STAT 412/512: Methods of/for Data Analysis II Assignment

- Mark Greenwood (Statistics, Department of Mathematical Sciences)

Assignment Notes: STAT 412/512 is a senior level/graduate co-convened course in statistical modeling and interpretation that is taken by both statistics students and students from other departments. A primary focus of the course is on communicating statistical results accurately, which includes framing results correctly based on the design of the study. This relates to how we write research questions as well as a discussion that we call "Scope of Inference" (SOI) where students address generalizability and causality of results based on the study design. Communicating statistical results starts with having clear and concise writing - if the writing is poor, can the reader really trust the statistical analysis results you are presenting were also done with care and attention to details?

Before students write their own report, they work with a "Demonstration Report" to learn about the intended structure of reports for the course, with some key sentences left for the students to fill in and the statistical analysis to complete to match the description in the report. The writing in the provided report is not perfect and could be improved but the improvements and edits might not be supported by the details in the results. This lab would occur a few weeks into the semester after they have seen/reviewed all of the statistical tools used in the models discussed.

License: This assignment is licensed under CC BY-NC 4.0

Purpose: This lab will involve using generative AI to potentially improve the introduction and to develop the SOI that would wrap up the report as well as to more easily generate citations to use in writing reports. By working with components of this demonstration report, students are more prepared to write their own reports in the desired style as well as learn about aspects of the study used to motivate the report. This would be completed in groups of 3 or 4 with circulating instructors available for clarifications/discussions.

Duration: 1 hour

Learning Objectives

By the end of this assignment, students will be able to:

1. How to use generative AI to edit and improve writing but also learn to use those suggested edits with caution.
2. How to write an SOI for a given study using generative AI by providing aspects of the study and then edit/modify results to match the study design.
3. How to use generative AI to aid in developing a references section for a report.

Materials Needed

- Personal computer, likely with R and RStudio installed and internet access.

Generative AI Tool for Assignment

For more in-depth information on this AI tool, see attached model card.

- Name of tool: https://gemini.google.com/app (suggested, but not required)
- Purpose: AI assistance, especially for writing support
- Who created it/Version used: Google AI/ Gemini 1.5 Flash
- Any known limitations/biases: Biases are based on the data used to train the model

Assignment Structure

Task 1:

Input the following paragraph from the introduction to the "Demonstration Report" into your selected generative AI with a request to modify/edit/improve some aspect or aspects of the provided writing:

"Streams play an important role in the emissions of greenhouse gases (GHGs) such as CO2. To better characterize this role, data were collected on three rivers on the Tibetan Plateau: the Yangtze (YZ), the Yarlung-Tsangpo (YT), and the Yellow (YL). Sampling was performed in 2014-2015, collecting partial CO2 pressure (pCO2) at each river site in µatm (Qu et al., 2017) as well as site elevation (in meters) and river name. Since both intra- and inter-river elevations variations may influence atmospheric pressure on dissolved gases, and river characteristics can impact CO2 emissions, we investigated how elevation impacts pCO2 and if those impacts vary across rivers."

Report the suggested edits to the paragraph. Then discuss each edit - would you accept them or not, do they suggest other edits you might consider? Do the suggestions match the state of (your) knowledge about the field and the study being considered?

- *Optional: Subtask: Ask for modified edits based on a persona (something like: revise those suggested edits as if you were a … working on …) and/or more detailed prompt (make the writing more … or condense the writing or ….). Do the suggested edits change? In what way did they change? You can use these modified versions or your original prompt results.*
- *Optional Subtask: Complete this part of the assignment using different generative AI platforms and compare results. Which platform provided the most useful edits?*

Task 2:

Provide information about the variables and the study design (is there random assignment and, if so, what variables were assigned? and was there random sampling? when and where were the data collected?) and request a sentence that addresses generalizability of the results (can you make inferences to a larger population or not?) and then request a second sentence that addresses whether causal inference is possible in this situation based on the similarly provided context of this study. Make sure you focus on a particular model for the discussion, as the causal aspects might change based on the variables being considered and how they are being used in the model you are focused on.

Report the suggested generative AI sentences and then edit/modify them to be versions you would want to include at the end of the report in the SOI section. Discuss how you had to edit the results.

Task 3:

Formatting citations can often be challenging when they come from disparate sources. Extract the noted citation information from various locations in the provided report using `citation("Rpackagename")` in R for obtaining citation information for the R packages used and other sources for other citations. Note that all citations were highlighted in the demonstration report to help you to identify the needed citations. Provide the information to your selected generative AI and use it to make sure the formatting is consistent with a particular citation style and in alphabetical order. Submit your citations/references section.

