Angewandte Regression — Serie 3

Für die ersten drei Aufgaben benützen wir den Datensatz `catheter`. Es handelt sich um Daten aus der Medizin. Die Variable `x1` ist eine charakteristische Länge am Körper (in cm), `x2` das Gewicht eines Patienten (in kg) und `y` die optimale Länge eines Katheters (in cm), der für eine Herzoperation verwendet wird. Man möchte gerne die Katheterlänge aus den Patientendaten schätzen.
In der Aufgabe 4 befassen wir uns mit dem Datensatz `spreng` aus der Vorlesung.

1. **a)** Untersuchen Sie die Verteilungen der 3 Variablen mit Hilfe von Boxplots und kommentieren Sie diese!

b) Betrachten Sie die zweidimensionalen Streudiagramme `y` gegen `x1`, `y` gegen `x2` und `x2` gegen `x1`. Was fällt Ihnen auf?
R-Hinweis: `pairs()`

c) Berechnen Sie die einfachen Regressionen von `y` auf `x1` und `y` auf `x2`. Geben Sie jeweils die Schätzungen für die Koeffizienten, $\widehat{\sigma}^2$ und R^2 an.

d) Testen Sie in beiden Modellen mit Hilfe des Regressions-Outputs die Hypothese $H_0 : \beta = 0$ gegen $H_A : \beta \neq 0$.

2. Passen sie das Modell $Y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + E_i$ an die Daten in `catheter.dat` an.

a) Gibt es einen gemeinsamen Einfluss von $x^{(1)}$ und $x^{(2)}$?

b) Testen Sie die Nullhypothese $H_0 : \beta_1 = 0$ resp. $\beta_2 = 0$ und vergleichen Sie mit Aufgabe 1.d).

c) Vergleichen Sie die Werte R^2 und $\widehat{\sigma}^2$ mit Aufgabe 1.c).

d) Führen Sie eine Residuen-Analyse durch.

3. Tabellieren Sie für das Modell in Aufgabe 2 die 95%-Vorhersage-Intervalle für alle Beobachtungen. In der Praxis würde man einen Vorhersagefehler von ± 2 cm akzeptieren. Lässt sich mit diesen Daten und diesem Modell die Katheter-Länge genügend genau vorhersagen?

Ist es sinnvoll, für die Bestimmung der Vorhersage-Intervalle alle Informationen (das volle Modell) auszunützen?

4. Der im Skript verwendete Datensatz der Sprengung steht unter `spreng.dat` zur Verfügung.

a) Passen Sie das Modell

$$\log 10(ersch)_i = \beta_0 + \beta_1 \log 10(dist)_i + \beta_2 \log 10(ladung)_i + \alpha_{Stelle_i} + E_i$$

mit der R-Funtkion `lm` an! Prüfen Sie mit Hilfe des `summary`, ob die Ladung einen signifikanten Einfluss hat und bestimmen Sie einen 95%-Vertrauensintervall für die Koeffizienten β_2.

b) Erstellen Sie aus dem `summary` eine zusätzliche Tabelle für `log10(dist), log10(ladung)` mit *t-Quotienten (signif)* und *stcoef*.

c) Lösen Sie die gleiche Aufgabe mit der Funktion `regr`.

d) Prüfen Sie, ob die Stelle einen signifikanten Einfluss hat, mit und ohne Benützung von `regr`. Ohne `regr` brauchen Sie dazu die Funktion `drop1`. (Sie sollten das gleiche Resultat erhalten.)

e) Prüfen Sie, ob eine Wechselwirkung zwischen der Stelle und der logarithmierten Distanz vorhanden ist, wieder mit und ohne `regr`. Welche Bedeutung hat diese Wechselwirkung?

f) Wenden Sie die Funktion `plot` auf die Ergebnisse von `lm` und von `regr` an. Ist die Annahme, gleicher Varianz für die Fehler E_i plausibel? Normalverteilung? Gibt es Hinweise auf Abweichung von der Regressionsfunktion?

g) Passen Sie das Modell auf die unlogarithmierten Zielgrösse `ersch` an. Welche Abweichungen sehen Sie in den Grafiken der Residuenanalyse?

R-Hinweis:
Um Graphiken und Outputs wie im Regressions-Skript zu erzeugen, brauchen Sie die von Werner Stahel geschriebenen und im Skript beschriebenen R-Funktionen `regr`, `g.res2x` und `plot.regr` und `print.regr`. Diese Funktionen finden Sie auf dem Internet unter:
`ftp://stat.ethz.ch/WBL/Source-WBL-2/R/regr.R`
Mit `source("ftp://stat.ethz.ch/WBL/Source-WBL-2/R/regr.R")` können Sie diese Funktionen laden. Natürlich können Sie sich die Funktionen auch auf Ihren Computer/Laptop runterladen, damit Sie später nicht immer Netzzugang brauchen, wenn Sie sie für Regressionsanalysen verwenden wollen.