STA 712 Project 2: Count regression

Due:

- Preliminary SAP: Friday, October 27 at 12:00pm (noon) on Canvas
- Project 2 report: Tuesday, November 21 at 12:00pm (noon) on Canvas

Overview

In Project 1, you explored research questions for a provided dataset. In Project 2, you will find your own data and and explore your own research questions.

Data

Find a dataset with a count variable (one which could be modeled with a Poisson, quasi-Poisson, negative binomial, ZIP, etc. model) which you are interested in modeling and investigating.

- If the observations in your data are not independent (i.e., if there is some correlation/group structure to the data) you will need to handle this too.
- Make sure your data has enough potential explanatory variables to ask an interesting research question.
- Make sure your data has enough observations to fit an interesting model (you probably want at least 30–50, and you will need more observations if you have more explanatory variables).
- Avoid data with too many missing observations.

Research question(s)

Develop at least one research question which you care about, and which can be investigated using your data and regression methods we have learned in STA 711/712. Your research question must involve a count variable as the response of interest, and several explanatory variables.

Requirements

You will submit two documents for this project:

- A preliminary statistical analysis plan (SAP) describing your data, your research questions, and the statistical methods you will use to investigate those research questions
- A final project report (like Project 1) describing the results of your analysis and addressing your research questions

I will give you feedback on your research questions and SAP to help guide your analysis.

Statistical analysis plan

This section should be approximately 1–2 pages (single-spaced), and should describe the data you will use and the analysis you will perform. Your analysis plan must address the following (Cressman & Sharp, 2022):

1. What is/are the research question(s) or objectives?
2. What are the variables that will be measured in the study? For each variable, what will the data look like?
    - This includes the response variable, the explanatory variable of interest, and any other variables to include in your analysis
3. What is the study design, and why is the study set up this way?
    - You (probably) aren't designing the data collection, so here I really want you to describe the following:
        - What is the source of the data, and how were they collected?
        - What is the size of the data, and what does a row in the data represent?
        - Are there any missing observations?
        - Is there any dependence (e.g. groups) between observations?
4. How will data and any patterns/results be summarized? (i.e., how you will do exploratory data analysis)
5. What will be examined statistically? Link directly back to the research questions.
    - Which models will you fit?
    - How will you assess the suitability of these models? (diagnostics, etc.)
    - How will you use the models to address the research question(s)? (hypothesis tests, confidence intervals, etc.)
    - What software will you use?
6. What alternative strategies should be considered, and when? (i.e., what will you do if assumptions are violated? How would you modify the model or statistical analysis?)
7. What are the statistical results that will be presented? (refer back to 4 and 5)

See the following article (particularly Table 1) for more guidance on crafting a statistical analysis plan:

> Cressman, K. A., & Sharp, J. L. (2022). Crafting statistical analysis plans: A cross-discipline approach. *Stat*, 11(1), e528.

Project report

For the project report, you will use your data to address your research questions. You will then communicate your findings in a written report. Your written report will be formatted similar to a submission to the academic journal *PLOS One*. Please read the *PLOS One* submission guidelines for full details, paying particular attention to the "Manuscript Organization", "Parts of a Submission", and "Statistical reporting" sections.

In particular, your manuscript must be organized with the following sections:

- **Abstract**: an overview of the main objectives and results
- **Introduction**: Background, motivation, a brief review of related literature, a summary of the research questions, and a summary of the main results
  - A brief literature survey of previous work related to the research questions (ballpark: 5–10 references) is required. All citations must be properly formatted.
- **Methods**: Description of the data, an explanation of how the research questions will be addressed, and a summary of the statistical methods used. Enough detail should be provided that a statistician could reproduce your analysis without access to the source code.
- **Results**: The main results of your analysis, including the fitted regression models, diagnostics and assessments of model performance, and hypothesis tests.
- **Discussion**: A discussion of the results in context of the research questions, and a conclusion summarizing the main findings of the analysis.
- **Supplementary information**:
  - A link to a GitHub repository containing the code you used in your analysis
  - Any supplementary figures or tables (e.g., additional EDA) not included in the main text
- **Acknowledgments**: Acknowledge any students with whom you discussed the analysis.
- **References**: A properly formatted bibliography for all references in the manuscript.

