Prof. Dr. M. Kateri, Dr. W. Herff RWTH Aachen, SoSe 2022

Publication: June 29th — Presentation of solutions: July 7th

Applied Data Analysis

Exercise Sheet 6

Exercise 21

Consider a GLM with stochastically independent, gamma distributed response random variables $Y_1, \ldots, Y_n$ and log-link function, i.e the natural logarithm function ln. This Exercise aims to show that the GLM described above leads to similar results as applying a normal linear model to $\ln(Y_1), \ldots, \ln(Y_n)$.

For this purpose, in this Exercise we use the class of gamma distributions with the following parametrization. For parameters $\mu > 0$, $\beta > 0$, let the corresponding density function $f(\bullet; \mu, \beta) : \mathbb{R} \longrightarrow [0, \infty)$ be given by

$$(*) \qquad f(y; \mu, \beta) \;:=\; \begin{cases} \dfrac{1}{\Gamma(\beta)} \left(\dfrac{\beta}{\mu}\right)^{\beta} y^{\beta-1} \exp\left(-\dfrac{\beta}{\mu} y\right) & , \quad y > 0 \;, \\ 0 & , \quad y \leq 0 \;. \end{cases}$$

(a) Show that the class of gamma distributions with parametrization given by $(*)$ is a member of the exponential dispersion family of distributions. Further, determine the corresponding canonical link function.

(b) Describe a practical situation (in the sense of properties of a data set), where modeling data with gamma distributions could be appropriate.

(c) Show: If for $i \in \{1, \ldots, n\}$, Y_i has a standard deviation $\sigma_i = \sqrt{\mathrm{Var}(Y_i)}$ that is proportional to its expectation $\mu_i = \mathrm{E}(Y_i)$, then $X_i := \ln(Y_i)$ has approximately a constant variance for small σ_i.

Hint: For $i \in \{1, \ldots, n\}$, use a Taylor approximation of $X_i := \ln(Y_i)$ around $\mu_i = \mathrm{E}(Y_i)$.

(d) The Gamma-GLM described above with log-link function ln refers to $\ln(\mathrm{E}(Y_i))$, whereas the ordinary linear model for the transformed response variable $X_i := \ln(Y_i)$ refers to $\mathrm{E}(X_i) = \mathrm{E}(\ln(Y_i))$ for $i \in \{1, \ldots, n\}$.

Show that if $X_i := \ln(Y_i) \sim \mathcal{N}(\mu_i, \sigma^2)$ for $i \in \{1, \ldots, n\}$ with $\mu_1, \ldots, \mu_n \in \mathbb{R}$ and $\sigma > 0$, then it holds:

$$\ln\big(\mathrm{E}(Y_i)\big) \;=\; \mathrm{E}\big(\ln(Y_i)\big) + \frac{\sigma^2}{2} \;.$$

Exercise 22

(a) Consider a Baseline-Category logit model according to Definition II.4.6 with 3 possible outcome categories and a single explanatory variable $x \in \mathbb{R}$. Then, according to II.4.6, the corresponding response probabilities fulfill the following equations:

$$\pi_j(x) \;=\; \frac{\exp(\beta_{j,1} + \beta_{j,2}\, x)}{1 + \exp(\beta_{1,1} + \beta_{1,2}\, x) + \exp(\beta_{2,1} + \beta_{2,2}\, x)} \;, \quad j \in \{1, 2\} \;,$$

and $\pi_3(x) \;=\; 1 - (\pi_1(x) + \pi_2(x))$ for $x \in \mathbb{R}$.

Assume that $\beta_{1,2} \neq 0$, $\beta_{2,2} \neq 0$ and show:

$$\pi_3 \text{ is } \begin{cases} \text{stictly decreasing}, & \text{if } \beta_{1,2} > 0 \text{ and } \beta_{2,2} > 0 \,, \\ \text{stictly increasing}, & \text{if } \beta_{1,2} < 0 \text{ and } \beta_{2,2} < 0 \,, \\ \text{not monotonic}, & \text{else} \,. \end{cases}$$

(b) Now, consider a Baseline-Category logit model for grouped data with $g \in \mathbb{N}$ observed groups and $J \geq 2$ outcome categories. Let n_i denote the number of responses in group $i \in \{1, \ldots, g\}$. Further, for $i \in \{1, \ldots, g\}$ and $j \in \{1, \ldots, J\}$, let y_{ij} denote the corresponding sample proportion of category j in group i (corresponding to the saturated model) and let $\hat{\pi}_{ij}$ denote the estimate of the probability π_{ij} of responding category j belonging to group i corresponding to the considered model.

