# Minute Paper after Class 11 Feedback

Thomas E. Love, Ph.D.

2023-02-22

## Table of contents

## 1 Sample

n = 56 of 56 students eventually responded. Thanks!

## 2 Most important thing you've learned recently in 432?

(*edited and grouped by TEL*)

- Fitting and evaluating logistic regression models
  - Describing Effect Sizes well using both `glm()` and `lrm()` fits
    - Fit your models with both `glm()` and `lrm()`
  - Plotting and summarizing effects in terms of odds ratios
  - Assessing discrimination with the C statistic and ROC curve
  - Assessing goodness of fit with Nagelkerke $R^2$, Brier score
  - Bootstrap validation of summary measures
  - In-sample comparisons of AIC and BIC across models for the same outcome
  - Assessing calibration with a calibration plot
- Adding non-linear terms to regression models
  - Spearman $\rho^2$ plots
  - Using and interpreting interactions
  - Using and interpreting splines
  - Plotting effects with ggplot(Predict) and plot(summary)
  - Using and creating nomograms
- Linear models with `lm()` and `ols()`
- Feature selection
  - The problems with stepwise procedures
  - The lasso and ridge regression approaches
- Imputation
  - Single and Multiple Imputation approaches
  - Sometimes, it doesn't make much of a difference what you do
- I have learned the importance of having clear code that people can follow and to add additional comments
- The challenge of finding relevant publicly available data
- How helpful it is to work ahead and start labs and projects early.
- Paying attention to details in a project
- Steps for cleaning a large data set

## 3 How confident do you feel about doing well on Quiz 1?

Scale is 1 = Not Confident at all to 5 = Very Confident

| Score | 1 | 2 | 3 | 4 | 5 | Mean | SD |
|---:|---:|---:|---:|---:|---:|---:|---:|
| Last Week | 1 | 6 | 25 | 21 | 2 | 3.31 | 0.79 |
| **This Week** | 1 | 10 | 23 | 18 | 4 | 3.25 | 0.90 |

## 4 How confident are you about successfully doing Project A?

Scale is 1 = Not Confident at all to 5 = Very Confident

| Score | 1 | 2 | 3 | 4 | 5 | Mean | SD |
|---:|---:|---:|---:|---:|---:|---:|---:|
| Last Week | 0 | 4 | 5 | 15 | 31 | 4.33 | 0.92 |
| **This Week** | 0 | 5 | 12 | 21 | 18 | 3.93 | 0.95 |

Perhaps the reactions to the initial drafts of the Project A Plans scared people a bit.

## 5 Would you prefer that we continue to require masks in class for 432?

| Response | Students (%) |
|---:|---:|
| Yes, please continue requiring masks for all. | 15 (27%) |
| I don't have a strong opinion one way or the other. | 27 (48%) |
| No - please stop requiring masks for all. | 14 (25%) |

I can't interpret this as a mandate to make a change, so we'll keep requiring you to wear masks, at least for the foreseeable future.

- I did get one impassioned comment from a student essentially begging me to not wear a mask. Perhaps I should have asked about my masking separately, but I'll stay with it for now.

# 6 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.*

This week, I answered almost all of the questions that I understood, or that I thought I could answer. If I didn't answer your question, I probably didn't understand it, but it's also possible that you asked about something where I don't know enough to help, or couldn't come up with a sufficiently brief answer for this forum. Sorry.

## 6.1 About Statistics / Data Science Issues

- When we add restricted cubic splines or quadratic functions in a linear regression model, why do we assess the residual plots the same as before? since we fit non-linear terms to our data, why would we assume linearity (or normality) from our residuals?
  - Because the residuals from our model still need to meet the same assumptions as before. The use of an interaction term, for example, doesn't change the requirements of the errors. It just changes how the predictors are put together. The need to check residuals didn't change when we transformed outcomes, either, for the same reason.
- Can you clarify again the differences between glm and lrm and when to use them or why to use both?
  - Use both. Read Chapters 20-22 of the Course Notes.
- Can we use the lasso in the context of logistic regression?
  - Yes, although the lars package probably isn't the best choice for doing that. https://bookdown.org/tpinto_home/Regularisation/lasso-regression.html has some relevant material using `glmnet`.
- How would you describe the relationship between statistics and information theory?
  - I wouldn't try. I don't know anything you cannot learn from googling about information theory, and haven't thought seriously about it in the past twenty years. Other statisticians I know use the ideas from information theory every day. To each their own.
- Are there further examples on how to best write your methodology out for linear regression for a manuscript ?

  - This is too broad a question for me to answer, really, other than to give the obvious response that you should be looking to existing manuscripts in your field to determine the state of the art.

