Module 11: Logistic Regression

Jeffrey Woo

MSDS, University of Virginia

Welcome

- Remind me to record the live session!
- Recommended: put yourself on mute unless you want to speak.
- Reminder: the raise hand button can be found under "Manage Participants".

Agenda

- Q&A
- A few comments about Module 11
- Small group discussion of guided question set
- Large group discussion of guided question set and other questions that popped up

Q&A

Questions?

Comment about Indicator Variables in R

- When using lm() or glm(), R converts factors to indicator variables. So, if your categorical variables are already coded as 0/1 indicator variables, they can be left alone.
- However, for other functions, 0/1 indicator variables may not work. Have to check documentation on how functions handle categorical variables.
- For example, using `glmnet()` for shrinkage methods and `tree()` for tree based methods.
- Most visualizations need categorical variables as factors and cannot be 0/1 coded, especially if using `ggplot2` functions.

If something in your output doesn't look right and you have categorical variables, make sure that your categorical variables are the correct type.

Logistic Regression Equation

Typically written as

$$\begin{aligned} \log \frac{\pi}{1-\pi} &= \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \\ &= \boldsymbol{X\beta} \end{aligned}$$

Alternate way of writing:

$$\begin{aligned} \pi &= \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)} \\ &= \frac{\exp(\boldsymbol{X\beta})}{1 + \exp(\boldsymbol{X\beta})} \end{aligned}$$

Reproducibility Poll

Run these 2 lines of code and record the 5 integers you get

- `set.seed(1)`
- `sample.int(100,5)`

I suggest you use set.seed() when you generate random numbers, so the same numbers are generated each time you run the code (or share the code for someone else to run).

Random Numbers

- If you got the wrong set of 5 integers, type `RNGkind(sample.kind = "Rejection")` on a line before `set.seed()`.
- Should use rejection sampler (and that should be the one that is used by default).
- Type `RNGkind(sample.kind = "Rejection")` or `RNGkind(sample.kind = "Rounding")` for needed sampler (type this on a line before using the set.seed() function).
- Old versions of R use the rounding sampler by default. So if you're following resources that were published a few years ago, they are likely using the old sampler.

Small Group Discussion

- Materials can be found under Module 11 Live session.
- Have the guided question set and corresponding data set open.
- Have R open.
- Recommended: have easy access to your notes, textbook, as well as the tutorial.
- You can see who your group members are. As well as some roles you will have in your small group. Roles will rotate each session.

Goodness of Fit Tests

Three Goodness of Fit (GOF) tests in logistic regression. In other words is the log odds a linear combination of coefficients and predictors:

1. Deviance GOF test
2. Pearson GOF test
3. Hosmer-Lemeshow test

Deviance and Pearson require grouped data (usually found in designed experiments); cannot have (a number of) observations that have unique combination of predictors. Difficult to implement in observational studies.

Residuals in Logistic Regression

Due to the binary nature of the response variable in logistic regression, examining residuals is not very helpful (due to discrete nature of response, the values of the residuals are typically discrete). Again, require data are grouped to reliably use the plots.

Where Are We Headed?

- Module 12: Validating the Logistic Regression Model.

Split data into training and test data. Use training data to build model. Use test data to assess predictive ability.

Upcoming

- Exam, due Apr 28. Covers Modules 3 to 10. Opens after our last “live session” on Apr 23.
- Project 2 due May 8 (bulk of work) and May 9.
- Please see my comments on your proposal. There are some that I approved based on an assumption. If my assumption is incorrect, we need to talk.