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
- A question from the guided question set
- Proj 2

Model Selection

Two main uses of regression models:

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

- Including more predictors or higher order terms can improve model fit, 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.

Model Selection

- 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 partial F test cannot be used.

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

Automated Search Procedure

- Only consider first order models (no interactions or higher order terms)
- Do not check if regression assumptions are met
- Do not guarantee the best model is identified

Comments on Data Splitting

- In data splitting, a data set is randomly split into two portions: the estimation data and the prediction data.
- The estimation data are used to build the regression model, and the prediction data are used to evaluate the predictive ability of the model.
- The estimation data and prediction data are also called training set and test set respectively.

PRESS

The PRESS statistic is a measure based on data splitting

$$PRESS = \sum_{i=1}^{n} \left[y_i - \hat{y}_{(i)} \right]^2 = \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}} \right)^2 .$$

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.

Q&A

Any questions from module 9?

Guided Question Set Discussion

Project 2 Intro

Parts 1 and 2, due November 13.
Breakout rooms: start working on Part 1: Group Expectations Agreement.

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)

Reminders

- There will be no class next Tuesday, Nov 8 (Election Day), per the University calendar.
- I will still hold office hours on Monday, Nov 7, and Thursday, Nov 10.
- No office hours on Monday, Nov 14.
- Module 10 live session on Tuesday, Nov 15.
- Module 11 live session on Tuesday, Nov 22.
- Module 12 live session on Tuesday, Nov 29.