Arbeitsblatt 4

Multiple lineare Regression (Einführung)

Aufgabe 1: Werbeausgaben

Der Datensatz `Advertisement.rda` enthält die Werbeausgaben (in 1000 US Dollar) für TV, Radio und Zeitung für 100 verschiedene Absatzmärkte. In `Sales` sind die Verkaufszahlen (pro 1000 Artikel) für jeden Absatzmarkt gegeben.

(a)

Passen Sie ein Modell mit Zielgrösse `sales` und den drei erklärenden Variablen `TV`, `radio` und `newspaper` an. Notieren Sie das Modell in mathematischer Notation.

(b)

Wie lauten die geschätzten Koeffizienten? Geben Sie eine Interpretation zum Achsenabschnitt und dem Koeffizient für die Variable `radio`?

(c)

Prüfen Sie mit einem geeigneten statistischen Test, ob mindestens eine der erklärenden Variablen einen auf 5% signifikanten Einfluss auf die Verkaufszahlen hat?

(d)

Zu welchem Prozentanteil lässt sich die Verkaufszahl mit diesem Modell erklären?

(e)

Berechnen Sie die 99% Vertrauensintervalle für den Anstiegs der Verkaufszahlen, wenn die Werbeausgaben für TV und Zeitung konstant bleiben und man die Werbeausgaben für radio um 1000 US$ erhöht.

(f)

Berechnen Sie eine Vorhersage für die Anzahl verkaufter Artikel, wenn die Werbeausgaben für TV bei 150000 US$, für Radio bei 40000 US $ und für Zeitungen bei 100000 US $ liegen. Geben Sie zusätzlich noch ein 95% Prognose-Intervall an.

(g)

Prüfen Sie mit einem geeigneten statistischen Test, ob die Werbeausgaben für die Zeitungen einen auf dem 5% Niveau signifikanten Einfluss auf die Verkaufszahl des Artikels hat?

(h)

Entfernen Sie die Variable `newspaper` aus dem Modell und visualisieren Sie die Situation mit einem 3D Plot. Verwenden Sie hierfür `scatter3d` aus dem R-Package `car`.

Aufgabe 2: Katheter

In dieser Aufgabe analysieren wir den Datensatz `catheter.dat`. Es handelt sich um Daten aus der Medizin. Die Variable `Groesse` ist die Grösse (in cm), `Gewicht` das Gewicht eines Patienten (in kg) und `y` die optimale Länge eines Katheters (in cm), der für die Untersuchung des Herzens eingesetzt wird. Man möchte gerne die Katheter-Länge aus den Patienten-Daten schätzen.

(a)

Untersuchen Sie den Datensatz mit Hilfe von Boxplots und zweidimensionalen Streudiagramme `y` gegen `Groesse`, `y` gegen `Gewicht` und `Gewicht` gegen `Groesse`. Was fällt Ihnen auf?

(b)

Berechnen Sie zwei einfache lineare Regressionen von `y` auf `Groesse` und `y` auf `Gewicht`. Geben Sie ausserdem jeweils die Schätzungen für die Koeffizienten und $\hat{\sigma}$ an.

(c)

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

(d)

Wir führen nun eine multiple lineare Regression durch, d.h. passen Sie das Modell

$$Y_i = \beta_0 + \beta_1 \texttt{Groesse}_i + \beta_2 \texttt{Gewicht}_i + E_i$$

an die Daten an. Kommentieren Sie den globalen F-Test und t-Test für die einzelnen Koeffizienten. Vergleichen Sie die Ergebnisse mit denjenigen der einfachen linearen Regression.

(e)

Vergleichen Sie die Schätzung für $\hat{\sigma}$ von der multiplen linearen Regression mit jenen von der einfachen linearen Regression.