Arbeitsblatt 6

Multiple lineare Regression (Interaktionen)

Aufgabe 1: Hausisolierung

Wir analysieren den Effekt einer Isolierungssanierung in einem Haus. Der Datensatz heisst `whiteside` im R-Package `MASS` (siehe Unterricht Woche 1). Der Datensatz enthält Daten über drei Variablen:

- Temp: Mittlere Aussentemperatur über eine Woche
- Gas: Wöchentlicher Gasverbrauch im Haus (in 1000 Feet^3)
- Insul (before/after): vor oder nach Einführung der Wärmedämmungsmassnahmen

(a)

Zuerst stellen wir die Daten in einem geeigneten Streudiagramm dar. Tragen Sie `Gas` auf der y-Achse und `Temp` auf der x-Achse auf. Zusätzlich färben Sie die Datenpunkte bezüglich der Variable `Insul` ein. Beschreiben Sie den Zusammenhang.

(b)

Führen Sie eine multiple Regression mit erklärenden Variablen `Insul` und `Temp` durch, um den Gasverbrauch zu modellieren. Zeichnen Sie das angepasste Modell in das Streudiagramm ein.

(c)

Prüfen Sie die Gültigkeit des Regressionsmodells mittels Residuenanalyse. Erzeugen Sie noch zwei zusätzliche Grafiken, einmal für die isolierten und einmal für die nicht isolierten Häuser, mit den Residuen auf der y-Achse und der `Temperatur` auf der x-Achse. Wie sollten die Punkte in dieser Zusatzgrafik verteilt sein, falls die Modellannahmen erfüllt sind? Ist dies der Fall?

(d)

Fügen Sie dem Modell zusätzlich eine Interaktion zwischen `Temp` und `Insul` hinzu. Zeichnen sie das Modell in ein Streudiagramm auf. Sind die Modellannahmen nun besser erfüllt?

(e)

Prüfen Sie die Signifikanz des Effekts einer Isolierungssanierung mit einem geeigneten Test auf dem 5% Niveau.

(f)

Wie hoch schätzen Sie, dass der Gasverbrauch in einem isolierten und einem nicht isoliertem Haus sein wird, wenn die Aussentemperatur über eine Woche um 0 Grad beträgt? Geben Sie einen Bereich an, in welchem der Gasverbrauch mit 95% Wahrscheinlichkeit enthalten sein wird?

Aufgabe 2: Berufsansehen

Der Datensatz `Prestige` stammt aus einer kanadischen Studie über das Ansehen von 98 Berufsgruppen in Kanada (Statistics Canada, 1971). Folgende Variablen stehen zur Verfügung:

`census`	ID für die Berufgruppe
`prestige`	mittleres Ansehen der Berufsgruppen
`education`	durchschittliche Ausbildungszeit (in Jahren)
`income`	mittleres Einkommen (in kanadischen Dollar)
`women`	mittlerer Anteil an Frauen (in Prozent)
`type`	Art der Beschäftigung: bc = Arbeiter; wc = Angestellte; prof = Selbstständige, Manager und Techniker

Die Daten stehen ihnen in `Prestige` im R-Package `car` zur Verfügung.

```
library(car, quietly = TRUE)
data(Prestige)
```

(a)

Passen Sie zuerst das folgende Modell an:

```
fit.prestige1 <- lm(prestige ~ income + education, data = Prestige)
```

und prüfen Sie mittels grafischer Residuenanalyse, ob die Modellvoraussetzungen erfüllt sind.

(b)

Bestimmen Sie alle Berufsgruppen mit grosser Hebelwirkung. Wenn Sie die Daten deskriptiv betrachten, können Sie sich erklären, warum diese Berufsgruppen eine grosse Hebelwirkung haben?

(c)

Passen Sie das Regressionsmodell mit und ohne die Punkte mit grosser Hebelwirkung an. Vergleichen Sie die Koeffizienten.

```
fit.prestige1b <- lm(prestige ~ income + education, data = Prestige[!s,])
coef(fit.prestige1)
```

```
 (Intercept)       income    education
-6.847778720  0.001361166  4.137444384
```

```
coef(fit.prestige1b)
```

```
(Intercept)      income   education
-7.70955035  0.00210334  3.81262803
```

(d)

Studieren Sie die Modelldefizite aus (a) anhand der partiellen Residuenplots genauer und zeichnen Sie die Residuen auch gegen die 2 (potentiellen) erklärenden Variablen, die noch nicht im Modell aufgenommen sind, auf.

Die partiellen Residuenplots lassen sich mit der Funktion `crPlots` aus dem R-Package `car` erstellen:

(e)

Passen Sie das folgende Regressionsmodell an:

```
Prestige$log_Income <- log(Prestige$income)
fit.prestige2 <- lm(prestige ~ log_Income + education + women
                    + type + log_Income:type + log_Income:women,
          data = Prestige)
```

und überprüfen Sie wider die Modelleignung mittels Residuenanalyse.

(f)

Wie ist die Interaktion zwischen `Income` und `type` zu interpretieren? Ist diese auf dem 5% Niveau signifikant?

(g)

Für die Berufsgruppe der Statistiker gibt es keine Messung des Ansehens. Wir nehmen folgendes an:

```
  log_Income women education type
1   10.59663    20        15 prof
```

Welches Ansehen geniessen die Statistiker? Geben Sie den Bereich an, in welchem das Ansehen mit 90% Wahrscheinlichkeit liegen wird. Ist das Ansehen höher als jenes der Mediziner (physicians)? Und wie ist das Ansehen im Vergleich zu jenem von Buchhaltern (accountants)?