Data Analysis Project, Report

BIOST 311: Regression Methods in the Health Sciences

Nina Galanter

Spring Quarter 2023; Submit on Canvas by 19 May 2023 (Draft) and 8 June 2023 (Final)

# Instructions

A report detailing your conclusions is an integral part of any statistical analysis. You must be able to describe your conclusions, along with the methods that you used to address your questions, to your collaborators.

This assignment is limited to **ten pages (including tables and figures)**, and your written text should use at least 12-point font. See the guidelines below for a list of components that need to be included.

Submit your *First Draft* on Canvas by 11:59pm on May 19, and your *Final Report* by 11:59pm on June 8. **Do not discuss your project with other students or post questions about your project on the Canvas discussion board.** But, feel free to ask us any questions in person/via Canvas instead!

This assignment is an important part of your grade in the course. Your first draft will be graded on the same rubric as your final draft. You will receive instructor and peer feedback on your first draft.

**The following checklist is not meant to serve as a list of sections the report should have or the order of contents**

# Checklist

## Scientific Context

### Background and Motivation

*Provide a brief description of your study to give perspective into the importance of the problem and the situation under study. In particular, why were the data collected?*

Your background/introduction should include…

* 1-2 paragraphs of scientific background
  + Set the stage for why this is an interesting overall scientific question to study
  + Why is this an important/relevant outcome to look at?
  + Why is this an important/relevant predictor/exposure to look at?
  + What other variables might be related to your outcome and/or predictor?

NOTE: You can re-use what you wrote for your *Statistical Analysis Plan,* but make sure to address any of the comments we gave you, and you may need to add additional detail to address the criteria above

### Scientific Questions

*You should lay out the scientific agenda of your analysis. In particular, you should state both the overall scientific goal and the specific scientific objectives.*

For this portion of the assignment, you should include…

* your overall scientific goal
* at least two specific scientific questions

NOTE: You can re-use what you wrote for your *Statistical Analysis Plan,* but make sure to address any of the comments we gave you.

### Data Description

*How were the data collected? How many subjects are in the study? Include all necessary information for interpreting your analysis.*

For this portion of the assignment, you should include…

* Where did you get your data (provide enough detail that someone else could go out and find it)
* who is being studied
* how many participants are in the study (i.e., sample size)
* what type of study is this (observational vs experimental; if observational, what type)
* where was this study conducted, where are the subjects from
* when was this study conducted
* how were data collected
* what every variable in your analysis means
  + how they were measured
  + their units

NOTE: You can re-use what you wrote for your *Statistical Analysis Plan,* but make sure to address any of the comments we gave you.

## Statistical Analysis

### Statistical Questions

*Connect the scientific questions listed above with statistical questions that you can test using your statistical tools. Make sure to list all variables that you will use in each analysis.*

This portion of the assignment, you should take each of your (2+) specific scientific questions and re-phrase them as “statistical questions” here. In particular, your statistical questions should be more precise versions of your scientific questions, which specify:

* what is the outcome variable
* what is the predictor of interest
* what other variables will you adjust for
* how are you defining association (e.g., difference in means via linear regression, ratio of odds via logisitic regression)

NOTE: You can re-use what you wrote for your *Statistical Analysis Plan,* but make sure to address any of the comments we gave you.

### Methods Description

*Describe the inferential analyses you will run (e.g., linear regression, logisitic regression), along with the estimates (e.g., which coefficients) and any other relevant values (e.g., confidence intervals) that you will interpret and report.*

For this portion of the assignment, you should include…

* a description of the model you will use to address each of your statistical questions
* for each model: what type of regression, what is your outcome, what is your predictor of interest, what variables are you adjusting for
* what numbers will you report/interpret; if this includes a p-value, what are your null and alternative hypotheses and what p-value threshold will you use for calling something “statistically significant”

NOTE: You can re-use what you wrote for your *Statistical Analysis Plan,* but make sure to address any of the comments we gave you. You will probably need to add additional details to address the criteria above.

### R code

*Upload your code in a separate file.*

For portion of the assignment, you should include…

* a or file with the code you used to complete your analyses

NOTE: This should **not** be a section in your report, but you must upload the code to Canvas along with your report.

### Appropriateness

Statistical methods should be appropriate in that

* They estimate the specific definition of association used in your statistical question (mean difference for linear regression, odds ratio for logistic regression, etc.)
* They include the outcome and exposure used in your scientific and/or statistical question
* Variables play the correct role in the analysis: the exposure is predicting the outcome and not the other way around for example
* They are appropriate for the type of variable (e.g. don’t use logistic regression for a continuous variable or a log transformation for a variable that can be negative)

## Results and Interpretation

### Descriptive analyses

*Actually run the descriptive analyses that you proposed in your Statistical Analysis Plan, and report the results in tables or figures, as appropriate.*

To receive full credit for this portion of the assignment, you should include…

* a table with summary statistics that describes your study population; the table should have:
  + an informative caption
  + meaningful row and column names
  + units
* an interpretation of interesting or informative differences between groups

### Inferential analyses

*Report and interpret the point estimates, confidence intervals, and p-values corresponding to each scientific question that you address.*

To receive full credit for this portion of the assignment, you should include…

* a full paragraph interpretation of results for each regression analysis, tied to each statistical question, including:
  + reporting and interpretation of the estimate, confidence interval, and p-value
  + a conclusion tying these results back to the statistical question

## Assumptions and Discussion

### Assumptions

* consider whether each assumption required for your analysis is met
  + e.g. LINE for linear regression, independence and sample size for logistic regression
* support discussion of assumptions with graphs as needed
* any graphs should be clear and well labelled

### Discussion

*Discuss what the statistical results you presented in the previous section mean in terms of the scientific questions that you posed at the beginning of the report.*

To receive full credit for this portion of the assignment, you should include…

* a 1–2 paragraph summary of the results of your analyses, in the context of your overall scientific goal (do not simply repeat your paragraphs from the *Results* section)
* a discussion of the limitations of your data or analysis, e.g.,
  + are there potential confounding variables that were missing from the data?
  + is the sample representative
  + are there any concerns about measurement error, events being underreported, or people responding untruthfully
  + unmet or potentially unmet assumptions
* a discussion of the following questions: Are you surprised by what you found? How do your results fit in with the larger scientific literature (do not do a full literature review, but interpret your results in the context of your understanding of the problem)?

## Writing and Organization

### Writing

* check for typos (use a grammar checking software such as word; the free version of grammarly is also good for this)
* take a break after you are finished writing your paper (or a section of it) and read it later to check if it is clear
* avoid passive voice - YOU CAN USE I (e.g. say “I retrieved the data from…” as opposed to “The data were retrieved from..”)
* vary your sentence structure and word choices
* be concise when possible (but don’t sacrifice correct interpretation even though it can be wordy)

### Organization

* organize your paper into sections
* a few ideas:
  + Introduction, Methods, Results, Discussion
  + Introduction, Dataset Background, Methods, Results, Discussion
  + Introduction, Methods, Results, Assumptions, Discussion

### Title

*Make up a title for your proposed study!*

Your title should…

* be interesting and informative (e.g., allude to your scientific goal)
* not contain any abbreviations

### Formatting

For this portion of the assignment, you should…

* use 12 point font
* write no more than ten pages (including your tables and figures)
* write in full sentences and paragraphs

You should not…

* include raw R code or output