

Generalisierte Lineare Modelle – Übungen: Blatt 1

1. Lese die Daten `bacteria.dat` in R ein. Dabei handelt es sich um die Ergebnisse aus einem einjährigen Monitoring (beginnend am 8. März 1995) von Bakterien in der Luft gemessen an 7 verschiedenen Plätzen im Großraum Graz, und zwar von Zwaring im Süden (`site = 1`) bis Mariatrost (`site = 7`) im Norden. Diese Messorte sind charakterisiert durch
 1. village zone, near big farms with liquid manure pits and dung-hills;
 2. grassland and arable land, without buildings;
 3. suburban area with one-family houses and small farms;
 4. busy crossing, near a slaughter-house;
 5. public park on top of the Schloßberg in the center of Graz;
 6. living area with apartment buildings and gardens;
 7. as for 6 but with compost arrangements.

Verwende im Folgenden die Information `site` ausschließlich als Faktor mit sieben Stufen.

Alle 14 Tage wurden an denselben sieben Stellen die Konzentrationen erhoben, sowie auch die zum Messzeitpunkt vorherrschende Temperatur und Luftfeuchtigkeit. Das dabei verwendete Messgerät war ein sechsstufiger Sammler der Firma Andersen. Die Variablen `b1`, \dots , `b6` geben jeweils die Anzahl der kolonienbildender Einheiten (cfu's) auf den Stufen j an, die in einer Menge von 128.3 Liter Luft vorhanden waren. Nähere Details zu den Daten findet man in der Beschreibung `bacteria.def`.

Generiere die Variable `bac`, welche die totale Bakterien-Konzentration in 1m^3 Luft beschreibt.

2. Lege nun Augenmerk auf die Response `bac` und analysiere die Abhängigkeit von der Luftfeuchte `humi`, der Temperatur `temp`, und dem Ort `site`. Das Datum der Messung sollte hinreichend durch die Temperatur und Luftfeuchtigkeit an diesem Tag beschrieben sein.

Finde das optimale multiple lineare Regressionsmodell für die Variable `bac`. Sind hierzu auch Interaktionen notwendig und was würden diese als Informationen ins Modell einbringen. Untersuche dabei auch den Grad der Erklärung der Variablen `temp**2` und `humi**2`. Ist der Einfluss der Wetter-Variablen auf die Response-Variable in allen Orten derselbe?
3. Prüfe das gefundene Modell auf Abweichungen von der Annahme einer *konstanten Varianz* mittels geeigneter Plots.
4. Führe eine optimale Box-Cox Transformation durch und teste auf die prinzipielle Notwendigkeit dieser Transformation sowie speziell auf die Adäquatheit der log-Transformation.
5. Vergleiche die Güte der Modellanpassung des linearen Modell mit dem des Box-Cox-Modell, wobei beide Modelle dieselben erklärenden Variablen beinhalten.
6. Ist die Struktur in den Residuen des Box-Cox-Modells jetzt besser (unauffälliger)?
7. Sei $Z \sim \text{Normal}(\mu, \sigma^2)$. Leite die Dichtefunktion sowie den Erwartungswert und die Varianz der Variablen $Y = \exp(Z)$ her (Momente der Log-normal Verteilung).