Minute Paper after Class 09 Feedback

Thomas E. Love, Ph.D.

2023-02-15

Table of contents

1 Sample

n = 55 of 56 students responded by the 2023-02-15 deadline. Thanks!

2 How Are You Feeling?

People feel **tired**. The most common response we saw was that people are tired. That's not surprising - we are now in the most stressful time of the semester in this course (which is why I asked) and the stress probably won't completely let up for many folks until the Project A plans are all approved and Quiz 1 is over. I am aware that Spring Break cannot come soon enough for some of you, but others have issues that a break alone will not resolve.

Other words/phrases I saw in the responses included:

- (a bit) overwhelmed, stressed out, sad, underpaid, burnt out
- (a little) nervous, anxious, not too bad, not confident, behind
- OK, all right, fine, gaining control, trying hard
- less stressed, pretty good, better, confident, engaged, glad
- well, excited, good, happy

I did see many more of the words listed in the last two bullets above than in the top two, but there was greater variation in the negative comments. If we can help you, let us know.

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

(*edited and grouped by TEL*)

- Logistic Regression
  - Assessing classification with confusion matrices and with area under the ROC
  - Using both `glm()` and `lrm()` to fit models
  - Obtaining predictions on the probability scale and on the linear scale (log odds)
  - Comparing two logistic regression models with in-sample comparisons of AIC, BIC, Nagelkerke R^2 and the C statistic
  - Bootstrap validating the Nagelkerke R^2 and Somers' d (and converting to C)
  - Plotting effects on the odds ratio scale with `plot(summary(modname))`
  - Interpreting effect sizes (see Chapter 22 of the Course Notes)
  - Checking calibration with `plot(calibrate)`
- Nomograms: building them and using them
- Spearman ρ^2: what it does, and what it doesn't do
  - Building non-linearity through interactions, restricted cubic splines, and (maybe) polynomials
  - Spending degrees of freedom wisely
- Single and multiple imputation
- How to formulate and execute a research question

- Coding is simple; however, interpreting and explaining to others is more difficult.
- Choosing variables for a regression can be difficult
- Cleaning strings is very annoying and datasets with over 9 million rows make my computer cry
- Learned about a lot of new databases.

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.

This week I actually think I answered all of the Questions that were posed, either below or in an email, but perhaps I missed something. So I'll leave the usual caveat (above) up.

4.1 About Statistics / Data Science Issues

- Would you have any recommendations/suggestions for a course to supplement/advance what we have learnt so far for the Fall semester?
  - I live in a department of Population and Quantitative Health Sciences, which has many courses on offer each semester, including my Observational Studies course (PQHS/CRSP 500) every Spring. There's no way you'll catch me not recommending any of our courses, but I would suggest that talking to other students who've taken those courses (like our TAs) is probably more helpful than talking to me, who hasn't taken any of them.
- After running a Spearman's rho-squared plot and seeing which variable(s) we may want to consider using non-linear terms with, is a check for collinearity appropriate for determining whether we should include an interaction term?
  - No. The whole point of doing the Spearman plot is to avoid looking at things that will bias your conclusions. Including an interaction term (or not) in a clinical model should be based on (1) your understanding of the clinical issues involved and/or (2) the suggestions of the plot. (1) is far more important in practical work.
- If I am looking to see if there is any variable has something to do with the outcome, do I need to build multiple models or is one model good enough?
  - I'm not sure why you'd be looking for that by building a multivariable regression model, but no, one model is (essentially) never enough for exploratory work of this type.

- I am still not very clear on how we decide what model we should use when we have a data set to analyze. Will you give a lecture summarizing all the models together?
  - Well, so far, it's just linear models (quantitative outcomes) and logistic models (binary outcomes), right? We'll see lots of other types of models this term, but it's not hard to identify what type of outcome you have, I think, between these two options.
- If a model has a "low" r-squared, does that reflect at all on the hypothesis?
  - Sure. It suggests that a hypothesis that a strong relationship would be found between this outcome and these predictors in these data isn't well-supported. But the problem could be with the hypothesis, the choice of outcome measure, the choice of how to measure the predictors, the sample you obtained and lots of other things, too.
- Is it possible to add non-linearity to categorical variables?
  - That's what interactions do. Other than interactions, no.
- Some mentors have stressed the use of "multivariable" over "multivariate" as the correct term when discussing regression models with a number of predictors for one outcome. I have observed others using these terms interchangeably because most of the time the audience will infer the correct information. Which side of this "debate" do you fall on?
  - My understanding from Google's English dictionary and Merriam-Webster is that this is a distinction without a meaningful difference *in common parlance*. "Multivariate" refers to a situation involving two or more variable quantities. The definition of "multivariable" is the same, and uses multivariate as a synonym.
  - But *in statistical work*, multivariate methods are those with more than one dependent variable (outcome) or methods that do not pre-specify outcomes and predictors, while multivariable methods have one dependent variable (outcome) and more than one independent variable. That said, I have probably been sloppy about making this distinction.

