Technische Universität München
Zentrum Mathematik
Lehrstuhl für Mathematische Statistik
Prof. Claudia Czado, Ph.D. (cczado@ma.tum.de)
Ariane Hanebeck (ariane.hanebeck@tum.de)

WS 2022/2023

Exercise Sheet 10
Generalized Linear Models

Discussion of the tutorial exercises on January 16 and 19, 2022

Preparations Download the dataset `insurance.dat` from Moodle.

Problem 1 The data set `insurance.dat` consists of motor insurance claims in Sweden from 1977. We use the *average claim size* as the response variable and we investigate the effect of the four covariates `Kilometres`, `Zone`, `Bonus` and `Make` on the response variable.

a) Load the data set, call it `ins.dat` and remove all observations with zero claims. Define the response variable Y_i^s as the *average claim size*, i.e., $Y_i^s := \frac{Payment_i}{n_i}$, where n_i is the number of claims for observation i.

b) Write down (in mathematical form, not in `R`) the assumed relationship between the response variable *average claim size* and the four covariates in the Gamma regression model if the log link function is used, and the one if the inverse link function is used.

c) Perform an exploratory data analysis to investigate the main effects using the function `cat_plot` and assuming the log link function. Merge categories with similar empirical log means.

d) Fit a gamma regression model with the main effects using the log link function. Note that the categorical covariates should be factorized (in `R`: `as.factor`) and metric covariates not. Make sure that you use **weights** in the Gamma regression model. What is the estimated value of the dispersion parameter?

e) Perform a residual deviance test to assess the model fit at $\alpha = 0.05$.

Problem 2 (Additional, Problem 1, Sheet 9 continued)

a) Perform an exploratory data analysis to investigate the interaction effects of the four covariates using the function `cat_plot`.

b) Fit a model `model.inter` with all pairwise interaction terms.

c) Select a model `model.inter2` by performing the stepwise AIC approach as follows: start with the model `model.main` from Sheet 9 and add interaction effects until the AIC is not reduced anymore. You may use the `R` function `step`. Compare the two interaction models `model.inter` and `model.inter2`. Furthermore, perform partial deviance tests for all nested models of `model.inter2` and interpret the result.

d) Perform a partial deviance test at $\alpha = 0.05$ for the models `model.main` and `model.inter`. Which one would you prefer? Further, use the residual deviance test to check if the model assumptions of the preferred model are correct.

e) Compute and plot Pearson and deviance residuals of `model.inter`. Interpret your plots.

f) Using the `R` function `persp`, draw a 3-dimensional plot for expected number of claims per year versus `Bonus` and `Kilometers` in case `Make=2` and the cases `Zone=1,2,3,4`. Interpret your 4 plots.

g) Is overdispersion present in `model.inter`?