Statistical Inference

HW Author: Reza Salamat
Instructor: Mohammadreza A. Dehaqani

Spring 2024

Final Project

Introduction

Welcome to your final project! Our main goal is to immerse you in a real-world data science experience. Throughout this course, we've emphasized that the ultimate goal is to help you advance even at least one step in both academia and industry. This project aims to assess your progress so you can better seek advice and support in areas where you need improvement whether from the teaching team or other sources you see fit.

Imagine you are a data scientist in a professional setting. Typically, you'd be handed a dataset or a database to work with. Here, however, you have the exciting task of sourcing your own data from trusted public sources. Choose a dataset that seems interesting to you (at least interesting enough!!) and can effectively tell a story through statistical inference.

To guide you in finding a good dataset and then doing the correct analysis, we've outlined a series of suggested TODO tasks in the next section. These will give you a clear idea of what you need to deliver in your data analysis.

Please select your dataset within one week from the upload date, have it approved by your TA team, and then add a new row to this Google sheet listing your chosen data.

Project TODO

You have the flexibility to choose up to 5 items from the visualizations section, 6 from each of the estimation and testing sections, and up to 3 from the regression section, giving you a total of up to 20 tasks to work on. However, don't worry about possible errors and potholes you might fall into. As long as you successfully complete and get 16 of them accepted, you will receive full credit for the TODO section. This allows you to focus on the tasks that interest you most while ensuring you cover a good range of what you have learned and also be forced to go beyond that. Remember, you have to have a good reason to choose any of the stated analyses. These reasons make up at least 15% of your grade. Now, this reason could also be EDA (Exploratory Data Analysis), but even for that, you'd need an acceptable hunch so you can explain it in your Conclusion and Discussion section.

(1) Visualization and Summarization of Dataset Variables

- Perform exploratory data analysis by visualizing categorical, numerical, and mixed variables. Use appropriate plots like bar charts for categorical data and histograms for numerical data.
- Analyze the distribution and relationship of variables. Create scatter plots, box plots, and QQ-plots to understand variable dependencies.
- Create pair plots to visualize relationships between multiple numerical variables simultaneously.
- Identify and report any dependent or independent factors discovered during analysis.
- Generate heatmaps to visualize the correlation matrix of numerical variables.

- Explore and visualize metrics that capture relationships beyond linear correlation, such as Spearman's rank correlation, Kendall's Tau, or mutual information.
- If the dataset comes from a published source, include and discuss any visualizations provided in the publication.

(2) Parametric Inference and Estimation

- Utilize interpolation methods to predict missing or unobserved data points, if necessary.
- Conduct parametric inference methods where applicable, explaining the choice of parametric tests.
- Apply estimation techniques to various variables, such as point estimation or maximum likelihood estimation.
- Conduct goodness-of-fit tests to determine how well the data fits a specified distribution.
- Calculate and interpret confidence intervals for the estimated parameters. (It might be a good idea to create plots to visually represent confidence intervals)

(3) Hypothesis Testing and Statistical Analysis

- Formulate and test hypotheses based on the dependency of variables. Clearly state the hypothesis, the rationale for testing, and the method used.
- Perform power analysis to determine the sample size required for detecting effects with desired power.
- Report all statistical findings, including test statistics, p-values, and conclusions drawn.
- Employ bootstrap or resampling methods for hypothesis testing, especially in cases of non-normal data distributions.
- Apply non-parametric tests such as the Mann-Whitney U test or the Kruskal-Wallis test for data that do not meet parametric assumptions.
- Conduct Analysis of Variance (ANOVA) tests to compare means of different groups. Use correction methods like Bonferroni or Tukey's HSD for multiple comparisons, if applicable.
- If the dataset comes from a published source, Critically analyze and discuss any tests presented in associated publications and compare them with your findings.

(4) Regression Analysis and Reporting

- Perform regression analysis to investigate relationships between variables. Explain the choice of regression model and interpret the results.
- Apply regularization techniques such as Lasso or Ridge regression to handle multicollinearity and enhance model performance.
- Perform diagnostic tests on regression models to check for assumptions like homoscedasticity, multicollinearity, and independence of errors.
- Include a detailed analysis of regression findings from the publication, if available.

  Note: If you haven't studied regression, you can instead choose 6 Visualization (Section **1**) tasks and 7 tasks from Estimation and Testing (Sections **2** and **3**)

Discussion and Conclusion

One of the most important human values is our spirit! Our ability to see new patterns, think outside the box, and be creative sets us apart. While many complex tasks are becoming automated, our creativity can enhance these automations or even lead to entirely new methods. In the spirit of keeping your creativity alive, 20% of this project involves telling a compelling story using the dataset you chose through the tasks you performed in the previous section.

Depending on the significance of your conclusions, we expect each story to be one or two paragraphs long (graphs or figures not included). This exercise is designed to help you develop the skill of telling compelling stories through statistics, which is invaluable in both academia and industry.

Each story should follow this structure:

- Begin with one or two visualizations. Describe what you observed and the hypothesis or question that arose from these visualizations.
- Explain the estimation or tests you performed to investigate your hypothesis. Include the rationale behind choosing these methods.
- Discuss any further statistical analysis you conducted, such as regression analysis, to confirm or refine your findings.
- Conclude with your final inference and its significance.

Feel free to be creative and thorough in your analysis. Your ability to connect different aspects of your analysis into a coherent narrative is just as important as the technical accuracy of your work.

Final Notes

- Students are required to submit a document (draft) detailing their project work, methodologies, and results. In academic terms, a 'draft' refers to a preliminary version of your work. It should comprehensively outline your project's progress and findings, though it may not be in its final, fully polished state. This draft serves as an essential step in developing and refining academic reports.
- This project will have an in-person (or online) hand-in in which you will be asked questions about what you turned in to assess how fluent you are in your own work. We understand that it's the end of the term, and one might not have enough time, so alternatively, you can create a 5-10 minute video presentation summarizing their findings and interpretations (i.e., record their screen while presenting their presentation). The presentation should be submitted along with the project document. It;s sincerely recommended that you chose live hand-in since it will probably save time and you can better explain your own work.
- Final Word: In addition to adhering to the instructions in the four outlined sections, students are expected to engage in further research and decision-making as data scientists on their own. Should certain aspects within these sections not be fully explained, you are encouraged to independently seek out the necessary information and decide on appropriate approaches. While support will be available as it's been so far in the course, your initiative in exploring beyond the provided guidelines and justifying your methodologies is essential and highly valued in this project. Think of this as your first work beyond the realm of "Courses and Grades." Do not worry about grades, and focus on doing work you are interested in. But remember, no effort shall go unrewarded (or ungraded!)

Good luck on your journey with the data,
Your Teaching Team