszw, patus, patdsg, aland* auf die Wahrscheinlichkeit für einen Einspruch bei einem Computertechnologie-Patent (`biopharm==0`) in einem Logit-Modell.

(c) Führen Sie die gleiche Schätzung mit zensierten Daten durch: Bestimmen Sie für die in (a) ausgewählten Einflussgrößen 99%-Quantile, und nehmen Sie nur die Datensätze in die Untersuchung mit auf, bei denen diese Merkmale kleiner gleich diesen Quantile sind.

Die erfassten Größen haben folgende Bedeutungen (1 = Ja, 0 = Nein):

einspruch: Einspruch gegen Patent
uszw: Es gibt ein US-Zwillingspatent
patdsg: Patentinhaber aus D, CH, GB
azit: Anzahl Zitationen des Patentes
ansp: Anzahl Patentansprüche
biopharm: Patent aus Biotechnologie-Branche
patus: Patentinhaber aus den USA
jahr: Jahr der Patenterteilung
aland: Länder in denen Patentschutz gelten soll