IDS 702: Module 1.11

Model building and selection

Dr. Olanrewaju Michael Akande

IDS 702

Which predictors should be in your model?

- This is a very hard question and one of intense statistical research.
- Different people have different opinions on how to answer the question.
- It also depends on the goal of your analysis: prediction vs. interpretation or association.
- We will not focus on answering the question on which is the best "overall".
- Instead, we will focus on how to approach the problem and the most common methods used.
- See Section 6.1 of An Introduction to Statistical Learning with Applications in R for more details on the methods we will cover.

IDS 702

What variables should you include?

- Goal: prediction
  - Include variables that are strong predictors of the outcome.
  - Excluding irrelevant variables can reduce the widths of the prediction intervals.
- Goal: interpretation and association
  - Include all variables that you thought apriori were related to the outcome of interest, even if they are not statistically significant.
  - This improves interpretation of coefficients of interest.

IDS 702

Model selection criterion

Model selection criterion

The most common are:

- Adjusted R-squared:

$$\text{Adj}.R^2 = 1 - (1 - R^2)\left[\frac{n-1}{n-p-1}\right]$$

- Akaike's Information Criterion (AIC):

$$\text{AIC} = n\ln(\text{RSS}) - n\ln(n) + 2(p+1)$$

- Bayesian Information Criterion (BIC) or Schwarz Criterion:

$$\text{BIC} = n\ln(\text{RSS}) - n\ln(n) + (p+1)\ln(n)$$

where n is the number of observations, p is the number of variables (or parameters) excluding the intercept, and RSS is the residual sum of squares, that is,

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

IDS 702

Model selection criterion

- Note:
  - Large $\text{Adj.}R^2$ = good!
  - Small AIC = good!
  - Small BIC = good!
- Notice that BIC generally places a heavier penalty on models with many variables for $n > 8$ since

  $$\ln(n)(p+1) > 2(p+1)$$

  for fixed p and $n > 8$.
- Thus, BIC can result in the selection of smaller models than AIC.
- *Note: the formulas for $\text{Adj.}R^2$, AIC and BIC in Section 6.1 of An Introduction to Statistical Learning with Applications in R take slightly different forms but are equivalent to those given here when comparing models.*

IDS 702

Common selection strategies

Backward selection

- Start with the full model that includes all p available predictors.
- Drop variables one at a time that are deemed irrelevant based on some criterion.
  - Drop the variable with the largest p-value (from nested F-test if categorical variable).
  - Drop variables (possibly all at once) with p-value over some threshold (for example, 0.10).
  - Drop the variable that leads to the smallest "change" in AIC, BIC, or $\text{Adj.}R^2$.
    You might even consider using average MSE from k-fold cross-validation if the goal is prediction.
- Stop when removing variables no longer improve the model, based on the chosen criterion.

Forward selection

- Start with the model that only includes the intercept.
- Add variables one at a time based on some criterion.
  - Add the variable with the smallest p-value using some threshold (for example, 0.10).
  - Add the variable that leads to the smallest value of AIC or BIC, or the largest value of $\text{Adj.}R^2$.
    Again, you might consider using average MSE from k-fold cross-validation if the goal is prediction.
- Stop when adding variables no longer improves the model, based on the chosen criterion.

Stepwise selection

- Start with the model that only includes the intercept.
- Potentially do one forward step to enter a variable in the model, using some criterion to decide if it is worth including the variable.
- From the current model, potentially do one backwards step, using some criterion to decide if it is worth dropping one of the variables in the model.
- Repeat these steps until the model does not change.

Model Selection in R

- step function (in base R): forward, backward, and stepwise selection using AIC/BIC.
- regsubsets function (leaps package): forward, backward, and stepwise selection using $\mathrm{Adj}.R^2$ or BIC.

IDS 702

Other options: shrinkage methods

- Fit a model containing all p available predictors, then use a technique that shrinks the coefficient estimates towards zero.
- The two most common methods are:
  - Ridge regression
  - Lasso regression (performs variable selection)
- We will not cover these methods in this course.
- Consider taking STA521 if you are interested in learning about how they work.

What's next?

Move on to the readings for the next module!

IDS 702