Model Selection & Data Splitting

Jeffrey Woo

MSDS, University of Virginia

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.

Comments on Automated Search Procedures

- `regsubsets` and `step` functions in R only consider 1st order models (no interactions or higher order terms).
- `regsubsets` and `step` functions do not check if the regression assumptions are met. You still need to check the residual plot.
- `regsubsets` and `step` functions do not guarantee the best model will be identified.
- `step` function can lead to different models if you have a different starting point.
- For the `step` function, R uses AIC to decide when to stop the search. The textbook describes using the F statistic.

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.