Task 4:

Reflect on the use of generative AI for each of the three tasks. How well did your chosen generative AI do the requested tasks? Rank your chosen generative AI in its usefulness on the three previous tasks and explain the reason for your ranking.

Task 5:

Moving forward, will you use generative AI for similar tasks in the future? What ways have you or do you expect to use generative AI in your academic career?

Submission Guidelines

- Format: Inline answers for each task, saved as a PDF document.
- Length: approximately 4 pages, might exceed based on prompts used.
- Citation Guidelines: No specific style, but consistency is important in the section on generating citations.
- Grading Rubric: Each task will be assessed on a scale from 0 to 5, with 0 being not attempted, 1 being minimal effort or incomplete reporting of results, 2 being partially incomplete or poor effort, 3 being a moderate effort but notable missing/incorrect on some parts, 4 being good work on some parts but missing/incomplete on others, and 5 being excellent and complete work.

Additional Resources

- Data for the report were extracted from https://static-content.springer.com/esm/art%3A10.1038%2Fs41598-017-16552-6/MediaObjects/41598_2017_16552_MOESM1_ESM.pdf

Additional pages below provide the demonstration partial report that motivates the previous tasks with additional coding and writing tasks for a follow-up assignment that would put the missing results and writing into the report, which relates to the indicated Qs that are incomplete in the following.

Impacts of Elevation and River on partial CO2 concentrations

I. Introduction

Streams play an important role in the emissions of greenhouse gases (GHGs) such as CO_2. To better characterize this role, data were collected on three rivers on the Tibetan Plateau: the Yangtze (YZ), the Yarlung-Tsangpo (YT), and the Yellow (YL). Sampling was performed in 2014-2015, collecting partial CO_2 pressure (pCO_2) at each river site in µatm (Qu et al., 2017) as well as site elevation (in meters) and river name. Since both intra- and inter-river elevations variations may influence atmospheric pressure on dissolved gases, and river characteristics can impact CO2 emissions, we investigated how elevation impacts pCO_2 and if those impacts vary across rivers.

II. Statistical Procedures Used

All analyses were conducted using R (R Core Team, 2024). The pCO_2 observations for the three rivers were visualized with enhanced strip charts from the catstats2 package (Greenwood, 2024) in Figure 1. Summary statistics for pCO2 and elevation by river are provided in Table 1 based on the modelsummary package (Arel-Bundock, 2022). The estimated means pCO2 of YZ and YL were similar at 1054.25 and 1083.364 µatm, respectively, while the mean of YZ was noticeably lower at 595.4 µatm. However, pCO_2 observations, particularly in the YT and YL rivers, had a wide range of pCO_2 values, from 300 to more than 1700 µatm. The number of sites varied by river with $n_{YZ} = 4$, $n_{YT} = 15$, and $n_{YL} = 11$, respectively, providing unbalanced design with respect to the rivers and a total sample size of n = 30. A scatterplot was utilized to examine the relationship between pCO_2 and elevation by river with both linear and nonparametric smoothing lines (Figure 2, ggplot2 package, Wickham, 2016). From Figure 2, a negative relationship between pCO_2 pressure and elevation was evident but the slope of the relationship for YT was more negative than for YZ or YL. There may be some curvature in the relationship for YT and increasing variance in the YT observations as elevation increases or what might be a possible outlier.

Linear models were used to model pCO_2 using the lm function. After fitting a preliminary model of $\mu\{pCO_2|River\&Elevation\} \sim River*Elevation$, a suite of diagnostic plots were generated using the ggResidpanel package (Goode et al., 2024, Figure 3). In the Residuals vs. Fitted plot (Fig. 3, upper left), the spread of the residuals increases as fitted values increase, which raises concerns about the assumption of constant variance and a slight curve may suggest a violation of the linearity assumption. From the QQ-plot (Fig. 3, upper right), there is evidence of a violation of the normality assumption with a clear right-skewed distribution of the residuals. To address the potential violations of linearity, constant variance, and normality assumptions, the response variable pCO_2 was log transformed (natural log) and the model was refit and new diagnostic plots were produced (Figure 4). After log transformation of the response variable, there was little or no evidence against the assumptions of constant variance (Residuals vs. Fitted does not show changing spread in the residuals as a function of fitted values), linearity is less clearly violated (limited curvature in Residuals vs. Fitted), and normality of residuals (QQ-plot of residuals does not show clear deviations from normality of the residuals). The partial residuals in the effects plots (Fox and Weisberg, 2018, Figure 5) for this model suggest that ...[Q1]

To address the question of interest, it was necessary to determine if a River by Elevation interaction term was needed in the model. To assess this, a Type II F-test (car package, Fox and Weisberg, 2019) was used on the $\mu\{\log pCO_2\} \sim$ River * Elevation model. Weak evidence was found against the null hypothesis of no interaction between River and Elevation on the log-pCO2 was found (F(2,24) = 1.0759, p-value = 0.357), so the interaction term was dropped from the model.

