Module 9: Model Selection & Data Splitting

Jeffrey Woo

MSDS, University of Virginia

Welcome Back

- 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

- A few comments about Module 9
- Q&A
- Small group discussion
- Large group discussion
- Proj 2

Model Selection Criteria

- R^2 should only be used when comparing models of the same size. Adding predictors to a model will always increase R^2 (since SS_R increases and SS_{res} decreases).
- Other measures such as adjusted R^2, Mallow's C_p, AIC, BIC are sometimes called **penalized-fit criteria**. A penalty is added when an extra term is added to the model to improve the fit of the model. E.g. for AIC

$$AIC = n \log(\frac{SS_{res}}{n}) + 2p$$

- These measures can be used to compare models when the general linear F test cannot be used.

Note: when using these model selection criteria, the response variable has to be the same across the models.

PRESS

$$PRESS = \sum_{i=1}^{n} \left(y_i - \hat{y}_{i(i)} \right)^2 \tag{1}$$

Motivated by assessing how well your model does in predicting new observations.

R^2_{pred}

$$R^2_{pred} = 1 - \frac{PRESS}{SS_T}$$

- R^2_{pred} can be interpreted as the proportion of variance in new observations the model might be able to explain.
- High values indicate a model that will perform well on prediction (test) data.
- Typically lower than R^2.
- R^2_{pred} much lower than R^2 indicative of overfitting.

Automated Search Procedures

- Be careful in viewing these procedures as producing the final model you want to use.
- Typically, if I use these, I use these to generate an initial model, then seek to improve the model with methods learned.
- These procedures do not check if assumptions are met.

Q&A

Any questions from module 9?

Small group discussion of Guided Question Set

Large group discussion

Project 2 Intro

Parts 1 and 2, due Apr 14.

Breakout rooms:

- Introduce yourselves to each other!
  - professional / academic interests
  - non professional / academic interests
- Start working on the group expectations agreement (Part 1). For today, at least discuss the following:
  - Mode of communication outside of class
  - Schedule next meeting
  - How to conduct meetings

Feel free to ask for help as we hop around. You may also start to work on other parts of Part 1 and 2.

Where Are We Headed?

- Module 10: how to transform predictors in MLR; how to detect outliers and influential observations. I will also have comments tying in everything you learned in linear regression.
- Module 11 & 12: Logistic regression (binary response variable)