

# 1 Analysis

So now that we've gone over each variable selection method, how do we know which one to use? We could see how each model (Lasso, P Value-based forward selection, BIC-based forward selection) performs from a predictive standpoint and use the model that has the most predictive power. However, the model that has the most predictive power is not necessarily the model that is the best at variable selection. The optimal model for variable selection is the model that is able to most accurately estimate beta coefficients for our explanatory variables. Unfortunately, given that we don't know each explanatory variable's true effect on graduation rate, there is no way for us to check how accurate our beta coefficient estimates are. As a result, there's no simple way for us to pick the model that is the best at variable selection.

All is not lost, however. While we may not be able to pick a single model to use for variable selection, we can just use all of 3 of our models. As a result, each model will identify a set of relevant explanatory variables. If a subset of these variables appear in all 3 of our models, we can identify these variables as extremely relevant to graduation rates. In the next section, we're going to go over how we used our Shiny R app to help us find the relevant variables for graduation rates.