Math 158, Spring 2022
Jo Hardin

One Model Building Strategy

Model building is an **art**. Unsurprisingly, there are many approaches to model building, but here is one strategy, consisting of seven steps, that can be used when building a regression model.[1]

The first step

Decide on the type of model that is needed in order to achieve the goals of the study. In general, there are five reasons one might want to build a regression model. They are:

- For predictive reasons - that is, the model will be used to predict the response variable from a chosen set of predictors.
- For theoretical reasons - that is, the researcher wants to estimate a model based on a known theoretical relationship between the response and predictors.
- For control purposes - that is, the model will be used to control a response variable by manipulating the values of the predictor variables.
- For inferential reasons - that is, the model will be used to explore the strength of the relationships between the response and the predictors.
- For data summary reasons - that is, the model will be used merely as a way to summarize a large set of data by a single equation.

The second step

Decide which explanatory variables and response variable on which to collect the data. Collect the data.

The third step

Explore the data. I can't possibly over-emphasize the data exploration step. There's not a data analyst out there who hasn't made the mistake of skipping this step and later regretting it when a data point was found in error, thereby nullifying hours of work.

- On a univariate basis, check for outliers, gross data errors, and missing values.
- Study bivariate relationships to reveal other outliers, to suggest possible transformations, and to identify possible multicollinearities.

The fourth step

(The fourth step is very good modeling practice. It gives you a sense of whether or not you've overfit the model in the building process.) Randomly divide the data into a training set and a validation set:

- The training set, with at least 15-20 error degrees of freedom, is used to estimate the model.
- The validation set is used for cross-validation of the fitted model.

[1] Taken from `https://online.stat.psu.edu/stat501/lesson/10/10.7`.

The fifth step

Using the training set, identify several candidate models:

- Use best subsets regression.
- Use stepwise, forward, or backward selection regression. Using different alpha-to-remove and alpha-to-enter values can lead to a variety of models.

The sixth step

Select and evaluate a few "good" models:

- Select the models based on the criteria we learned, as well as the number and nature of the predictors.
- Evaluate the selected models for violation of the model conditions.
- If none of the models provide a satisfactory fit, try something else, such as collecting more data, identifying different predictors, or formulating a different type of model.

The seventh and final step

Select the final model:

- A small mean square prediction error (or larger cross-validation R^2) on the validation data is a good predictive model (for your population of interest).
- Consider residual plots, outliers, parsimony, relevance, and ease of measurement of predictors.

And, most of all, don't forget that there is not necessarily only one good model for a given set of data. There might be a few equally satisfactory models.

Getting the Variables Right

Underspecified

A regression model is underspecified if it is missing one or more important predictor variables. Being underspecified is the worst case scenario because the model ends up being biased and predictions are wrong for virtually every observation. Additionally, the estimate of MSE tends to be big which yields larger confidence intervals for the estimates (less chance for significance).

Extraneous

The third type of variable situation comes when extra variables are included in the model but the variables are neither related to the response nor are they correlated with the other explanatory variables. Generally, extraneous variables are not so problematic because they produce models with unbiased coefficient estimators, unbiased predictions, and unbiased MSE. The worst thing that happens is that the error degrees of freedom is lowered which makes confidence intervals wider and p-values bigger (lower power). Also problematic is that the model becomes unnecessarily complicated and harder to interpret.

Overspecified

When a model is overspecified, there are one or more redundant variables. That is, the variables contain the same information as other variables (i.e., are correlated!). As we've seen, correlated variables cause trouble because they inflate the variance of the coefficient estimates. With correlated variables it is still possible to get unbiased prediction estimates, but the coefficients themselves are so variable that they cannot be interpreted (nor can inference be easily performed).