Minute Paper after Class 07 Feedback

Thomas E. Love, Ph.D.

2023-02-09

Table of contents

1 Sample

n = 51 of 56 students responded by the 2023-02-08 deadline. Thanks!

2 Which of the following tasks have you accomplished in the development of your Project A plan?

Available responses (check all that apply: n = 51 respondents) were:

Students	Task
44	I've read the Project A instructions.
34	I've decided to work on this alone, rather than with a partner.
15	I've agreed to work with someone as my Project A partner.
26	I've identified a data set I will use for Project A.
15	I've started to work completing the Quarto template for Project A.
12	I've successfully loaded my data into R.
4	I've cleaned my data in R and now have a tidy tibble of my information.
2	I've built a codebook for my data in Quarto.
6	I've identified an outcome and predictors for my linear regression model.
5	I've identified an outcome and predictors for my logistic regression model.
1	I've completed the Project A plan and submitted it to Canvas.

3 Most important thing you've learned recently in 432?

(*edited and grouped by TEL*)

- Adding non-linear terms to a linear model
  - interaction terms, polynomials, and restricted cubic splines
  - the importance of spending degrees of freedom wisely
  - Spearman's ρ^2 and potential predictive punch
  - Doing some things with `ols()` and others with `lm()`
- Logistic regression models
  - Fitting them with `glm()` and with `lrm()`
  - Understanding the fundamentals of logit, odds ratios, etc.
  - The logistic link function
- Using multiple imputation within a regression model
  - via `aregImpute()`
  - via `mice`
  - Using single imputations for some modeling work
- Evaluating and Comparing Data that are Publicly Available

- Validation in regression via `validate()` and via splitting data into testing and training samples
- Survey weights
- The value of attending TA office hours!
- The Linear Probability Model

4 Your Questions: My Answers

I edit these questions a bit for clarity, and answer some, but not all that are asked. If I didn't answer your question, you are welcome to try again, perhaps on Campuswire, in person or during TA office hours, or just with a rephrased version in the next Minute Paper. Those of you without questions this week, try not to make a habit of never asking a question here, but don't worry about not asking anything if you really don't have any questions.

I also got to fewer questions than usual this week. Sorry about that.

4.1 About Statistics / Data Science Issues

- Can you explain more about the importance of order of predictors to models?
    - Nearly all regression output produces identical results regardless of the order in which you insert predictors into a model. A notable exception to this rule is the sequential ANOVA table, which shows p values describing hypothesis tests about the variables in order (so the first one you enter is tested first, then the second assuming the first is already in the model, and so forth.)
- I have read references that state that the number of events (outcomes) per variable in a multivariable linear regression model can be as little as 10-20.
    - Please share that reference (or references) with me. I would be eager to see them, so I can react to them in their proper context. I also don't really know what you mean by outcomes per variable.
- For a manuscript, how much should you show of the models that you create in the linear and logistic regression.
    - It would depend mightily on the manuscript, the specific questions I was trying to address, the types of models fit, and who it was aimed at.
- How can imputation add any new information to our dataset/analysis?
    - Who suggested that it did? The choice is not between imputing and not imputing. The choice is between using only the data that is complete, or also (in a careful way) being able to use the data that was gathered but is not complete.

- How do we know which regression model to use, linear or logistic?
  - Logistic models are for binary outcomes, exclusively. Linear models are for quantitative outcomes.
- Is there a rule of thumb that can be used when deciding between using a polynomial term and a restricted cubic spline term?
  - Um, always use a spline?
- What is a spline and what are knots mathematically? As in I know the code and when to apply it, but I don't actually know what the math behind it is?
  - Check out https://www.nature.com/articles/s41409-019-0679-x and Frank Harrell's Regression Modeling Strategies book from our Sources page.
- Someone asked questions about several terms that are not discussed in this course, including: "hazard-function multiphase parametric modeling", "competing risk methodology" and "bootstrap bagging".
  - I don't really have time or inclination at the moment to try to address things that are way out of the scope of this class. Doesn't mean I never will, but not soon. Sorry.

4.2 About Project A

- In case we are working on surveys, how do the weights take into account the follow up variations of the subjects?
  - If you are using survey data in Project A, **do not** use the weights.
- Does it matter if project partners are different (one undergraduate and one post-graduate)?
  - No. Anyone enrolled in the class can partner with anyone else enrolled in the class.
- Can I merge the BRFSS and CHR data for my project A?
  - That would require you to have BRFSS data at the county level.
- Do you have a suggestion of some other publicly available datasets other than NHANES?
  - I made many such suggestions in the Project A instructions.

4.3 About Other Aspects of the 432 Course

- If you'd like to generally review content from a class or two, are TA office hours the best places to do that?
  - Not really, no. They're better equipped to handle specific questions. I suggest you review something on your own, perhaps trying to replicate something I did in new data, to generate such questions.

4.4 About R/RStudio/Quarto/Coding

- Do binary categorical variables have to be encoded as 0/1 for glm/lrm to work properly, or can they still be encoded as factors with more meaningful names?
  - It depends on their role. As an outcome, I usually recode as 0-1 because it helps me interpret the result correctly. If the binary variable is a predictor, it doesn't matter.
- For similar functions that effectively do the same (ex. glm() and lrm() for logistic regression) is there anything besides the differences in summary results that you use as deciding factors for which to use?
  - Again, it is a false choice. Use both is always the best option.
- I don't understand why `geom_jitter()` was used in o-ring example on 2023-02-07.
  - Solely to make the visualization clearer: specifically that there were several temperature readings which saw multiple shuttle launches, which was hidden when we used `geom_point()`.
- What does `#| echo: true` do in Quarto?
  - It means include the source code in your HTML output. You should always be doing that in anything you send to me, but you also don't need to specify it when you're building an HTML.
  - When I build the PDF slides, Quarto defaults to `echo: false`, so I have to tell it to show the code when I want you to see it. This is one reason I preferred building HTML slides, but the PDF ones are clearly more appealing to some students.
- What is the quiz format?
  - You'll see the instructions next week. It's a set of 25-30 questions in a PDF which you'll answer in a Google Form, like the Minute Papers. I'll give you several data sets to support the work. It is open book, open notes, although you can only discuss it with Dr. Love and the TAs, and only in specific ways. You'll have four days to complete it.

4.5 Miscellaneous Questions

Lots of people thanked me, or the TAs. We really appreciate that.

- Any plays coming up?
  - Come to "The Play That Goes Wrong" at https://www.auroracommunitytheatre.com/ this weekend or next - we close February 18. It is the most enjoyable play (for an audience, and for me) that I've been in here in Ohio, and it will most likely be my only time on stage in 2023.
- How's the show going thus far?
  - Incredibly well. I am delighted.
- How did you feel about your first play and what was the play?
  - I had been onstage in a few different revues as a kid, and had tiny parts in things my parents were also involved in. My first real role was as Oliver in the musical Oliver. I was in sixth grade. It was a lot of fun.
- How did your toothache go?
  - Thanks for asking. It's a nightmare. I have a recent implant that will need to be removed so that they can make a plan to see what they need to do (root canal being the best of the options) with the molar behind it where the damage has occurred. I am on Advil every waking moment right now, and this will likely remain an issue for at least six more weeks.