Regression I — Serie 2

In dieser Serie wird der Datensatz `catheter` benützt. Es handelt sich um Daten aus der Medizin. Die Variable `x1` ist die Grösse (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 Katheter-Länge aus den Patienten-Daten schätzen.

1. Einfache Lineare Regression

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?
Nützlicher R-Befehl: `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. Multiple lineare Regression

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` 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).

3. Vorhersage mit linearen Modellen:

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?

Theorie: Siehe Skript: "Lineare Regression", Kapitel 2.4

R Hinweise: Die R-Funktion `predict(..., se.fit=T, ...)` liefert eine Liste mit 2 Vektoren (Vorhersage $\hat{\eta}$ an bestimmten Stützstellen und zugehörige Standardfehler $\mathrm{se}^{(\hat{\eta})}$) und zwei Zahlen (Anzahl Freiheitsgrade der Regression, geschätzte Standardabweichung des Fehlers).