Aufgabenblatt

Die Aufgaben können in **2er Gruppen** bearbeitet werden. Für jede Aufgabe sollen Sie ein R Skript erstellen, das Sie als „`Aufgabennr.Vorname1.Vorname2.R`“ speichern, also z.B. „`3.2.Bernd.Ute.R`“ für das R Skript zur Aufgabe 3.2. Abgabe: **06.10.2010 bis 17:00 Uhr** per Email.

8 Lineare Modelle

Aufgabe 8.1. (4 Punkte)
Verifizieren Sie die Formeln für $\hat{\alpha}$ und $\hat{\beta}$ durch eine Kurvendiskussion der SQT-Funktion.

Aufgabe 8.2. (2 Punkte)
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 8.3. (5 Punkte)
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 Ohmschen Gesetz gilt die Beziehung

$$U = R \cdot I,$$

wir können also lineare Regression durchführen, um Werte für den Widerstand R 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.

Aufgabe 8.4. (3 Punkte)
Nutzen Sie die Ergebnisse des Fragebogens, um die Zielgröße *Gewicht* auf einen linearen Zusammenhang mit der Einflussgröße *Körpergröße* zu untersuchen. Sind die Zusammenhänge signifikant?
Inwiefern sind die Ergebnisse realistisch, welche Verzerrungen können durch das Umfragedesign aufgetreten sein?

Aufgabe 8.5. (8 Punkte)
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 Streudiagrammen (`plot`) 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?

(b) Wählen Sie drei Variablen aus, von denen Sie den größten Einfluss vermuten. Berechnen Sie dazu mittels `cor()` die entsprechenden Korrelationskoeffizienten. Beachten Sie, dass nicht alle Messwerte vorliegen, und schauen sie in der Hilfe nach, wie mit `NA`-Werten umgegangen werden kann.

(c) Erstellen Sie ein Regressionsmodell für die Ozonbelastung abhängig von den in (b) ausgewählten Einflussgrößen. Gehen Sie dabei schrittweise vor, indem Sie mit der signifikantesten Variable beginnen und jeweils die nächstsignifikante Variable zum Modell hinzufügen.

Aufgabe 8.6. (8 Punkte)
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 generalisierten linearen Modell, indem Sie wie folgt vorgehen:

(a) Untersuchen Sie zunächst 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) : *einspruch* - Einspruch gegen das Patent, *biopharm* - Patent aus der Biotechnologie-Branche, *uszw* - es gibt ein US-Zwillingspatent, *patus* - Patentinhaber aus den USA, *patdsg* - Patentinhaber aus D, CH, GB, *jahr* - Jahr der Patenterteilung, *azit* - Anzahl Zitationen dieses Patentes, *aland* - Länder, in denen Patentschutz gelten soll, *ansp* - Anzahl Patentansprüche.