Diagnostic plots of the additive model with River and Elevation did not indicate further issues with the normality or constant variance assumptions (Figure 6). However, one observation had a Cook's distance value of XXX [Q2] in the Residuals vs Leverage Plot (Figure 6, lower right), qualifying as potentially influential observation. Examining that YL observation more closely, this point had both the highest elevation (4091 m) and pCO_2 (1771 µatm) measurement in the data set. This was contrary to the generally observed trend of decreasing pCO_2 values with increasing elevation (see Figure 2). However, without information about this observation that explicitly showed it to be in error, we did not exclude it from the data analysis.

Due to the sampling of sites with multiple observations in each river and some closer or further apart geographically, there might be an issue with a violation of the independence assumption as some sites might be more similar than others even after accounting for river and elevation information. If river is not included in the model, the repeated measures on a river would create a clear violation of the independence assumption. Because samples were taken sequentially in time in the study years, there could be an additional violation of independence by some observations being taken closer in time and others taken later. The results may be biased because the sampling locations in the rivers were not randomly selected and easy-to-access sites may have been selected and they might have systematically higher or lower pCO_2 on average than the population of sites.

III. Summary of Statistical Findings

The final estimated model was: $\hat{\mu}\{log.pCO2 \mid River, Elevation\} = 7.788 - 0.000228Elevation - 0.615I_{River=YT} - 0.195I_{River=YL}$, where River is a three-level categorical variable represented by the indicator variables $I_{River=YT}$ (which takes on a value of 1 if the River is YT, and 0 if not), and $I_{River=YL}$ (which takes on a value of 1 if the River is YL, and 0 if not). This means that the third level of River, YZ, is treated as the reference.

A Type II F-test was generated to assess including River in the model. Accounting for elevation, there is very strong evidence against the null hypothesis of no difference in the true mean log.pCO_2 for all three rivers (F(2, 26)= 6.75, p-value = 0.0044), so we would conclude that there is some difference in mean log.pCO_2 across the rivers. Accounting for the river, there is moderate evidence against the null hypothesis that elevation is not linearly related to log.pCO_2 pressure (2-sided t-test, t(26)= -2.19, p-value = 0.0381), so we would also conclude that there is a linear relationship between elevation and pressure after accounting for rivers. The model has an R-squared of XXX, which suggests that a model with … and … explains ……., which suggests that this model is …….. [Q3]

For two otherwise similar locations that differ in elevation by 1 m in elevation, the median pCO_2 pressure of the higher elevation location is 0.99977 times as much as the lower elevation, controlling for river (95 % CI: 0.99956 to 0.99998). Controlling for elevation, the median pCO_2 pressure in river YT is 0.54 times as much as YZ (95 % CI: 0.368 to 0.795) and the median pCO_2 pressure in river YL is 0.82 times as much as YZ river (95% CI: 0.53 to 1.27). To visualize the impact

of both River and Elevation on the response log.pCO_2, effects plots for the additive model with partial residuals are displayed in Figure 7.

IV. Scope of Inference

To be completed with assistance from generative AI based on final selected model and study details [Q4]

References:

[Q5 – complete remainder of references section, remember – alphabetical order and consistent citation style.]

Greenwood, M. (2024) catstats2: Upper Level Statistics for Montana State University Bobcats. R package version 0.2.

R Core Team (2024) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Figures:

[Q6 – make figure 1 and insert here]

Figure 1. Enhanced stripchart of pCO2 by river.

[Q7 – make figure 2 and insert here]

Figure 2. Scatterplot of …

[Q8 – make figure 3 and insert here]

Figure 3

[Q9 - make figure 4 and insert here]

Figure 4

[Q10- make figure 5 and insert here]

Figure 5

[Q11 - make figure 6 and insert here]

Figure 6

[Q12 - make figure 7 and insert here]

Figure 7

Tables:

[Q13 – make table 1 and insert here]

Table 1. Table of …