3.2: Variable Selection

Dr. Bean - Stat 5100

Why might we prefer simpler models to complex ones? *Should* we prefer simpler models to more complex ones?

- Simpler models are easier to interpret/describe.
- Simpler models are harder to overfit, which means models that fit the sample data very well but generalize to new data poorly.
- *Conversely,* simpler models may fail to describe a complex problem.

Why is variable selection not something we would normally want to use in an experimental setting?

Observational studies are usually searching to find *something* interesting, while in an experiment, we wish to test whether *specific things* are interesting. In experiments, we should have decided beforehand what factors we were going to control for.

Clearly, it would be better to compare all possible models, rather than a subset (in stepwise methods). Why do stepwise methods even exist?

When P gets large, fitting 2^{P-1} models quickly becomes unrealistic computationally. Also, chance of a "consensus" among measurement techniques as to which is the best model becomes unlikely.

In what ways might multicollinearity "mess-up" stepwise selection approaches?

Stepwise selection techniques rely heavily on the p-values associated with the coefficient t-tests. Multicollinearity inflates the variance of the estimated coefficients which renders the p-values useless. Stepwise selection techniques could potentially reject two variables that are related to each other that have significant t-tests when considered individually, but insignificant results when considered separately.

(Open-ended discussion, no right answer) How do we decide on the "right" model when each variable selection technique suggests a different model?

One question you might ask is: how difficult is it to obtain values for the predictor variables in question? If it is difficult to collect the information for a candidate predictor variable AND the inclusion of that variable in the model is questionable, we may choose to exclude it. If the information is easy to collect, we may choose to include it.