Module 6: Categorical Predictors

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
- Practical considerations with categorical predictors
- Small group discussion of guided question set
- Large group discussion of guided question set and other questions that popped up

Q&A

Practical Considerations with Categorical Predictors

- Categorical predictor with **many classes**
- Consideration of **interaction terms**
- Variances **not equal** across all classes

Categorical Predictor with Many Classes

- With many classes, output can be daunting to look at.
- Are you really interested in exploring the differences in the mean response across all the classes?
- Is there a logical way to **collapse** some classes together that still answers your research question and reduces the number of classes?
- Having more parameters than needed leads to **overfitting**, which typically leads to poor performance on test data.

Consideration of Interaction Terms

- Typically, people start with an **additive first order** model.
- Interactions considered at the start if:
  - Exploring interactions is part of your research question
  - An interaction makes sense contextually, or is well-established in the literature

Consideration of Interaction Terms

Can use some EDA to explore possibility of interactions:

- Interaction between a categorical predictor and a quantitative predictor: create scatterplot of response against quantitative predictor, create separate regression lines for each class.
- Interaction between two categorical predictors: side by side boxplots, or **interaction plot**.
- Interaction between two quantitative predictors. If one is discrete, can use scatterplot approach, with separate regression lines for each value of the discrete variable. If both are continuous, you will have to create separate lines for various fixed values of one of the predictors.

A resource to create these plots easily with ggplot2:
`https://cran.r-project.org/web/packages/interactions/vignettes/interactions.html`

Variances not Equal Across all Classes

Additional assumption: variances are **equal** across all classes.

- With balanced sample sizes, the model is robust to this assumption.
- With unbalanced sample sizes, check ratio of largest variance with lowest variance. If not more than 1.5, ok.
- Collapsing classes can help, especially if it makes sense to do so and results in balanced sample sizes.
- Can fit separate regressions for the classes with different variance.

Breakout Rooms

Work through questions 1, 2, 3, 5, 4 in this order.

Pairwise Comparisons

- Bonferroni method is even more conservative than Tukey but easier to work out "by hand".
- I've added an additional pdf (mod6_pairwise.pdf) under the tutorial for module 6. This pdf goes over how to use the Bonferroni method. This pdf should be read after going over the tutorial. You will need this for one of the questions in HW 6.

Upcoming

- HW 6 next Tuesday Oct 13 as usual.
- Project 1 due Oct 19.