STA303 Methods of Data Analysis 2

Assignment 1 - Due February 7th, 2025 by 11:59PM ET

LEARNING OUTCOMES

By completing this assignment, students are demonstrating they are able to do the following:

- ☐ Select an appropriate generalized linear model for a given dataset.
- ☐ Fit generalized linear models of different types on data using R.
- ☐ Recognize and describe situations in which overdispersion is present.
- ☐ Interpret results (i.e. coefficients, parameters, etc.) from generalized linear models correctly.
- ☐ Write up an analysis, including existing knowledge, methods applied, results, and conclusion.

ASSIGNMENT SUMMARY

This assignment requires material up to and including Week 4 (end of GLMs).

Students (either individually or in groups of two) will be asked to consider a research question for a given dataset. The question will require fitting an appropriate generalized linear model (GLM) for the provided data to address the research question or goal. Students will need to search the existing peer-reviewed literature to understand what is known about this research area/problem and to highlight the importance/relevance of this problem.

Students will be responsible for displaying model results from their analysis and results of exploratory data analysis of their data in a professional manner. A short report will be submitted for assessment that describes and justifies the statistical approach used, the results obtained, and interpretation and contextualization of results to the research goal.

RESEARCH GOAL

The goal of this assignment is to answer the following question: ***"How do literacy and age of a marriage affect family size?"***. By European standards, Portugal is a poor country, and in 1980, it had the same GDP per capita as Mexico. It is also well known that families are larger in rural areas. Therefore, in addition to estimating the effects of various factors on family size, you must also determine how much variation (if any) there is in birth rates after accounting for known explanatory variables.

The data you will use is from a fertility survey conducted in Portugal in 1979.

An expert has reached out and informed you that there is no need to consider any zero-inflation for these data, as they don't fit well, probably because birth rates are lower than Fiji, so many zeros are expected and the likelihood is flat.

Data

The data can be accessed through the Quercus assignment page. The link will send you to JupyterHub where you will find an RData file as well as code provided to read and clean the dataset.

For more information about these data:

- More information about the World Fertility Survey, Portugal 1979-1980
- Data source
- Data dictionary (although we have isolated the variables you will work with)

REPORT REQUIREMENTS

Introduction

The purpose of the introduction is to provide your reader with the importance of the results to come and an overview of what is already known about the topic and/or research question/area. Your introduction should therefore include:

- A strong and relevant motivation as to why it is important to study this particular area or research question (with a supporting peer-reviewed citation).
- Summaries of the main results from three peer-reviewed articles on the same or related research areas/questions, with a focus on understanding which variables have been shown to be related to the response variables in your data.

Methods

The purpose of the methods section is to outline the statistical procedures that will be used to address the research question. Your methods should therefore include:

- The choice of generalized linear model that will be fit or those that will be considered.
- A statistical justification for the choice of model that is in alignment with the data and research goal.
- The process to decide on the appropriate model for the data.

Results

The purpose of the results section is to introduce the data and properties of variables being considered and to present the results of the procedures outlined in the methods section, with emphasis on how these provide the answer for the research question. Your results should therefore include:

- Statistical summaries of the data and variables being considered (e.g. table of numerical summaries of the variable distributions, or plots of univariate or multivariate distributions of variables)
- Discussion of the data distributions and their preliminary implications regarding the research question.
- Presentation of model results (e.g. estimated coefficients, standard errors, CIs) and evidence supporting inclusion/exclusion of predictors.
- Discussion of the process and results involved in developing a suitable model that addresses the research goal.

Conclusion

The purpose of the conclusion is to interpret and contextualize the results in the previous section to provide the reader with a complete answer to the question. Your conclusion should therefore include:

- A formal statistical interpretation of a coefficient of at minimum one of the most important or relevant predictors.
- A general summary of what your model as a whole says about the research question.
- A comparison to the literature results summarized in the introduction section, commenting on what is similar and what is different.

TECHNICAL REQUIREMENTS

Your report must meet the following requirements

- Be written in R Markdown or LaTeX and knitted to a PDF document.
- No R code should be present in the knitted PDF report (other than appendices/Rmd file).
- All figures and tables must be of professional quality
    - Have informative captions outlining 1) what is presented, and 2) what is the key take-away
    - Avoid using R's variables and names (i.e. say "the 20-25 age group" instead of "age20_25".
- Include a bibliography/reference list containing the full citations for the peer-reviewed articles (in an appropriate citation format like APA)
    - In-text references to items in the bibliography should be of a consistent style to the bibliography.
- Reports should **not exceed 1500 words** in length. Be sure to edit your writing for conciseness and clarity.
- Reports should contain **no more than 5 figures/tables** (e.g. 2 tables and 3 figures).

WORKING IN PAIRS

Students have the option to work in groups of two (i.e., pairs) for this assignment. You can work with any student from any section of the course. To create your group:

- Go to the People tab in the navigation bar on Quercus
- Click on Groups
- Find an empty group (or a group that already contains the person you wish to work with) and click Join.

Students should finalize their group/pair as soon as possible to avoid delays in beginning the assignment.

NOTE: If you wish to work independently, you do not need to join a group (i.e. you do not need to do anything).

WHAT TO SUBMIT TO QUERCUS

If working in a pair, only one student needs to upload the below required components. A complete submission includes:

- ☐ The written report, as a PDF, that satisfies the above technical requirements.
- ☐ The Rmd file that contains your full written report and code.

HELPFUL RESOURCES

Writing Resources:

- Tip sheet with advice for writing a research report
- Writing centres on campus with bookable appointments to help with writing
- General advice on writing introductions and conclusions, and on organizing a paragraph of writing.

