Exam 2019 Generalized Linear Models

B-KUL-G0A18A

Emmanuel Lesaffre

L-Biostat, K.U.Leuven, Leuven

June 2019

Included is the description of 2 exam projects. The data are put in separate files, also attached.

Practical arrangements:

1. Deliver your exam project on a **pdf** file. The document should have a title page with heading: **2018-19 GLM course KULeuven,** the names of the members of the group and their email addresses. Allocate a contact person (one that submits the project) and put that name first in the list. Send the document via **email to martial.luyts@kuleuven.be** at the latest on **Saturday June 1, 2019 12.00h (noon).**

   Note that this is AN ABSOLUTE DEADLINE, submissions after this deadline will NOT be taken into account.

2. **Send also your programs** to me in the same email, so that we can check that the programs really work! **But do not include them in the main pdf file!**
3. Use the software that was used in the course, to solve the questions, don't use any other software.
4. For clarity, **first repeat the question** in your document and then give your solution.
5. Describe the flow of your procedures and the reasoning behind. Remember that there exist many good statistical analyses, most important is that you give a clear motivation of performing the analysis in your way.
6. Mention each time the software that you have used
7. Annotate your output, but not everything that you have done needs to be put in the report. **Limit your report to 20 pages (excluding title pages, but including tables and figures)! You can add an appendix for additional, but not essential, info.**
8. Print a version for yourself and bring it with you at the oral defense and **DO NOT write anything on your version. So DO NOT write on YOUR VERSION!!!**

9. **Study also the course material**; you will get general questions on the course material at the oral exam.
10. **Good luck!**

The projects

Project 1: Bike sharing systems.

Bike sharing systems are new generation of traditional bike rentals where whole process from membership, rental and return back has become automatic. Through these systems, user is able to easily rent a bike from a particular position and return back at another position. Currently, there are about over 500 bike-sharing programs around the world which is composed of over 500 thousands bicycles. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues.

This dataset (**bike_sharing.xlsx**) contains the daily count of rental bikes in the year 2011 in Capital bikeshare system in Washington, DC with the corresponding weather and seasonal information. The following variables are in the dataset:

instant	Record index
dteday	Date
season	Season (1: Springer, 2: Summer, 3: Fall, 4: Winter)
mnth	Month (1: January, 2: February, ..., 12: December)
holiday	If day is holiday or not (extracted from http://dchr.dc.gov/page/holiday-schedule)
weekday	Day of the week (0: Sunday, 1: Monday, ..., 6: Saturday)
workingday	If day is neither weekend nor holiday is 1, otherwise is 0
weathersit	Weather situation (1: Clear, Few clouds, Partly cloudy, Partly cloudy; 2: Mist+Cloudy, Mist+Broken clouds, Mist+Few clouds, Mist; 3: Light Snow, Light Rain+Thunderstorm+Scattered clouds, Light Rain+Scattered clouds; Heavy Rain+Ice Pallets+Thunderstorm+Mist, Snow+Fog)
temp	Normalized temperature in Celsius. The values are divided to 41 (max)
atemp	Normalized feeling temperature in Celsius. The values are divided to 50 (ma x)
hum	Normalized humidity. The values are divided to 100 (max)
windspeed	Normalized wind speed. The values are divided to 67 (max)
casual	Number of casual users per day
registered	Number of registered users per day
cnt	Number of total rental bikes per day including both casual and registered

Tasks:

- Extract the following variables from the data set: weathersit, workingday, temp, hum, windspeed, season and cnt.
- **Analysis:**
    - Set up a Poisson regression model that predicts the number of total rental bikes (cnt) in a frequentist manner with the R software used in the course.
    - Select the most predictive regressors in a classical GLM manner and interpret your results.
    - Also select the most predictive regressors in an extended GAM manner.
    - Do the necessary checks to verify that the chosen model(s) fit the data well, e.g. check the link function, the scale of the covariates, etc. In other words, choose the most appropriate procedure. Illustrate your findings with appropriate graphics and illustrate how good prediction.
    - Fit a negative binomial model and a quasi-Poisson model in a frequentist manner if there is overdispersion.

Reference:

Fanaee-T, Hadi, and Gama, Joao, ``Event labeling combining ensemble detectors and background knowledge'', Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg, doi:10.1007/s13748-013-0040-3.

Project 2: Internet Movie Database

A commercial success movie not only entertains audience, but also enables film companies to gain tremendous profit. A lot of factors such as good directors, experienced actors are considerable for creating good movies.

The dataset (**IMDb.xlsx**) is composed from IMDb (Internet Movie Database), an online database of information related to films, television programs, etc., with fan reviews and ratings. It contains 26 variables for 312 English movies, spanning from 2005 till 2008.

The variables in the data set are:

movie_title	Title of the movie
duration	Duration in minutes
director_name	Name of the director of the movie
director_facebook_likes	Number of likes of the director on his facebook page
actor_1_name	Primary actor starring in the movie
actor_1_facebook_likes	Number of likes of the actor 1 on his/her facebook page
actor_2_name	Other actor starring in the movie
actor_2_facebook_likes	Number of likes of the actor 2 on his/her facebook page
actor_3_name	Other actor starring in the movie
actor_3_facebook_likes:	Number of likes of the actor 3 on his/her facebook page
num_user_for_reviews:	Number of users who gave a review
num_critic_for_reviews:	Number of critical reviews on IMDb
num_voted_users:	Number of people who voted for the movie
cast_total_facebook_likes:	Total number of facebook likes of the entire cast of the movie
movie_facebook_likes:	Number of facebook likes in the movie page
plot_keywords:	Keywords describing the movie plot
facenumber_in_poster:	Number of the actor who featured in the movie poster
genres:	Film category (`Animation', `Comedy', `Romance', `Horror', `Sci-Fi', `Action', `Family')
title_year:	The year in which the movie is released (2005:2008)
country:	Country where the movie is produced
content_rating:	Content rating of the movie
gross:	Gross earnings of the movie in dollars
budget:	Budget of the movie in dollars
profit:	Profit of the movie in dollars
imdb_score:	IMDB Score of the movie on IMDB

Tasks:

- Extract the following variables from the data set: budget, profit, director_facebook_likes, content_rating and duration.
- **Analysis:**
    - Make a descriptive analysis to look at the relationship between the covariates and the profit. Covariates used in this analysis are content_rating, budget, director_facebook_likes.
    - Model the profit as a function of budget only, use the following techniques:
        - Polynomial regression model
        - Truncated polynomial splines of degree 2 (consider k=2, 3 and 5 knots)
        - B-splines of degree 2 (consider m=3, 5 and 8 knots)
        - Cubic P-splines (consider k=5, 8 and 20 knots)
    - Include the other covariates in the model and determine what variables show an impact on the profit.
    - A movie is defined successful when the profit is positive. Fit a model that relates the probability of success and the covariates considered above.