I got several questions about the general topic of degrees of freedom, which I've gathered below…

- I am a little unclear about the choice of number of variables with regards to degrees of freedom. Since categorical variables uses more d.f, when we determine the max number of variables do we count levels for categorical variables or the categorical variable counts as just one?
  - Levels.
- How to decide appropriate DF or predictor variables for a logistic regression to not overfit model.

  - So far, I've given you specific instructions on how many degrees of freedom to consider adding to a model. Stick to those ideas in work for the class.
  - More generally - this is a topic we haven't yet really discussed, but we will have more on this later this term.

- The concept of degrees of freedom is slightly vague to me. I do understand the effect of predictors and non-linearity on it but do not understand the reason behind presenting it. I just think of it to be analogous to power in models. Am I right?

  - It's the best way to summarize how much information you're obtaining from the data. It's not really the same thing as statistical power.

4.2 About Project A

- Why do you want us to choose the data set instead you give us one? It takes some time to select the data set.

  - You've sort of answered your own question there. My goal is not just to have you analyze something - you need to go through the whole process of creating a project from nothing. If I give you the data, that eliminates important work on your end - skills that I want you to develop.

- How early should we start working on the analysis part of the project?

  - As soon as I approve your Project A plan.

- Is it OK if we make small deviations from the Project A plan after we actually do the analyses?

  - Yes, but you should probably do this less than you're anticipating. For instance, I wouldn't materially change the plan just because a model had a poor R^2 value.

- Is it possible to change our predictor variables for Project A if in the process of doing additional data analysis we find an interesting variable we'd like to explore more?

  - Sure, so long as your set of candidate variables meets our requirements for the project.

- In the project A proposal, you ask us to use formulas to determine the maximum number of predictors in your two models. Why is this important?

  - It's a way for me to ensure that people fit relatively small "main effects" models that have some chance of validating reasonably well. I make no claims that this approach will work well outside the confines of this Project.

4.3 About The 432 Course (Other Than Project A)

- Will we learn Kaplan Meier curve creation in 432?
  - Yes, starting in Class 14. See the Calendar.
- Will we work with ordered multinomial logistic regressions in 432?
  - Yes, in Classes 17-20. See the Calendar.
- Are we going to work on hierarchical data? Are we also going to learn about k-means clustering?
  - Not in any serious way, no.
- Is the course notes for 431 and 432 available as pdf?
  - I've answered this question before in previous Minute Papers. For 431, yes - just click on the Adobe logo on the top left of the document. For 432, not before May, and maybe not even then, because producing a PDF (as opposed to HTML) is an incredibly intensive activity, because several of the things I'm doing (most especially equatiomatic) work only with HTML at the moment, and specifically not with a PDF book.
- How do I apply to be a TA in 431 next Fall?
  - I'll be in touch with everyone who successfully completes 432 in June to tell you how to do this, if you're interested. This is something I work on in the summer. If you're interested in more details, let me know.
- I was wondering if there are other resources that are useful to learn about applied data science that are not explicitly not mentioned in the course.
  - If I find something useful, I usually add it to the Sources page or one of the Class READMEs or both. There's a lot of material out in the world - if you find something that helps you, let us know.
- What is the advice you have for quiz 1?
  - Review the Labs and complete as much of your Project A analyses as possible to prepare. I would also look at the Minute Paper lists of "Most Important" topics and make sure I was familiar with those.
- What do you think will be the most challenging part of quiz 1 for students?
  - Paying attention to little details usually seems to be the biggest problem people mention after seeing the answer sketches to my quizzes. Read the whole question carefully before you try to answer it.
- Is there anything else that can make us lose points from the minute paper, except for not completing it?

  - So far, just completing it late. I may change that plan if I stop getting enough good questions to answer, but that's not the situation now.