Searching for peer-reviewed papers:

- How to search for academic articles related to your topic
- How and Why to cite your sources
- List of citation generators to ensure you have the right format

R Markdown Resources:

- Cheatsheet for common R codes to manipulate and read data
- Cheatsheet for common R Markdown functions
- R for Data Science book with helpful information about visualizing and working with data
- R Markdown Cookbook with helpful information about adding captions and other formatting options

RUBRIC

E = Excellent (1, satisfies the criterion well with little to no revisions)
S = Satisfactory (0.5, satisfies the criterion somewhat but would benefit from some revisions)
NR/A = Needs Revision/Absent (0, major revisions needed to satisfy the criterion)

Criteria for Assessment	E	S	NR/A	Total
Introduction The introduction should be about 2-3 paragraphs in length. The introduction is self-contained and tells the reader everything they need to know to understand why this study is being conducted. This includes:				
1) A strong/convincing argument for the importance of this study, supported by peer-reviewed evidence (i.e. why this study is relevant to other people).				
2) Summaries of at least three peer-reviewed articles on the same area, topic, or question as this study. These should concisely describe the main conclusion/result in the context of the study sample.				
3) The research objective, goal, or question of this study is explicitly stated (i.e., what this study is going to do) and a strong connection is explicitly drawn between the literature summaries and this goal (i.e., how does the work of others fit into this goal/study).				
4) A brief but correct overview of the general statistical techniques that will be employed in the paper (i.e., what types of procedures will be used) and how these allow the research goal to be met.				
Methods The methods section should be about 2-3 paragraphs in length and should not present any results or data summaries. The methods section is self-contained and tells the reader what statistical procedures will be used in the study and why these are chosen. This includes:				
1) Correctly mentioning what type of statistical model(s) will be considered using appropriate course terminology.				
2) Correctly justifying why the statistical model(s) is/are appropriate for the problem, with specific reference to the type of data being considered and any other data properties that are relevant.				
3) Correctly explaining what predictor variables are considered of interest to the question and why other predictors are being included in the model.				
4) Correctly stating how it will be determined from the model(s) which predictors are significant and how much variation exists in the response.				

Results The results section introduces the data used and highlights important attributes or characteristics of the variables. It also walks the reader through the key steps in applying the statistical procedures mentioned in the methods to the data with the aim of addressing the research goals. This section includes:				
1) Statistical summaries of the variables considered in the analysis. These are appropriate/correct for the data being presented. Summaries should be univariate in nature (i.e., each variable separately) but may be supplemented by multivariate summaries if desired.				
2) A discussion of the statistical summaries presented, that correctly comments on at least one of the centre, spread or shape of each variable in the context of the sample population. Specific emphasis should be placed on the response distribution with correct commentary on all three of centre, spread and shape. If multivariate summaries are also presented, the discussion should describe how the relationship presented is relevant to the research problem.				
3) Presentation of the results of all statistical procedures applied to address the research question. Estimates from models should be presented with a corresponding measure of error/variability. Results of hypothesis tests should include the test statistic, the corresponding p-value and the significance level used.				
4) A discussion of the process involved in developing a suitable model that addresses the research goal. When referencing multiple models, descriptive language is used to aid the reader in distinguishing between them (e.g. avoid labelling models as “model 1, model 2”). Variables should be referred to conversationally and not using the names presented in R. Statistical procedures used should be explicitly stated, consistent with the methods section, and correctly applied. Evidence that supports decisions made towards developing a model that addresses the research goal must be referenced explicitly and used correctly. The discussion should flow logically, beginning with an initial model based on the cited literature and ending with the result that will be interpreted in the conclusion.				
Conclusion The conclusion section should be about 1-2 paragraphs in length and should refer to figures/tables from the results section as needed (generally no new evidence is presented here). The conclusion summarizes and interprets the results from the previous section and describes whether the results are consistent with what is already known about this area. This includes:				
1) A formal statistical interpretation of the effect on the response of at least one of the most important/relevant predictors in the chosen model and should be correct and in the context of the data.				
2) An informal interpretation of the model results, focusing on what the chosen model says about the relationship between predictors and response as a whole in this project. This should be consistent with the model results and in the context of the data.				

3) A comparison between the effects suggested by the model and the effects in the literature that was summarized earlier. The comparison should correctly address whether the effects from the model are supported by or contradictory to the literature, with in-text citations directing the reader to the relevant sources.				
4) A closing statement that highlights how the results of this study can be used to address the stated motivation from the introduction.				
Figures, Tables, and References Figures, tables and references provide the reader with evidence and supporting information to understand the work described. They should be well-formatted and easy to understand, and in the case of figures and tables, should act as standalone information that can be understood without reading the core text. Figures, tables and references should satisfy:				
1) The reference list/bibliography contains all source materials or articles mentioned in the written report. The list is well-formatted using an appropriate format (e.g., APA). In-text citations are included whenever information is summarized or mentioned from an external source (e.g. peer-reviewed article) and are in a consistent format as the reference list/bibliography.				
2) No R code is present in the written report (e.g. code that produces tables/figures), and R output (as displayed in the R console) is not present.				
3) Tables and figures include informative captions that cover 1) what is presented and 2) what is the key message the reader should take away. Captions should ensure that the reader does not need to have read the report itself to understand what is presented in the table/figure.				
4) Tables and figures are well designed (visually appealing, cohesive, and valuable). The information presented in each table/figure is easy to read and understand (e.g. avoids variable names from R, legible axis labels, etc.), makes sense to be presented in a single standalone unit, and is referenced and discussed in the main report text.				