Show that the deviance given by

$$G^2 \;=\; 2 \sum_{i=1}^{g} \sum_{j=1}^{J} n_i \, y_{ij} \, \ln\left(\frac{n_i \, y_{ij}}{n_i \, \hat{\pi}_{ij}} \right)$$

coincides with the corresponding likelihood-ratio test statistic for testing the considered Baseline-Category logit model against the saturated model.

Exercise 23

For each task of this Exercise, let P denote the probability measure of the underlying probability space.

(a) Suppose $Y_1, \ldots, Y_n$ are stochastically independent counts that satisfy a Poisson GLM with parameters $\mu_1, \ldots, \mu_n > 0$ and log-link function, i.e.:

$$\ln\big(\mathrm{E}(Y_i)\big) \;=\; \ln(\mu_i) \;=\; \sum_{j=1}^{p} x_{ij} \, \beta_j \;\;, \;\; i \in \{1, \ldots, n\}$$

with explanatory variables $x_{1,1}, \ldots, x_{n,p} \in \mathbb{R}$ and model parameters $\beta_1, \ldots, \beta_p \in \mathbb{R}$.

(i) In this part, for $i \in \{1, \ldots, n\}$, assume that the corresponding observed response merely indicates whether Y_i is positive or not, modeled by the indicator random variable $Z_i := \mathbb{1}_{(0,\infty)}(Y_i)$.
Show that $Z_1, \ldots, Z_n$ satisfy a binary GLM with complementary-log-log link function, i.e.:

$$\ln\Big(-\ln\big(1 - P(Z_i = 1)\big)\Big) \;=\; \sum_{j=1}^{p} x_{ij} \, \beta_j \;\;, \;\; i \in \{1, \ldots, n\} \,.$$

(ii) In this part, assume that $x_{i1} = 1$ for $i \in \{1, \ldots, n\}$, and thus, that the model contains an intercept. For $j \in \{1, \ldots, p\}$, let $\widehat{\beta}_j$ denote the maximum likelihood estimate of β_j and for $i \in \{1, \ldots, n\}$, let $\widehat{\mu}_i$ denote the corresponding estimate of μ_i.

Show that the average rates of change in the estimated means satisfy the following equations:

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\partial \widehat{\mu}_i}{\partial x_{ij}}(x_{ij}) \;=\; \widehat{\beta}_j \, \overline{y} \;\;, \;\; j \in \{1, \ldots, p\} \,.$$

(b) Let X be a positive random variable with $\mathrm{E}(X) = 1$ and $\mathrm{Var}(X) = \tau$ for some $\tau \in (0, \infty)$. (Thus, especially, we asumme the existence of the second moment of X.) Further, let $\lambda > 0$ and let Y be a random variable with values in N_0 fulfilling

$$P^{Y|X=x} \;=\; \mathcal{P}(\lambda\, x) \;\;,\;\; x \in (0, \infty)\;,$$

i.e. for $x \in (0, \infty)$, the conditional distribution of Y under the condition $X = x$ is a Poisson distribution with parameter $\lambda\, x$.

Show

(i) $\mathrm{E}(Y) = \lambda$,

(ii) $\mathrm{Var}(Y) = \lambda + \tau\, \lambda^2$.

Hint: Use the following property for dealing with conditional expectations:

$$\mathrm{Var}(Y) \;=\; \mathrm{E}\big(\mathrm{Var}(Y|X)\big) \;+\; \mathrm{Var}\big(\mathrm{E}(Y|X)\big)$$

where $E(Y|X)$ denotes the conditional expectation of Y given X and $\mathrm{Var}(Y|X) \;=\; \mathrm{E}(Y^2|X) - \big(\mathrm{E}(Y|X)\big)^2$ denotes the conditional variance of Y given X.

(c) Now, in the situation of part (b), especially, let $\lambda := 1$ and let X have a gamma distribution with parameters $\mu > 0$ and $\beta > 0$ according to the parametrization given by $(*)$ in Exercise 21. Thus, with $\lambda = 1$, we assume:

$$P^{Y|X=x} \;=\; \mathcal{P}(x) \;\;,\;\; x \in (0, \infty)\;.$$

Show that the marginal distribution of Y is a negative binomial distribution with parameters $p = \frac{\beta}{\mu+\beta}$ and β, i.e. the corresponding probability mass function of Y is given by

$$P(Y = y) \;=\; \int_0^\infty P(Y = y \,|\, X = x)\, f(y; \mu, \beta)\, dx \;=\; \frac{\Gamma(y+\beta)}{\Gamma(\beta)\, y!}\, p^\beta\, (1-p)^y \;\;,\;\; y \in \mathbb{N}_0\;.$$