Prof. Dr. E. Cramer, Prof. Dr. M. Kateri,
L. Kaufmann, M.Sc., T. van Bentum, M.Sc.,
A. Müllenmester

RWTH Aachen, SS 2021
Release date: June 1st, 2021
Solution: June 8th, 2021

Applied Data Analysis

R-Laboratory 6

Exponential Dispersion Family – Response Transformation

Task 19

(a) Write two R functions which allow the definition of a probability mass function (pmf) or probability density function (pdf) from the exponential dispersion family and from the k-parametric natural exponential family for a (univariate) random variable Y.

(i) The arguments of the first function should be the functions a, b, c, and the dispersion parameter ϕ (see Definition II.2.3 in the lecture). The function should return another function with the arguments y and ϑ whose return value is the value of the pmf/pdf of Y at y for the natural parameter ϑ.

(ii) The arguments of the second function should be the functions T, h and B (see Definition II.2.7 with $\eta_j(\boldsymbol{\vartheta}) := \vartheta_j$, $j = 1, \ldots, k$ for $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_k) \in \Theta \subset \mathbb{R}^k$). The function should return another function with the arguments y and $\boldsymbol{\vartheta}$ whose return value is the value of the pmf/pdf of Y at y for the natural parameter vector $\boldsymbol{\vartheta}$.

Remark: Remember that the symbols `c` and `T` are already used in R. If you define those two functions, give them a name different from `c` and `T`.

(b) A pmf of a random variable with values in $\mathbb{N}_0$ is given by a member of the exponential dispersion family with

$$\begin{aligned} a&\colon \mathbb{R}_+ \to \mathbb{R}_+, && \phi \mapsto \phi, \\ b&\colon (-\infty, 0) \to \mathbb{R}, && \vartheta \mapsto -\ln(1 - \exp(\vartheta)), \\ c&\colon \mathbb{N}_0 \times \mathbb{R}_+ \to \mathbb{R}, && (y, \phi) \mapsto 0 \end{aligned}$$

and $\phi = 1$. With these functions and $\vartheta = -0.8$ plot the resulting probability mass function using your own function from (a) and compare it with known ones. What do you observe? Furthermore, draw a random sample of size $n = 200$ from this known distribution using the correct parameters and compare the sample with the probability mass function using a barplot.

(c) A pdf of a random variable with values in $\mathbb{R}_+$ is given by a member of the exponential

dispersion family with

$$\begin{aligned} a\colon \mathbb{R}_+ &\to \mathbb{R}_+, & \phi &\mapsto \phi, \\ b\colon (-\infty, 0) &\to \mathbb{R}, & \vartheta &\mapsto -\ln\left(\frac{(-\vartheta)^3}{2}\right), \\ c\colon \mathbb{R}_+^2 &\to \mathbb{R}, & (y, \phi) &\mapsto \ln\left(y^2\right) \end{aligned}$$

and $\phi = 1$. With these functions and $\vartheta = -2$ plot the resulting density using your own function from (a) and compare it with known density functions. What do you observe? Furthermore, draw a random sample of size $n = 200$ from this known distribution using the correct parameters and compare the sample with the density function using a histogram.

(d) A pdf of a random variable with values in $\mathbb{R}_+$ is given by a member of the 2-parametric natural exponential family with

$$\begin{aligned} T\colon \mathbb{R}_+ &\to \mathbb{R}^2, & y &\mapsto (\log(y), y) =: (T_1(y), T_2(y)), \\ B\colon (-1, \infty) \times (-\infty, 0) &\to \mathbb{R}, & \boldsymbol{\vartheta} &\mapsto \ln(\Gamma(\vartheta_1 + 1)) - (\vartheta_1 + 1)\ln(-\vartheta_2), \\ h\colon \mathbb{R}_+ &\to \mathbb{R}, & y &\mapsto 1, \end{aligned}$$

where $\Gamma(\,\cdot\,)$ denotes the gamma-function. Find parameters ϑ_1 and ϑ_2 such that this pdf coincides with the pdf from (c). Compare the pdfs graphically using the functions from (a).

Task 20

(a) Download the data set *Sim1.csv* from the RWTHmoodle space and load it as data frame into your workspace.

(b) Fit the linear model $\mathsf{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$, where the design matrix $\boldsymbol{X}$ contains an intercept and two additonal columns of the variables `x1` and `x2` from the *Sim1.csv* data set. The observed values of $\boldsymbol{Y}$ are stored in the variable `y` in the *Sim1.csv* data set. Check the fit of the linear model using the discussed techniques of the previous R-Labs and the lecture. Is this an appropriate model?

(c) As alternative approach, fit a linear model with intercept for the transformed response variable `log(y)` and the explanatory variables `x1` and `x2`. Check the fit of this linear model. If this model is appropriate, test on the significance level $\alpha = 0.05$ if the variables `x1` and `x2` have a significant influence on the expected value of the response variable. Fit a final linear model only containing the important explanatory variables.

(d) Try to use the final linear model of (c) to estimate $\mathsf{E}(\boldsymbol{Y})$. For comparison consider the actual values $\mathsf{E}(Y_1) = \exp(1)$ and $\mathsf{E}(Y_4) = \exp(1.5)$. What do you observe?