Bernd Klaus (bernd.klaus@imise.uni-leipzig.de)
Verena Zuber (verena.zuber@imise.uni-leipzig.de)

http://uni-leipzig.de/~zuber/teaching/ws10/r-kurs/

1 Aufgabe: Lineare Regression mit Faktoren

Wir betrachten erneut den Datensatz "*toycars*", den man im Paket `DAAG` findet. Er beschreibt die Wegstrecke, die 3 verschiedene Spielzeugautos zurückgelegt haben, nachdem man sie in unterschiedlichen Winkeln eine Rampe herunterfahren ließ.

- *angle*: Winkel der Rampe
- *distance*: Zurückgelegte Strecke des Spielzeugautos
- *car*: Autotyp (1, 2 oder 3)

(a) Wenn noch nicht geschehen, installieren und laden Sie das Paket `DAAG`.

(b) Speichern Sie den Datensatz "*toycars*" in einem dataframe `data` ab und wandeln Sie die Variable "`car`" dieses Datensatzes in einen Faktor um.

(c) Erstellen Sie drei Boxplots, die die zurückgelegte Strecke getrennt nach dem Faktor "`car`" darstellen.

(d) Schätzen Sie die Parameter des folgenden linearen Modells mit Hilfe der Funktion "`lm()`"

$$\text{distance}_i = \beta_0 + \alpha_{i,\text{car2}} + \alpha_{i,\text{car3}} + \beta_1 \cdot \text{angle}_i + \varepsilon_i$$

Dabei bezeichnen $\alpha_{i,\text{car2}}$ und $\alpha_{i,\text{car3}}$ die Einflüsse der Autotypen 2 und 3 auf die Referenz β_0 für Autotyp 1. Interpretieren Sie die Regressionsparameter $\alpha_{i,\text{car2}}$ und $\alpha_{i,\text{car3}}$.

(e) Überprüfen Sie deskriptiv den Fit der Modelle, indem Sie die Koeffizienten des Modells mit dem Boxplot aus Aufgabe (c) vergleichen. Deutet das R^2 auf eine gute Anpassung des Modells hin?

(f) Führen Sie weitere deskriptive Diagnosen mit Hilfe der `plot()` Funktion durch.

2 Aufgabe: Interaktion von stetigen Variablen

Wir betrachten den Datensatz "*Suess.csv*". Er beschreibt die Auswirkung von Feuchtigkeit und Süße auf den Geschmack einer Süßigkeit.

- *Geschmack*: Geschmackspunktzahl (Integer)
- *Feuchtigkeit*: Feuchtigkeitspunktzahl (Integer)
- *Suesse*: Süßegrad (Integer)

(a) Laden Sie den Datensatz "*Sues.csv*" und speichern Sie ihn in einem dataframe `sues` ab.

(b) Benutzen Sie die Funktion `coplot( )` um einen Plot von *Geschmack* abhängig von der *Feuchtigkeit* der Süßigkeit, bedingt auf den Süßegrad zu erstellen. Benutzen Sie für eine bessere Darstellung die Optionen `pch = c(5,18)`, `rows = 1` und `columns = 3`. Gibt es bei gegebener Feuchtigkeit einen Einfluss der Süße auf den Geschmack?

(c) Fitten Sie das Modell

$$\text{Geschmack}_i = \beta_0 + \beta_1 \cdot \text{Feuchtigkeit}_i + \beta_2 \cdot \text{Suesse}_i + \varepsilon_i$$

(d) Plotten Sie die Residuen gegen Feuchtigkeit $*$ Suesse. Ist ein Zusammenhang erkennbar?

(e) Bestimmen sie nun die Parameter des Modells mit Interaktionsterm

$$\text{Geschmack}_i = \beta_0 + \beta_1 \cdot \text{Feuchtigkeit}_i + \beta_2 \cdot \text{Suesse}_i + \beta_3(\text{Feuchtigkeit}_i * \text{Suesse}_i) + \varepsilon_i$$

(f) Verbessert sich die Anpassung des Modells an die Daten? Ist der Koeffizient β_3 der Interaktion signifikant von 0 verschieden? Plotten Sie erneut die Residuen gegen Feuchtigkeit$*$Suesse und vergleichen Sie die Ergebnisse mit dem Plot aus Aufgabe (d).

3 Aufgabe: Diagnose

In dieser Aufgabe soll mittels simulierten Daten untersucht werden, wie eine Verletzung der Annahmen des linearen Modells in den Diagnoseplots zu erkennen ist.

- Erstellen Sie eine Hilfsvariable `h1` der Länge $n = 181$, die das Intervall von $[1, 10]$ in 0.05 Schritten abdeckt.
- Simulieren Sie die n Beobachtungen der erklärenden Variable X als `X=h1+` eine normalverteilte Zufallsgröße mit Erwartungswert 0 und Standardabweichung 1.

Berechnen Sie für die folgenden drei Szenarien das lineare Modell und untersuchen Sie die Annahmen mit den geeigneten Diagnoseplots. Versuchen Sie die Verletzung der Annahmen in den Diagnoseplots zu erkennen.

(a) Simulieren Sie einen normalverteilten Fehler `epsilon1` der Länge $n = 181$ mit Erwartungswert 0 und Standardabweichung 1 und konstruieren Sie die Zielgröße `Y1` als

$$Y1 = log(X) + epsilon1$$

(b) Simulieren Sie einen Cauchy-verteilten Fehler `epsilon2` der Länge $n = 181$ mit dem Befehl `rcauchy(n, location=0, scale=1)` und konstruieren Sie die Zielgröße `Y2` als

$$Y2 = X + epsilon2$$

Plotten Sie die Kerndichteschätzer für `epsilon2` und `epsilon1` in einer Graphik. Wie unterscheiden sich Cauchy und Normalverteilung?

(c) Simulieren Sie einen normalverteilten Fehler `epsilon3` der Länge $n = 181$ mit Erwartungswert 0 und einer Standardabweichung, die ein Zehntel des entsprechendes X-Wertes ist (Hinweis: Der Funktion `rnorm` können bei der Option `mean` und `sd` Vektoren identischer Länge n übergeben werden. Dann werden auch n normalverteilte Zufallszahlen mit dem in `mean` und `sd` spezifizierten Parametern erzeugt). Konstruieren Sie die Zielgröße `Y3` als

$$Y3 = X + epsilon3$$