Minute Paper after Class 05 Feedback

Thomas E. Love, Ph.D.

2023-02-01

Table of contents

1 Sample

n = 54 of 56 students responded by the 2023-02-01 deadline. Thanks!

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

(*edited and grouped by TEL*)

- Linear Regression Models
  - Adding non-linear terms
    - Restricted cubic splines
    - Polynomial terms (raw and orthogonal)
  - New tools for fitting and evaluating linear models
    - Using `ols` to fit a linear model
    - Using `validate` to obtain bootstrap-validated R^2 and RMSE
    - Building and reading a nomogram
    - Using Spearman's ρ^2 plots to help suggest non-linear terms
  - Analysis of Variance with and without interaction
    - Interaction plots
  - Analysis of Covariance (adding a quantitative covariate to an ANOVA)
- Dealing with Survey Weights
- Splitting a quantitative variable into categories can lose a lot of information
- Things from "How To Be A Modern Scientist"
  - I'll save these for now, since not everyone's read the book yet.

3 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.

3.1 About Statistics / Data Science Issues

- Can you point us towards a textbook or online resource that would break down the last few class topics in more basic detail? (splines, ols, etc.)
  - I did this in the Class 05 README.
  - There are lots of resources on our Sources page as well.
- I'm trying to understand how we validate summary statistics of an ols fit model.
  - If the examples in class today and Tuesday are insufficient, I suggest you look at the Course Notes, especially this section on model validation and correcting for optimism in an `ols` fit.
  - You may also want to look at some of the Harrell-verse resources provided in the Class 05 README.
- How frequently do you use splines in your real work?
  - Well, teaching is real work, I promise.
  - In research and quality improvement work, I don't know - perhaps 15 or 20 times per year. But not all of those models wind up getting published - only a tiny fraction, in fact.
- Do you have a favorite reference to explain limitations of using p-values
  - The first two Key Articles on our Sources page: specifically the 2019 and 2016 ASA reports.
- Over winter break, I was finishing up a manuscript with my old PI. They really wanted to include phrasing about statistical significance, and I tried to gently provide some of the points you gave us last semester. It was only minimally effective. Do you have any suggestions for communicating the problems with p-values when someone outranks you and/or is stuck in their ways?
  - Try to be the change you want to see in the world, but sometimes it's best to go along to get along. Some fights aren't worth pursuing sometimes.
- How do you determine a cut-off for a "good" model?

  - All models are somewhat incorrect. Some are more useful than others.
  - I encourage you to stop looking for an easy distinction between a "good" model and a "bad" one. This is over-simplifying a complex series of challenges involved in assessing the quality of a model.

- Is there a way to specify what the comparison is within an ols model summary rather than just using the distance between the .25 and .75 quartiles?

  - That's the main approach that `summary` in an `ols` fit is designed for. It is probably possible to look at other things, but I'm not sure I ever have with those plots.

- What is the difference between the calibration plots shown in class 5 and the residual plots?

  - A calibration plot has nothing to do with regression residuals. So they have nothing to do with each other.
  - A calibration plot helps us determine whether the predictions made by the model match up well with observed data, and uses bootstrap validation of predictions to help us do this.
  - Residual plots help us assess regression assumptions like linearity, constant variance and Normality.

- What is the interpretation of R^2 values when we have categorical variables in our model. Am I correct to assume R^2 value is most impacted by and is most useful in evaluating relationship between two continuous variables?

  - Not really, no. If your outcome is quantitative, then R^2 is pretty much equally useful regardless of what types of predictors are included in your model. In each case, it can be interpreted as the proportion of variation in the outcome accounted for by the model, or as the square of the correlation between predicted and observed values of the outcome.
  - Here's a tweet from Tom Carpenter with some related explanations

- Is there a way to change the labels on plot legends without recoding the plotted variables?

  - That's the way I usually do it in ggplot2 plots, but I don't know that it's the best solution in all cases.

- When making the enormous assumption of MCAR with regards to missing data, are there any graphical representations of data and/or statistical tests that can back up this assumption when proceeding with a complete case analysis?

  - Well, no. Plots of the non-missing data can't tell you anything helpful about the missing data. MCAR is an assumption based on how the data were gathered, rather than an assumption based on what the non-missing data look like. So what is one to do? Well, if you're going to do a complete-case analysis, the most important thing is to make it clear that's what you're doing, and describe the amount of missingness you are dropping.

- Why do we treat survey weights numbers to decimal points if they represent people? Shouldn't we be using whole numbers?
  - No. If your survey sampled 100 people in class A and 50 people in class B, but your population is 74,023 people in Class A and 106,794 people in Class B, for example, the weights would be different for the Class A subjects than the Class B subjects, and none of them would be integers.

