Generalised Linear Models – 1st Homework Assignment

1. Download the data on bacteria counts in the air in and around Graz. You find the data either in `http://www.stat.tugraz.at/courses/files/BacteriaData.xlsx` (sheet name `bacteria`) or by clicking on `Bacteria Data` at `http://www.stat.tugraz.at/courses/glmLjubljana.html` (only for those of you experiencing troubles with the use of `read.xls`).

2. The data resulted from a one year study in which bacteria (colonies forming units, cfu's) in the outdoor air were monitored at 7 different sites characterized as follows:

   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.

   Use the information `site` as a factor with 7 labels.

   Every 2 weeks the concentration of airborne bacteria (and fungi) was observed. Also observed was the temperature (`temp`) and the humidity (`humi`) at this time. The gauge (measurement equipment) was a six stages microbial air sampler (Andersen). The variables `b1`, ..., `b6` describe cfu counts observed on every stage $j = 1, \ldots, 6$ of the gauge from 128.3 liter air.

   Define the variable `bac` as the total number of cfu's (sum of `b1`, ..., `b6`) in 1m^3 air.

3. Concentrate on the response variable `bac` and analyze its linear relationship with `humi`, `temp`, and `site`. Don't consider `date` because this information should be sufficiently described by temperature and humidity of the same day.

   Find the best linear regression model for the response variable `bac`. Also check for a necessary interaction between temperature and humidity. Don't forget to additionally check the relevances of the quadratic effects `temp^2` and `humi^2` in your model. Such effects will help to account for some optimal temperature and/or optimal humidity which bacteria like most.

4. Assess the resulting linear regression model with respect to departures from the assumption of *constant variance (homoscedasticity)* by means of suitable plots.

5. Search for the optimal Box-Cox transformation and test on the general necessity of such a transformation ($\text{H}_0 : \lambda = 1$) as also on the adequacy of a log-transformation ($\text{H}_0 : \lambda = 0$).

6. Compare the goodness-of-fit of the linear regression model with that of the Box-Cox-model, where both these models contain the same set of predictors.

7. Has the structure in the residual plot from the Box-Cox-model now improved (compared with that from the multiple linear regression model from before)?