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 8
Generalized Linear Models

Discussion of the tutorial exercises on December 12, 2022

There will be no exercise session on December 15.

Problem 1 (*) We consider the model `small.model` defined as follows

`kredit` $\sim$ `moral + laufkont + laufzeit`

Group the data according to the three variables `moral`, `laufkont`, and `laufzeit` using the function `aggregate`. Transform ordinal and nominal variables to factor variables.

1. How many observations does the aggregated data set have?
2. Fit a binomial regression model `small.model.agg` to the aggregated data: assume each observation Y_i is binomially distributed $Y_i \sim Binom(n_i, p_i)$
3. Do the following regression diagnostics for the model `small.model.agg`.
   (a) Compute the leverage h_{ii}^L for each observation using the `R` function `hatvalues()`. Plot the leverage and identify all points with high leverage.
   (b) Compute
      i. the Pearson residuals
      ii. the deviance residuals
      iii. the adjusted residuals

      and plot them. Use the function `resid` and specify the option `type` for (i) and (ii). For (iii), use the definition from the lecture. Interpret the results.
   (c) Compute the Cook's distance using the definition from the lecture and plot the results. Are there any influential observations?

Problem 2 (Additional) Consider a logistic regression model for the response variable `kredit` of `credit.dat` with the four main effects `moral`, `laufzeit`, `alter` and `laufkont`.

Define the partial residuals for `laufzeit`. Generate a partial residual plot for `laufzeit` and interpret it.

Problem 3 (Additional) We consider again the data set `credit.dat` using the covariates `laufzeit`, `moral`, `laufkont`, and `alter`. To predict the category (`kredit=1` or `kredit=0`), use a threshold function

$$\widehat{Y}_i = \begin{cases} 1 & \widehat{p}_i \geq \tau \\ 0 & \widehat{p}_i < \tau, \end{cases}$$

where $\tau \in [0, 1]$ is the threshold. A common choice is $\tau = 0.5$.

a) Load the data and transform the categorical covariates into `factor` data.

b) Set the seed of R's random number generator to obtain reproducible results:

```
set.seed(234)
```

c) Partition the data set into a *training* data set of size 700 and a *test* data set of size 300 by defining a vector `train` that contains the indices of the training set:

```
train = sample(1000, 700)
```

Important: Use only the training data to estimate your models in the rest of this exercise. Use both the training and testing data to evaluate your estimated model.

d) Define the following GLM (`model1`) using the training data:

```
model1 = glm(kredit ~ laufzeit + moral + laufkont + alter,
             data = credit, family = binomial(link = "logit"),
             subset = train)
```

e) Evaluate the estimated probabilities $\widehat{p}_i = \widehat{P}(\text{kredit}_i = 1)$ with the whole data set using the function `predict`. Then, predict the response variable $\widehat{\mathbf{Y}}$ of 1s and 0s with the threshold $\tau = 0.5$. What percentage of the training examples is predicted incorrectly (this is called the *training error*)? What percentage of the test examples is predicted incorrectly (this is called the *test error*)?