Prof. Dr. M. Kateri,
Dr. W. Herff,
L. Kaufmann, M.Sc.

RWTH Aachen, SS 2022
Release date: May 23rd, 2022
Solution: Week starting from May 30th, 2022

Applied Data Analysis

R-Laboratory 6

Exponential Dispersion Family – Response Transformation

Task 22

(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 $\eta_j(\boldsymbol{\vartheta}) := \vartheta_j$, $j = 1, \dots, k$ for $\boldsymbol{\vartheta} = (\vartheta_1, \dots, \vartheta_k) \in \Theta \subset \mathbb{R}^k$). 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 23

(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 ontaining only the explanatory variables with a significant effect on the response.

(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?

```
#######################
########TASK22########
#######################

#a) i)
# define the exponential family function
my.exp.fam = function(a, b, my.c, pi){

  return(my.func = function(y, mu){
    return( (exp(y*mu - b(mu))/(a(pi)) + my.c(y, pi)))
  })
}
#a) ii
# define the EXP family by using T, h, and B
my.exp.fam2 = function(my.T, h, B, eth){
  return(my.func2 = function(y, mu){
    return(h(y, mu)*exp(sum(eth(mu) * my.T(y)) - B(mu)))
  })
}
# define the a function
my.a = function(pi){

  if(pi >= 0){
    return(pi)
  }else{
    warning("it should be bigger than 0!")
  }
}
# define the b function
my.b = function(theta){

  if(theta < 0){

    return(-log(1-exp(theta)))
  }else{
    warning("this function expects to have theta less than 0")
  }

}
# define the c function
my.c = function(y, pi){

  return(0)

}
```

```
theta = -0.8
# run the function
new.expfam(x, theta)

# generate a matrix
comb.table = matrix(0, 2, 7)
# set the column names with 0,1,2,...,6
colnames(comb.table) = x
# set the row names with own pdf and real
rownames(comb.table) = c("expfam", "dgeom")
# set column values with the densitys
comb.table[1,] = new.expfam(x, theta)
comb.table[2,] = dgeom(x, 1-exp(theta))
# barplot
barplot(comb.table,
        beside = TRUE,
        col = c("blue", "red"),
        ylab = "probablity",
        xlab = "value",
        legend.text = TRUE,
        args.legend = list(x="topright"))

n = 200

# we now get the 200 random sample
# and estimate the density directly
# retrieve the sample
geom_sample = table(rgeom(200, 1-exp(theta)))
# filter the unique values
observed.vals = as.numeric(names(geom_sample))

plot.vals = 0:max(observed.vals)
# generate the combination table to plot
# set the size of table
comb.table = matrix(0, 2, length(plot.vals))
# set the columns
colnames(comb.table) = plot.vals
# set the row names
rownames(comb.table) = c("rgeom", "own geom")
# assign the density values into the matrix
comb.table[1, observed.vals + 1] = geom_sample/200
comb.table[2, ] = new.expfam(plot.vals, theta)
# plot'em
barplot(comb.table,
        beside = TRUE,
        col = c("blue", "red"),
        ylab = "probablity",
        xlab = "value",
        legend.text = TRUE,
        args.legend = list(x="topright"))
```