3.2 About the Course

- Does this course covers logistic regression?
  - Yes, as indicated on the Calendar, this subject is the focus of (at least) Classes 7-10.
- I know last semester, labs were lined up such that its content was typically taught at least one week prior to its due date. Can we expect the same this semester?
  - Mostly, yes. Things are a little quick now, with three labs at the start. That will be less of an issue as we move forward and work on projects, too.
- Is there a list of r packages learned in 431 so I can ensure I learn all of them?
  - I don't know every function from every R package that we used in 431, so this is sort of a fool's errand. Every important package we used in 431 is also used in 432, and the most important packages are the ones we use most often in 432. The list of packages I had people install for 431 is still available.
- Should we be adding residual plots or some version of a distribution visualization to check regression assumptions to each of our lab questions when we are asked to build a regression model?
  - Not if that's not what we're asking for, no. Not sure what you have in mind, particularly. It doesn't hurt to check assumptions, but usually we're pretty focused in what we ask for in a Lab.
- Will you spend some time in an upcoming class talking more about Project A?
  - Yes, today (Class 06), I'll talk about the Project A plan. I'll discuss other elements of Project A after the Plans are submitted and reviewed.
- Will we get any bonus labs this semester, just like how we had them last semester?
  - Yes, one for sure. It'll be posted around March 1 and be due in May.
- Do you recommend working with a partner on the project or not? How can we approach for a classmate for working on a project together? I did not have a partner in project B for 431.

  - I think there are advantages and disadvantages to working with a partner. I have created a Campuswire post to help people self-identify as people looking for a partner for the Project.

- I'm thinking about using a data set for project A or B that's related to the work I do in the NDGE lab, and I was wondering if that would be feasible.
  - The instructions provide some details on what makes a useful data set in each case. I don't have much to add to those comments, but if you need more after reviewing them, let me know. Certainly the key differences between Projects A and B are (1) that the data in project A needs to be available to all, while the data in project B does not, and (2) the data in project A need to be appropriate for different modeling types than project B.

- Do you write the Labs to be deliberately tricky?
  - No. I do write them assuming you will pay close attention to details. That's a critical part of science in general, and statistics/data science inparticular.

3.3 About R/RStudio/Quarto/Coding

- Can we use R or Quarto for social media?
  - Social media is such a broad concept that I don't know how to answer this question. Quarto can certainly be used to create a blog, for instance.

- Is Quarto slower than a regular markdown file?
  - Not for me, no. If anything, it's a little faster to process.

- Why did we make the switch from R markdown to Quarto?
  - See the feedback to the Minute Paper after Class 03.

- When should we use OLS and when should we use lm in modeling?
  - It's a false choice. There's never a situation where you cannot use both.

3.4 Miscellaneous Questions

- What is your favorite part about being in the performance?
  - Two things.
    * When the audience laughs.
    * Feeling like I am part of a team.

- If you could inhabit a fantasy world of your choosing, which setting would you choose?

  - I'm not a science fiction or fantasy fan generally, so it's hard to come up with something. I suppose I like some of the ideals in the Star Trek universe, but I have no desire to spend my life on a starship.

- When did you realize you wanted to work in academia?
  - I wanted to teach, and I wanted to do research, by the time I was finishing my first Masters degree, the year after I finished college. This seemed to be the best way to approach those dual goals.

- I really enjoy reading the books you recommend for us for class 431 and 432. I feel they are really helpful. Do you mind to share a book list to me, not just limited to statistics, could be fiction or others.
  - Thanks. I read a fair number of mysteries (I've always enjoyed the Nero Wolfe series, for example, and I like Richard Osman's recent trilogy of books), and I also read lots of books about some of my particular interests (baseball, other sports, theater, comedy, television, games and puzzles) so I'm not sure how helpful that might be. If you're interested in something in particular, I could make a few suggestions.
  - If you're looking for a good book on statistical thinking, you should read Nate Silver's *The Signal and the Noise*.

4 Have you thought at all yet about Project A for this course?

Students	Response
2	I've done substantially more than just read the directions for the Project.
20	I've read some or all of the directions for the Project, but that's about it.
32	I've yet to think about it in any substantial way.

OK. We'll talk about it today in class.

5 Have you finished reading Jeff Leek's book "How To Be a Modern Scientist"?

Students	Response
1	I've not even started reading it.
38	I've started the book, but not finished it.
15	Yes, I've finished the book.

6 How did Lab 2 go for you?

Students	Response
2	I was unable to complete the Lab on time.
8	I had more than a few substantial problems but eventually completed the Lab on time.
26	I had between one and a few substantial problems but eventually completed the Lab on time.
17	I didn't have substantial problems with the Lab, and completed it on time.

7 Did you seek out help from us for Lab 2?

Sorry for the silly mistake in the question, where I referred to Lab 1 when I meant Lab 2. I hope that wasn't too confusing.

Students	Hours	Campuswire	Slides/Videos	Notes
3	No	No	No	No
3	No	No	No	Yes
7	No	No	Yes	No
15	No	No	Yes	Yes
2	No	Yes	No	No
3	No	Yes	No	Yes
1	No	Yes	Yes	No
1	No	Yes	Yes	Yes
5	Yes	No	No	No
3	Yes	No	No	Yes
2	Yes	No	Yes	No
7	Yes	No	Yes	Yes
1	Yes	Yes	No	Yes
1	Yes	Yes	Yes	Yes
Total	**19**	**9**	**34**	**35**

Available responses (check all that apply) summarized above were:

- (**Hours**) I went to TA office hours.
- (**Campuswire**) I asked for help on Campuswire.
- (**Slides/Videos**) I rewatched part or all of Dr. Love's class recordings and/or reviewed his slides.
- (**Notes**) I read part of Dr. Love's Course Notes to help me answer Lab 2's questions.