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 7
Generalized Linear Models

Discussion of the tutorial exercises on December 5 and 8, 2022

Preparations Download the data set `credit.dat` from Moodle. The data set describes 1000 consumer loans of a German bank. The response variable of interest is `kredit` and it is given in dichotomous form: 1 indicates that the loan was repayed properly, 0 indicates the opposite. We consider the covariates `laufzeit`, `moral`, `laufkont`, `alter`, and `beruf`. A detailed description of the data set can be found at `https://data.ub.uni-muenchen.de/23/1/DETAILS.html`.

Problem 1 (*)

a) We are working with the aggregated data, i.e., aggregate the data first to obtain binomial data instead of binary data. For this, you have to define a new variable `no_kredit=1-kredit`. Then, characterize the scale (quantitative, ordinal, nominal, dichotomous) of each of the 5 selected variables of the data set. Transform all nominal and ordinal covariates into `factor` variables in `R` using the function `as.factor`.

b) Perform an explorative data analysis (EDA). For this, do the following two steps.

   i) How many debtors repay their loan? Use the function `table` on the original data.

   ii) Determine the influence of each covariate on the response. Use the function `cat_plot` for categorical covariates and the function `gam` for quantitative covariates. An example is given in the example from the lecture. Do you need to transform quantitative covariates or merge categories of categorical covariates?

c) Fit a logistic regression model to the data using the transformations or merging you decided on in b).

   `glm(cbind(kredit,no_kredit) ~ ... , family=binomial(link = "logit"))`

   Using the step function with `direction="forward"`, decide if some covariates are not important. We call the resulting model `my.glm.step.main`.

d) Perform partial deviance tests for all the nested models using the `R`-function

   `anova(my.glm.step.main, test="Chisq")` and explain the outcome with $\alpha = 0.05$. What is the null hypothesis and the rejection rule for this test? Decide if you want to remove another covariate from the model.

e) Perform a residual deviance test at the $\alpha = 0.05$ level (without using a pre-defined `R`-function). Explain the outcome. What is the null hypothesis and the rejection rule for the residual deviance test?

f) Investigate the pairwise interaction effects of the covariates. For this, make interaction plots for all possible interactions. Using the `R` function step, sequentially add interaction effects by

model.inter = step(my.glm.step.main, .^ 2, direction="forward).

Compare with your interaction plots if the selection makes sense.

For interaction plots between a quantitative and a categorical covariate, use `interact_plot`. For interaction plots between two categorical covariates, use `cat_plot`. To obtain an interaction plot between two quantitative covariates, first group them using `cut(alter, breaks = quantile(alter, probs =c(0, 0.2, 0.4, 0.6, 0.8, 1), na.rm = T))`.

Problem 2 (Additional) Consider the complementary log-log regression model

$$\begin{aligned} Y_i|\boldsymbol{X}_i = \boldsymbol{x}_i &\sim Bernoulli(p(\boldsymbol{x}_i)) \qquad \text{independent} \\ p(\boldsymbol{x}_i) &= 1 - \exp\left(-\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})\right), \quad i = 1, \ldots, n, \end{aligned}$$

where $\boldsymbol{x}_i^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_i$.

a) Determine the log-likelihood function ℓ of this model.

b) Determine the scaled score equation $s(\boldsymbol{\beta}) = s(\beta_0, \beta_1) = (\frac{\partial \ell}{\partial \beta_0} \frac{\partial \ell}{\partial \beta_1})^T$.

c) Determine the scaled Hessian matrix $\boldsymbol{H}(\boldsymbol{\beta})$ of the log-likelihood,

$$\boldsymbol{H}(\boldsymbol{\beta}) = \begin{pmatrix} \frac{\partial^2 \ell}{\partial^2 \beta_0} & \frac{\partial^2 \ell}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 \ell}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 \ell}{\partial^2 \beta_1} \end{pmatrix}$$

d) Determine the deviance of this model.