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 3
Generalized Linear Models

Discussion of the tutorial exercises on November 7 and 10, 2022

Problem 1 (*) We consider the USCRIME data again. Perform the following regression diagnostics for the linear model: `my.model <- lm(R ~ Ex0 + X + Ed + Age + poly(NW,3) + U2 + poly(LF, 2) + N)`

a) Plot the raw residuals `my.model$residuals` versus `X` and `Ed` separately. Do these two covariates have a nonlinear influence on the response?

b) Plot the internally studentized residuals versus the observation number. Add appropriate bands in which the vast majority of the studentized residuals (≈95%, 99.7%) should lie. Are there any unusual observations? If so, which ones are they, and what do they indicate about the model?
Hint: use `rstandard` to obtain the internally studentized residuals.

c) Plot the internally studentized residuals versus the fitted values `my.model$fitted.values`. Interpret the plot.

d) Draw a QQ-plot of the internally studentized residuals. Add the line that represents the theoretical relationship, assuming that the model assumptions are satisfied. Interpret the result.
Hint: use `qqnorm` or `qplot`.

e) Compute the hat matrix of `my.model`. Determine points with high leverage and list their corresponding observation index.

f) Draw bivariate scatter plots of two predictor variables and indicate the x-outlier(s) from part e) with a different color. Interpret the plots.
Use the pairs: (`Age`, `Education`), (`Age`, `X`), (`Ex0`, `X`).

g) Compute Cook's distance for each observation using `cooks.distance()`. Determine the observations with a high value and list them. Why do these points differ slightly from the leverage points?

Problem 2 (Additional) Consider a linear model $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with uncorrelated errors $\boldsymbol{\epsilon}$. Assume that the error variances are inhomogeneous and define $Cov(\boldsymbol{\epsilon}) := \boldsymbol{V} = diag(\sigma_1^2, \ldots, \sigma_n^2)$ and $w_i := \sigma_i^{-2}$.
An estimator $\boldsymbol{\beta}^*$ is called a weighted LS estimator, if $\boldsymbol{\beta}^*$ minimizes

$$SS_{Res,\boldsymbol{V}} := (\boldsymbol{Z} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{V}^{-1} (\boldsymbol{Z} - \boldsymbol{X}\boldsymbol{\beta}).$$

Provide the normal equations for the weighted LS estimation in vector form, write down the formula for $\boldsymbol{\beta}^*$ and show that with $\eta_i^* = \sum_{s=1}^p x_{is}\beta_s^*$ the normal equations can be written as

$$\sum_{i=1}^n w_i x_{ij} \eta_i^* = \sum_{i=1}^n w_i x_{ij} z_i, \qquad j = 1, \ldots, p.$$