You will follow the *PLOS One* guidelines for manuscript style and format, with a few exceptions to make the manuscript easier to read:

- Figures and tables *should* be included in the manuscript
- You may use BibTex and a .bib file to format references
- Supplementary figures and tables may be included at the end of the manuscript, in the same file (you do not need to submit separate files)

Requirements

Content

Your report must contain:

- A brief literature survey of previous work related to the research questions (ballpark: 5–10 references). All citations should be properly formatted
- A description of the data, including
  - The source of the data
  - The size of the data, what a row in the data represents, what variables are available, and whether there is any missing data
- Any exploratory data analysis needed to address the research questions, including
  - A description of any transformations or manipulations (e.g., creating a new variable, removing missing observations, etc.)

  - A (well formatted) table summarizing the relevant variables (as an example, see Table 1 in the dengue paper from class)
  - Any essential figures informing your data analysis. **Note:** do not include all figures you create during your exploration of the data! Pick the most important 1–3 figures. Other figures can be included in the supplementary materials, if necessary.

- An explanation of how you chose the final model(s) to answer the research questions
- A (well formatted) table of coefficients for your final model(s) (see, for example, Table 3 in the dengue paper from class)
- Model diagnostics and assessments for your final model(s)
- When testing a hypothesis to answer a research question, clearly state the null and alternative hypotheses, report both a test statistic and p-value, and use the p-value to make a conclusion about the research question
- In the supplementary information, a link to a GitHub repository containing all code needed to reproduce the analysis in your report

Format and style

- The report should be written like an article or research paper: in full sentences and paragraphs, with headings for each section. You should not write your report with question numbers or as a list of bullet points. Scientific articles are generally written in third person, though "we" can also be acceptable ("we can see from Figure 2.1...").
- No R code or raw R output may be included in your report. Only properly formatted figures and tables may be included.
- Your document must be prepared using the LaTeX template provided, and submitted on Canvas as a pdf.
- Figures should have labeled axes, and should be clear and easy to read. Figures should also be captioned and numbered. Captions should provide enough information to understand what is being plotted, but interpretation can be left to the main text. Refer to figures by their number in the text. Make sure that any figures you include are discussed in the text.
- Tables should be nicely formatted, and have a number and caption. As with figures, refer to tables by their number in the text.

Grading

The project will be graded on a mastered/not yet mastered basis. To master the project, you must:

- Submit a statistical analysis plan on time (the SAP will be graded on completion only)
- Master the final project report

Mastery of the project report requires meeting all of the Criteria for Publication for *PLOS One* (see the link for full details on the criteria). I will grade your manuscript as a reviewer for *PLOS One*. I will make one of the following decisions, and provide feedback on your initial submission:

- *Accept*: the manuscript can be accepted as-is; no revisions are necessary.

- *Minor revisions:* the manuscript is in good shape, but some minor changes (formatting, some additional explanation, etc.) need to be made before acceptance
- *Major revisions:* the work has potential promise, but major changes (e.g., a revised or expanded analysis) need to be made

You will have one opportunity to re-submit the manuscript if you do not receive an "Accept" on your first submission. To master the project report, you must either:

- Receive an "Accept" on your initial submission, OR
- Receive either an "Accept" or "Minor revisions" on your second submission

Collaboration

You are welcome to work with other students on this project, but everyone needs to submit their own report, and each student needs to find their own data (you may not use the same dataset as a classmate). Acknowledge any collaboration in the "Acknowledgments" section of your report.

Resources

Finding data

There are lots of great places to find data. Jo Hardin at Pomona College has a great list here if you need help getting started.

GitHub

You are required to host your code for this project in a GitHub repository. See the GitHub docs for guidance on getting started.

LaTeX

Your written report must be created using LaTeX. See the course website for resources on getting started with LaTeX. Overleaf also has a lot of good tutorials to help you get started.

Google Scholar is useful for getting BibTex references. In Google Scholar, search for an article, then click on "Cite", then "BibTex". Copy the resulting BibTex citation into your .bib file.

LaTeX template

A LaTeX template for this project is available in a ZIP file on the course website.

- Download and extract the ZIP file
- Create a new project on Overleaf (e.g. "STA 712 Project 2")
- Upload all files and folders from the extracted template folder
- Compile
- Make edits to the manuscript, references, and images as needed

Figures and tables

- Examples are provided in the LaTeX template
- See the Overleaf tutorial on including figures and tables in LaTeX documents
- The `xtable` R package is useful for creating LaTeX tables
- The online LaTeX table generator is also helpful
- Consider the `gridExtra` and `patchwork` R packages for arranging multiple R figures into a single image