- Are there real world instances where you have a continuous outcome where you consider making it binary to use a logistic regression model?
  - Only if a cutoff point is of critical scientific interest and agreed to by all involved as being more important than the value itself. And even then, it's probably a bad idea because of the damage it does to your statistical power.
- In 431 and again in class today, we learned about the drawbacks to stepwise regression procedures. Are there any situations in which you would recommend using this tool or is it just better to be avoided in almost all cases?
  - I can't provide an example where it's the best choice.
- In a hypothetical dataset with a few variables, what percentage of missing values of a variable would be your cutoff for you to discard the variable as a potential predictor? Is disregarding a variable because of how many values it's missing a good practice?
  - This is too hypothetical for me. In general, no, it's not good practice to drop variables because you're missing data, but sometimes all you can learn about something is that it's missing in lots of people and you might want to try to figure out why that's the case. If you know why, the answer to whether to include it or not might be very different depending on that reason.
- Is a log transformation recommended for right-skewed data?
  - Sure. It's not the *only* transformation that helps in that setting, and sometimes it doesn't help enough, but it's the most common choice.
- Is there regression models for multicategorical outputs?
  - Sure, and we'll start learning about them in Class 17, according to our Calendar.
- When should you use single imputation and multiple imputation, instead of just doing a complete cases analysis?
  - When you're not willing to assume the missing data are missing completely at random. See section 7 of the Course Notes.
- When using imputation in a scientific article, does your table 1 have the information on the original data set or on your imputed data set, or both?
  - The original data set.
- When describing the effect estimate for a continuous predictor in a binary logistic regression model, do you think it's preferable to describe the effect as a one unit increase in the predictor (such the odds ratio value from a glm model) or as an increase from the

25th percentile to the 75th percentile of the predictor (such as the odds ratio value from an lrm model)?

  - No. That is, I don't feel it's universally preferable to look at this in only one way. I feel it's important to be able to do either in practice.

- When we have large data sets that we need to clean for some sort of consultation, how do we ensure we are cleaning the data correctly? Should we check with the client every step of the way?
  - Yes, certainly.
- When you impute, do you have to have less than 5% missing values?
  - No, why would you think that?
- Do i always have to make sure my outcome is coded 0/1 in R for logistic regression model? and what is the best way to get R to show the value labels of a variable
  - As opposed to 1/2, yes.
  - As opposed to using a factor - no, but it's easy to get the factor flipped around and be predicting Pr(not X) instead of Pr(X), so I usually use 0/1 to avoid this problem.

### 6.2 About The 432 Course

- Is the Quiz similar to the 431 Quizzes in terms of format?
  - A complete draft of instructions for the Quiz has been available to you for some time. I encourage you to read it to get a sense of the answer to this question.
- I know the quizzes change year to year, but based on data you've probably collected in prior years, about how long does it take people to complete quiz one?
  - This is answered in the instructions to the Quiz.
- Are there extra labs to help with the final grades?
  - There will be one, and it will become available March 1.
  - There are many many ways to improve your grade - an obvious option is participating on Campuswire.
- I noticed you reviewed five project A plans and accepted them all, so I'm wondering if the TAs have more stringent standards for reviewing the projects.
  - No, I reviewed more than 20 of the initial versions of Project A plans, and accepted five of those. The others I stopped reviewing as soon as I couldn't accept them. The TAs are undoubtedly more strict on some things and less strict on others. That's why I make all of the final decisions to approve a project.

- The class has been quite fast paced so far. Will it continue to be so until the end of the semester?
  - I'm not anticipating a big change in speed.
- When the project A plan gets graded, does every revision get a separate grade that's put together or the most updated version will get a grade?
  - The current (most up to date) version is the one that is graded, each time. I don't go back to prior versions of your work, although I do check that you've done what we asked you to do in revisions. Grades on the project plan are 20 points for successful and on-time first versions, 18 points for successful and on-time second versions, and 16 points for versions 3 and later.

## 6.3 About R/RStudio/Quarto/Coding

- Are R packages verified by an entity or are their performance and validity checked by the community? In other words, how much can we trust R packages?
  - Trust them to do what, exactly? There are tens of thousands of R packages just on CRAN alone - while CRAN does necessitate the passing of certain milestones and test the packages daily on multiple systems, some of those packages are more well-verified than others. Here's a primer on CRAN - https://kbroman.org/pkg_primer/pages/cran.html
- Do you know of R packages for generating plots of points on maps of the body? I have a dataset of acupuncture points and am looking to summarize them visually.
  - I do not know of such a thing, and haven't ever looked. Good luck.
- Have you written any books? Did you use R or quarto?
  - Do you consider the Course Notes, written in Quarto, a book?
- What is the reason that we need to do `set.seed()` more than once? I feel like it would make sense to set the seed at the beginning of the code chunk and that the seed would stay, but on the latest lab I found that I needed to set the seed multiple times within the code block or my answer would change.
  - If you run multiple calls to random numbers within a single code block, why would you expect the machine to produce the same set of random numbers if you just set a seed at the start as it would if you set the seed again each time you wanted new numbers?
- What did `scale()` do again in your ridge regression code?
  - Center each observation within each variable by subtracting the mean of the variable, and dividing by the standard deviation of the variable, across all subjects in the sample.

### 6.4 Other Questions and Comments

Several people thanked me, the TAs or both for their help. Thank you!

- What course(s) (CWRU, online, etc.) do you recommend to learn advanced theoretical statistics?
    - I don't give global recommendations on courses, since I haven't taken any recent versions of any courses offered at CWRU or online. My department gives several courses on theoretical statistical methods. I don't know how relevant they would be to your particular needs.
- Following up on above: 432 is an excellent applied data science course, but I find myself wondering about the theoretical underpinnings of many of the topics we discuss in class.
    - Great. I'm succeeding in turning you into a future Ph.D. student.
- What techniques do you use to help you focus on days you feel like your being unproductive?
    - I make short lists using Workflowy and try to get things done by reducing them to smaller tasks. I have used other "tricks" in the past but that's the one that sticks. I also try to take a walk (even just for five minutes) when I get stuck or lazy.
- Do you prefer to perform in plays or musicals?
    - Yes. It all depends on the role I'm playing, where I'm doing the show, and with whom. But I prefer performing to being backstage or in the audience.