Variable Selection & Model Building

Dr Austin R Brown

Kennesaw State University

Introduction

- Some of the materials used in today's lecture were adapted from those created by Dr. Taasoobshirazi and my former professor, Dr. Khalil Shafie (thanks Drs S & T!).

Introduction

- To this point, we have assumed that we, the people actually building the regression models, know that our included predictor variables are important and ought to be included.

- Our general process has been is: (1) fit the full model; (2) check all assumptions including residual analysis; (3) perform any necessary transformations/use different methods; (4) perform all relevant tests (omnibus and individual predictor); (5) reevaluate if necessary.

- This is well and good, but what happens if we have a lot of candidate predictor variables, but aren't sure which would be most important to include?
  - This is a common problem!

Introduction

- As we've somewhat discussed earlier in the semester, when we're building a regression model, we want as many predictor variables as possible in order to better inform the variability in our outcome variable.

- On the flip side, we want as few regressors as possible seeing as an increase in regressors can also create an issue with multicollinearity, and consequently, an inflated estimate of σ^2.

- