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

RWTH Aachen, SS 2021
Release date: June 8$^{\text{th}}$, 2021
Solution: June 15$^{\text{nd}}$, 2021

Applied Data Analysis

R-Laboratory 7

Newton Raphson – Model Selection – Gamma Regression

Useful packages and functions:

- `AIC()`
- `BIC()`
- `step()`
- `glm()`

Task 21

Let $Y \sim \Gamma(\alpha, \beta)$ be gamma distributed with shape parameter $\beta > 0$ and rate parameter $\alpha > 0$, i.e. Y has the pdf

$$f(y\,;\alpha,\beta) = \frac{\alpha^\beta}{\Gamma(\beta)} y^{\beta-1} \exp(-\alpha y), \qquad y > 0,$$

where Γ denotes the gamma-function.

(a) Implement an R-function with your own implementation of the Newton-Raphson algorithm (see Algorithm II.2.26 in the lecture) to compute the maximum likelihood estimator of α for an iid sample $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \Gamma(\alpha, \beta)$, $n \in \mathbb{N}$, where β is assumed to be known. As a termination criterion, check if $|\alpha^{(t)} - \alpha^{(t+1)}| < \varepsilon(1 + |\alpha^{(t)}|)$ with $\varepsilon = 10^{-8}$ holds, where $\alpha^{(t)}$ denotes the value assigned to α at iteration $t \in \mathbb{N}$. Implement a second termination criterion $\delta|\alpha^{(t)}| < |\alpha^{(t+1)}|$ for $\delta > 0$ to check if the algorithm diverges. The function should get the following input variables:

- The observed sample $\boldsymbol{y} = (y_1, \ldots, y_n)$ for an arbitrary $n \in \mathbb{N}$,
- the starting value $\alpha^{(0)} > 0$,
- the termination criterion values $\varepsilon, \delta > 0$ with default values $\varepsilon = 10^{-8}$ and $\delta = 100$.

The function should return a list with the following output variables:

- a boolean indicator if the algorithm converged or diverged,
- the final guess for α.
- the number of iterations T required for convergence

Test the function on a simulated $\Gamma(2, 2)$ sample of size $n = 20$ generated by the `rgamma(20,2,2)` function.

(b) For the sample sizes $n = 10$, 100, 500, 1000, create 1000 data sets of the $\Gamma(2, 2)$ distribution and estimate the rate parameter $\alpha = 2$ using the MLE and the Newton-Raphson algorithm of (a). Compare the mean of the MLEs with the true value of α and $\mathcal{I}^{-1}$ with the sample variance of the MLEs for all n, where $\mathcal{I}$ denotes the Fisher information at $\alpha = 2$.

Hint: If the Newton-Raphson algorithm diverges in some cases, then only take the values of $\hat{\alpha}$ for which the algorithm converges to compute the sample mean and variance of the $\hat{\alpha}$s, reporting the percentage of times the algorithm diverged.

Task 22

(a) Download the file *Windmill.dat* from RWTHmoodle and load it as a data frame into your workspace. Transform the attribute `bin1` to type factor.

(b) Divide the data in training and testing data randomly. The training data should consist of about $\frac{2}{3}$ of the rows of *Windmill.dat*.

(c) Fit on this data frame a liner model with formula

$$\begin{aligned}\texttt{Cspd} \sim\ & \texttt{Spd1} * \texttt{Spd1Lag1} + \texttt{Spd2} * \texttt{Spd2Lag1} + \texttt{Spd3} * \texttt{Spd3Lag1} + \texttt{Spd4} * \texttt{Spd4Lag1} + \\ & \texttt{Spd1sin1} + \texttt{Spd1cos1} + \texttt{bin1} + \texttt{Dir1}.\end{aligned}$$

and compute the AIC and BIC for this model.

(d) Next compute the AIC and BIC for the linear model with formula

$$\begin{aligned}\texttt{Cspd} \sim\ & \texttt{Spd1} + \texttt{Spd1Lag1} + \texttt{Spd2} + \texttt{Spd2Lag1} + \texttt{Spd3} + \texttt{Spd3Lag1} + \texttt{Spd4} + \texttt{Spd4Lag1} + \\ & \texttt{Spd1sin1} + \texttt{Spd1cos1} + \texttt{bin1} + \texttt{Dir1}.\end{aligned}$$

Compute the AIC and BIC. Which of the models considered in (c) and (d) do you prefer?

(e) (i) Search for the linear model with lowest AIC value out of all models nested in the model (c) using a backward, a forward strategy and both. What do you observe?

Hint: You may use the function `step` with the model of (c) considered the most complex model.

(ii) Now search for the linear model with lowest BIC value analogously.

Hint: If n denotes the number of rows of the data set, you can search with respect to BIC by setting `k=log(n)` in the `step` function.

(f) Compare the models selected in (e) (i) and (ii) using the testing data and compute the so called *predicted residual sum of squares* (PRESS)

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 ,$$

where $\hat{\mu}_1, \ldots, \hat{\mu}_n$ denote the predicted values of a model considering the test data set. Which of the models do you prefer based on this criterion?

Task 23

Let $X \sim \mathcal{U}(20, 80)$ be uniformly distributed on $[20, 80]$ and let $Y|X = x \sim \Gamma(1, \beta(x))$ be gamma-distributed conditional on X with constant rate parameter $\alpha = 1$ and shape parameter specified by $\log(\beta(x)) = -2 + 0.08x$.
Randomly generate $n = 25$ independent observations from this model. Fit the model in R using the `glm` function. Calculate $\text{corr}(y - \hat{\mu}, \hat{\mu})$. Do the same for $n = 100$, $n = 1000$ and $n = 10000$ and summarize how the computed correlations depend on n.