4.4 About R/RStudio/Quarto/Coding

- How can I use Quarto to write a book and produce attractive presentations.
  - The Quarto website is where I got started: books here, presentations here.
- Do you have personal preferences or criteria for choosing between two functions that do the same (ex. using lrm() vs glm())?
  - `glm()` and `lrm()` fit the same model, but they don't really do the same things. I use both, all the time. There's no reason not to.
  - More generally, if there's a tidyverse way of doing something, and a non-tidyverse way, I usually choose the tidyverse way. This doesn't apply (yet) to tidymodels stuff other than `rsample` and `broom` for me. I still use `caret` a lot.
- I noticed you used "NULL" within a recode statement of the NHANES data to tell it to code those values as NA rather than writing "NA." Is that usually how you do that?
  - Yes. NA doesn't work in some settings to convince R that the value is missing. NULL does.
- What are the limitations of R? In particular, when it comes to doing programming/statistical and analytic work in professional spaces, when would R be used and when would other software be more appropriate for quantitative work?
  - I have used R for literally everything I have done in terms of statistical programming and analytic work so far in 2023, and I have many colleagues who do, as well, but that doesn't mean there aren't other good tools in the world.
- Is a Quarto doc more or less efficient than an R markdown?
  - No, it isn't meaningfully more or less efficient if you're working in R exclusively.

4.5 Miscellaneous Questions/Comments

To everyone who thought of me as I celebrated my birthday or wished me a happy birthday, I really appreciate it. Thanks for making my day a bit brighter. While I have your attention, I encourage you to go out and give some blood, get rid of some guns, or just be nice to someone who needs some help. Thanks.

Several people also thanked me, or thanked the TAs. We appreciate that very much. I want to express my personal thanks to the TAs for reviewing all of the Project A plans so quickly.

- How was your birthday celebration?
  - My wife is out of town, unfortunately, so it was just me and the cat. I had a pretty quiet evening, but ate well. Probably too well. Thanks for asking.
- I hope your tooth pain is improving!
  - Thanks, I have good and bad days. Nothing ibuprofen can't handle, so far. I am scheduled for a first procedure Thursday the 23rd after class.
- Mustache looks cool, feel free to trim to your comfort, but you could try and keep it.
  - Thanks, but the life of this mustache after the curtain falls on The Play That Goes Wrong can be measured in dozens of minutes, or maybe even less.
- Do you have a favourite area/topic of health/medical research?
  - I guess I don't, really. I don't spend as much time as I used to reading other people's work. That's a downside of being busy.
- I really appreciated seeing the examples of people's visualizations in the Lab 1 answer sketch…. It also helped to read your short descriptions of what you specifically thought made them worth highlighting. I know it is extra work to do those things and you already have a lot on your plate (you give us A LOT of resources already), but would you be willing to continue sharing a few examples from time to time of things students in 432 have done lately that you found particularly clever or useful?
  - Thanks for the feedback. Some of the more recent labs haven't really been suitable for this sort of thing. Project A is the next obvious opportunity, but I won't make any promises just yet.
- How applicable is a degree in biostatistics to jobs that are in applied statistics or data science?
  - Very? I don't really know how to answer this question well. It would completely depend on the job.
- Any advice for how to build a successful career?
  - A successful academic career, or some other kind of career? I'd be happy to chat about this with you as the semester moves forward, but useful general advice is essentially to find what drives you, and then try to craft a job (or these days, a series of jobs) around that.

5 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
Count	1	6	25	21	2	3.31	0.79

6 Are you working with a partner on Project A?

We have 47 projects, including 38 individuals and 9 teams of two.

7 Complete Project A Plan submitted?

At 5 PM on Wednesday 2023-02-15, we have submitted plans for all but one Project A.

8 How satisfied are you with your Project A Plan?

Scale is 1 = Not satisfied at all to 5 = Very Satisfied

Score	1	2	3	4	5	Mean	SD
Count	0	5	15	24	11	3.75	0.89

- The mean rating across the 17 respondents working with a partner was 4.06
- The mean rating across the 38 respondents working alone was 3.61

9 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
Count	0	4	5	15	31	4.33	0.92

- The mean rating across the 17 respondents working with a partner was 4.24
- The mean rating across the 38 respondents working alone was 4.37

It was interesting to me to note that the mean rating for "work so far" was higher with a partner, but for "work to come" it's higher without a partner.

10 Cross-Tabulation of These Last Two Items

- Rating of Satisfaction with your Project A Plan is in the rows.
- Rating of Confidence in completing Project A is in the columns.

–	A = 2	A = 3	A = 4	A = 5	Total
PlanA = 2	2	1	0	2	**5**
PlanA = 3	2	2	8	3	**15**
PlanA = 4	0	2	6	16	**24**
PlanA = 5	0	0	1	10	**11**
Total	**4**	**5**	**15**	**31**	**55**

The Spearman correlation of these responses is 0.527; the Pearson correlation is 0.533.