Module 12: Logistic Regression 2

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

- Large group discussion of guided question set.
- Q&A.
- End of semester memos.

Large Group Discussion

The confusion matrix with 0.5 as the cutoff:

```
> table(test$chd69, preds>0.5)
    FALSE
  0  1449
  1   128
```

- Among males who do not have heart disease, 0 out of 1449 are incorrectly classified as having heart disease. We have a false positive rate of 0.
- Among males who do have heart disease, 0 out of 128 are correctly classified as having heart disease. We have a true positive rate of 0.
- Where are these values of FRP and TPR on the ROC curve?
- Overall error rate of $\frac{128}{1577} = 0.081$.

Large Group Discussion

Compare with a “useless” classifier that guesses FALSE for every observation. Same results.

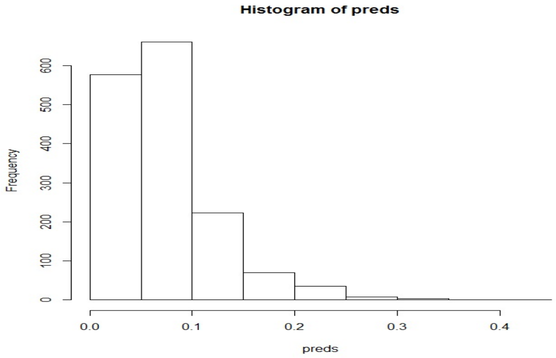
Large Group Discussion

The confusion matrix with 0.08 as the cutoff:

```
> table(test$chd69, preds>0.08)
    FALSE TRUE
  0   940  509
  1    36   92
```

- Among males who do not have heart disease, 509 out of 1449 are incorrectly classified as having heart disease. We have a small false positive rate, but the false positive rate has increased.
- Among males who do have heart disease, 92 out of 128 are correctly classified as having heart disease. The true positive rate has improved (but still isn't great).
- Where are these values of FRP and TPR on the ROC curve?
- Overall error rate of $\frac{509+36}{1577} = 0.346$.

Estimated Probabilities



- The predicted probability of having heart disease is small among middle-aged men.
- Using a cutoff of 0.5 to classify an observation as having heart disease may be unrealistic in this context, since we are modeling a rare event.

Large Group Discussion

- A threshold of 0.5 minimizes the overall error rate, on average.
- Depending on your situation, you may be more concerned with controlling the false positive or false negative rate more than the overall error rate. Can adjust the threshold accordingly.
    - Use context to guide your decision, and / or consult with subject matter expert.

Cautions

- Accuracy / error rate may be a misleading measure when you have unbalanced sample sizes of the two classes. Error rate may be low but either FPR or FNR may be high.
- The ROC curve shows the true positive and false positive rates as the threshhold is varied. It does not immediately inform you of the true positive and false positive rates for your specific threshold.
- Similar comment with the AUC.
- Very often, I see people report the accuracy / error rate, display the ROC curve and report the AUC and are happy. This is not enough! With unbalanced sample sizes confusion matrix (or at least FPR and FNR) must be checked.

Cautions

Some workarounds include

- adjusting the threshold
- finding better predictors than can distinguish between the two classes
- adjust the population of interest (so response variable is more balanced, e.g. instead of considering men aged 39 to 59, only consider men aged 55 to 59)

Goals

Uses of models:

- prediction: given predictors, what is the predicted response?
- inference: how do predictors relate to the response variable?

With unbalanced data, prediction may get more challenging. However, we can still gain insights into how the predictors relate to the log odds of “success”.

Separation

Separation: when predictors (almost) perfectly predict the binary response variable. Consequence:

- Estimated standard errors of coefficients get large.
- R will produce a warning message that some predicted probabilities are 0 or 1.

With separation, inference gets challenging. Confidence intervals associated with coefficients are very wide. However, prediction works fine. Some workarounds:

- If the dataset is small, sometimes collecting more data helps break the separation. Does separation exist in the population?
- If categorical predictor has many classes, collapsing some classes may help break the separation.

Assignments

- HW 11 & 12, due Monday Dec 5.
- Group Evals for Mods 9 to 12 discussions, due Monday Dec 5, via Test & Quizzes.
- Exam, due Sunday Dec 4, via Assignments. Opens right after today's live session.
    - Any question regarding the Exam MUST be posted on the discussion forum. So my answers are available to all. Questions via email will be ignored.
- Project 2:
    - Parts 3 and 4 (report and recording of presentation), due Wednesday Dec 14. One upload per group.
    - Parts 5 and 6 (feedback for other group and group eval), due Thursday Dec 15. Everyone completes.

Office Hours

- Thursday, Dec 1, 7:30 to 8:30pm as usual.
- Monday, Dec 5, 7:30 to 8:30pm as usual.
- Thursday, Dec 8, 7:30 to 8:30pm.
- Friday, Dec 9, 7:30 to 8:30pm.
- Sunday, Dec 11, 8:30 to 9:30pm.
- Wednesday, Dec 14, 7:30 to 8:30pm.

Dec 8 onwards will be Project 2 consultations. Sign up for 15-minute consultation for your group. Details on Overview page.