Aufgabenblatt
Abgabe: bis Montag, 11.01.2010; 10:00 Uhr

Diese Zettel sollen Sie mit dem Statistikprogramm R vertraut machen. Zunächst werden die grundlegende Bedienung und wichtige Befehle erklärt. Im Anschluss daran finden sich Aufgaben.
Durch das regelmäßige Bearbeiten dieser Aufgaben kann ein Teilnahmeschein erworben werden.
Alle Aufgaben sind Programmieraufgaben. Bitte kopieren Sie die eingegebenen Befehle und die Ausgaben von R in eine Textdatei (Kommentare können hinter der Raute `#` eingegeben werden), und senden Sie diese zur Abgabe per e-Mail an
`mentemeier@uni-muenster.de`
Eine schriftliche Abgabe ist nicht erforderlich.

Das Statistikprogramm R ist frei verfügbar, und kann von `http://www.r-project.org/` heruntergeladen werden. Zum Starten in der Uni siehe gesonderte Anleitung.

Theorie

Lineare Modelle

(Homoskedastische) Lineare Modelle umfassen lineare und polynomiale Regression, aber auch Modelle mit qualitativen Faktoren; gemeinsam ist, dass nur linear auftretende Koeffizienten zu schätzen sind.

Bei polynomialer Regression wird zwischen zwei Datensätzen x und y ein funktionaler (polynomialer) Zusammenhang vermutet, d.h. ein Zusammenhang der Form

$$y_i = \theta_0 + \theta_1 x_i + \theta_2 x_i^2 + \cdots + \theta_k x_i^k + \epsilon_i,$$

wobei die Fehler ϵ_i Zufallsgrößen mit Erwartungswert 0 und unbekannter Varianz σ^2 sind, die bspw. Messfehler in der Physik repräsentieren.

*L*ineare *M*odelle werden mit `lm()` untersucht, als Argument übergeben wir unsere Vermutung über den Zusammenhang: Schätzer für das lineare Modell

$$y_i = \theta_1 + \theta_2 x_i + \theta_3 x_i^2 + \theta_4 x_i^3 \epsilon_i$$

berechnen wir beispielsweise mit

`lm(y~I(x)+I(x^2)+I(x^3))`.

Beachten Sie, dass wir die Verschiebung θ_0 nicht erwähnen müssen, und dass jeder Term, vor dem ein Koeffizient θ_i auftaucht, mit `I()` umschlossen wird. Als Ergebnis erhalten wir

zunächst eine Wiederholung unserer Eingabe, und dann die geschätzten Werte für θ_1 bis θ_4. Sie werden als Koeffizienten des zugehörigen Monoms aufgeführt, d.h. θ_1 als Achsenabschnitt steht unter `(Intercept)`, und θ_2, als Koeffizient von `x`, steht unter `x` usw. Ergänzen wir im Befehlsaufruf noch `+0` auf der rechten Seite, so wird eine Regressionskurve durch den Nullpunkt bestimmt.

Mit `plot(x,y)` können wir die Punkte zeichnen lassen. Die Regressionskurve können wir in den Plot nun mittels

```
matlines(x,predict(lm(y~I(x)+I(x^2)+I(x^3))))
```

einfügen. Wie der Name schon andeutet, berechnet die Funktion `predict()` aus den geschätzten Werten für θ die Werte an den Stellen aus x.

Auch multivariate Modelle können mit `lm()` untersucht werden, eine Vermutung

$$y_i = \theta_0 + \theta_1 x_i + \theta_2 z_i + \epsilon_i$$

würden wir beispielsweise mit

```
lm(y ~ x+z)
```

untersuchen. Beachten Sie, dass hier, da nur zwei unterschiedliche Einflussgrößen auftauchen, auf `I()` verzichtet werden kann.

Aufgaben

Eine Aufgabe zu polynomialer Regression finden Sie bei Interesse auf den Übungen zur Statistik, Blatt 10.

Aufgabe 1. (4 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.

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

Hinweis: Teildatensätze eines data.frames erhalten Sie mit dem Befehl `subset(Datensatz, Auswahlbedingung)`. Beachten Sie, dass der Vergleichsoperator in `R` das doppelte Gleichheitszeichen `==` ist.

Aufgabe 2. (6 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 3. (5 Punkte)
Im Rahmen der Erforschung von Arbeitsbedingungen in der Glasindustrie sollte untersucht werden, ob sich die Belastung der Arbeiter am Ende der täglichen Arbeitszeit von der zu Beginn unterscheidet. Als Indikator wurde die Arbeitspulsfrequenz während des ersten und des letzten Schichtdrittels gewählt.

Die Ausgangsfrage wird nun dahingehend spezifiziert, dass zu prüfen ist, ob die Differenzen der Arbeitspulsfrequenzen im Mittel gleich null sind, d. h. formal, ob gilt:

$$\mu = 0.$$

Als Basis für die Prüfung dienen die Messungen der Differenz X der Arbeitspulsfrequenz bei 28 Arbeitern, die Sie in der Datei `Arbeitspulsdiff.txt` finden.

Die Daten können als realisierte Stichprobe aus einer Normalverteilung angesehen werden. Wählen Sie einen geeigneten Test und testen Sie (mit Signifikanzniveau $\alpha = 0.05$), ob die Nullhypothese $\mu = 0$ verworfen werden muss. Kann die Nullhypothese (zum selben Signifikanzniveau) verworfen werden, wenn man die Nullhypothese $\mu \geq 0$ wählt und einseitig testet?

Aufgabe 4. (5 Punkte)
In einem Isotopenlabor werden die Zerfallszeiten (die Zeit zwischen Messbeginn und Zerfall) von 100 Atomen eines instabilen Isotops gemessen. Die Ergebnisse dieser Messungen (in 10^{-6} Sekunden) finden sich in der Datei `Zerfall.txt`. Theoretische Vorüberlegungen haben ergeben, dass die Lebensdauer eines Teilchens exponentialverteilt ist. Aufgrund der Gedächtnislosigkeit der Exponentialverteilung können die in der Datei angegebenen Zeiten als Lebensdauern (die Zeit zwischen Entstehen und Zerfall) aufgefasst werden. Der Zerfallsparameter $\lambda > 0$ ist unbekannt. Kann zum Niveau $\alpha = 0.05$ absichert werden, dass durchschnittlich weniger als einen Zerfall pro 10^{-6} Sekunden vorliegt, d. h. dass $\lambda \geq 1$ gilt?