

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

<http://uni-leipzig.de/~zuber/teaching/ws09/r-kurs/>

## 1 Aufgabe: Einfaktorielle ANOVA

Wir betrachten den Datensatz “*HotDog.csv*”. Er beschreibt den Salz- bzw. Kaloriengehalt von Hot Dogs verschiedener Marken in Abhängigkeit von der im Hotdog enthaltenen Fleischart.

- *Type*: Fleischart (“Beef”, “Meat”, “Poultry” )
  - *Calories*: Kalorien des Hot Dogs
  - *Sodium*: Salzgehalt des Hot Dogs
- (a) Lesen Sie den Datensatz mit Hilfe des `read.table()`-Befehls ein.
- (b) Untersuchen Sie mit Hilfe von zwei Varianzanalysen den Einfluss der Fleischart sowohl auf die Kalorien als auch auf den Salzgehalt.
- (c) Welches der beiden Anova-Modelle liefert einen signifikanten *F*-Wert?

## 2 Aufgabe: Mehrfaktorielle ANOVA

Wir betrachten den Datensatz “*Noise.csv*”. Er beschreibt die Auswirkung von drei Faktoren auf das Fahrgeräusch verschiedener Automodelle.

- *NOISE*: Geräuschpegel
  - *SIZE*: Größe des Autos: klein (1), mittel (2), groß (3)
  - *TYPE*: Art des Abgasfilters: (1) oder (2)
  - *SIDE*: Lage des Abgasfilters: links (1) oder rechts (2)
- (a) Lesen Sie den Datensatz mit Hilfe des `read.table()`-Befehls ein.
- (b) Erstellen Sie ein Anova-Modell mit 3 Faktoren. Welcher der 3 Haupteffekte scheint isoliert betrachtet der Wichtigste zu sein? Welcher der 3 Haupteffekte trägt am wenigsten zur Erklärung bei?
- (c) Erstellen Sie zwei weitere Anova-Modelle ohne den unwichtigsten Faktor, einmal mit und einmal ohne Interaktionseffekt. Ist die Interaktion signifikant?
- (d) Sollte das Modell aus der letzten Teilaufgabe noch weiter reduziert werden?

### 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

$$Y1 = \log(X) + \text{epsilon1}$$

- (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

$$Y2 = X + \text{epsilon2}$$

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 entsprechenden  $X$ -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

$$Y3 = X + \text{epsilon3}$$