Assignment 4

Ethan Tenison

4/13/2020

# Q1: Model Selection in the abstract (20 points)

The decennial census is a massive undertaking that involves hiring thousands of short-term employees and, locally, volunteering opportunities. The Census’ 2016 Planning Database Block Group Data contains 344 variables describing 220,357 block groups. The variable Mail\_Return\_Rate\_CEN\_2010 is the 2010 Census Mail Return Rate – suppose that when this rate is low in a block group, more people are needed to help with the count. We know the return rates, along with a lot of other characteristics of block groups, for 2010 but a lot has changed in ten years. Understanding the most important drivers of low return rates in 2010 could help with planning for 2020.

### (a) Describe how you might use forward vs. backward stepwise selection to find the best model of

low return rates. Explain the advantages and disadvantages of one approach over the other. Explain how these approaches compare to best subset selection.

Using forward and backward stepwise selection will help us decide which features are most important in our model, especially when n < p. Both are computationally superior to best subset selection, where 2^p models are created. While FSS is computationally superior, there is still a chance that the selected model will be the best. FSS adds predictors one at a time, and each successive step might not include the same vital predictors as the previous one. BSS does not have this problem because all predictors are added into the model first, and then substracted. The downside of BSS is that n must be greater than p.

### (b) How would you use model-level comparison metrics to determine when one model is better

than another? Explain how a “penalty” works and how they manifest differently in different comparison metrics.

There are essentially two methods for comparing models, make adjustments to the training error to account for overfitting, and directly estimating test error through cross-validation. Since R^2 and MSE increase and decrease monotonically as predictors are added, metrics such as Cp, AIC, BIC, and adjusted R^2 are used. Each of them has a penalty term which increases as does the number of predictors.

**Cp** = (RSS - 2d), where d is the number of predictors, and is an estimate of the error associated with each response measurement. THis is done to adjust for the decline in RSS that occurs when more predictors are added.

**AIC** = (RSS - 2d), which uses maximum likelihood and is proportional to *Cp*

**BIC** = (RSS - log(n)d), which uses a log instead of 2d, resulting in a larger penality for a high number of predictors.

**Adjusted**  = , the result being that adding predictors that are just noise will only slightly decrease the RSS.