Module 7: Sums of Squares and Multicollinearity

Jeffrey Woo

MSDS, University of Virginia

Welcome

- You can download this set of slides. Find with the materials for the live session in Module 7.7.
- Remind me to record the live session!
- Recommended: put yourself on mute unless you want to speak.
- There is a “raise hand” button for you. Click on “Reactions” in the panel at the bottom.

Agenda

- Some comments on Module 7
- Q&A
- Small group discussion of guided question set
- Large group discussion of guided question set plus other questions that pop up

Uses of Regression Models

Two main uses of regression models:

1. Prediction
2. Explore relationship between response and multiple predictors simultaneously.

- Including more predictors or higher order terms typically improves model fit (in fact R^2 never decreases), but also make the model more difficult to interpret.
- Making a model more complicated than needed can result in **overfitting**, which leads to poor predictive performance on new data.

Partial F Test

The partial F test allows us to assess if multiple predictors can be dropped simultaneously from the model. The partial F statistic measures the change in the SS_R (or SS_{res}) with the removal of these predictors from the model.

- Essentially, is the improvement in SS_R (and hence R^2) large enough that it is worth the extra complexity?

Sum of Squares

- As long as we have the same response variable, SS_T is constant, regardless of the number and form of predictors used.
- $SS_T = SS_R + SS_{Res}$
- Each time predictors are added to the model, the SS_R increases and the SS_{Res} decreases by the same amount, since SS_T stays constant.

Model Selection

- The partial F test only works when comparing two models, the parameters of one being a subset of the parameters of another.
- Other measures are used in other situations (module 9).

Reading ANOVA Output from R

Let's look at the example from tutorial 7.

Issues with Multicollinearity

When predictors are nearly linear dependent on each other. Issues:

- High variance with estimated coefficients: the estimated coefficient may be very different from the true value.
    - Caution with interpreting estimated coefficients in the usual manner.
    - Estimated coefficients tend to be large.
    - Algebraic sign of coefficients different than what is known theoretically.
    - Adding or removal of one or more data points results in large changes in the estimated regression coefficients.
- Predictions are fine but must be very careful with extrapolation.

Detecting Multicollinearity

- Insignificant t tests for predictors that are known to be useful in predicting the response variable, and significant ANOVA F test.
- High VIFs (exceeds 10).
- High correlation between pairs of predictors.

Solutions when Multicollinearity is Present

- Consider a subset of predictors (among those that are linearly dependent).
- Shrinkage methods.
- Dimension reduction methods.

Small Group Discussion

What is coming up...

Module 8: Categorical predictors.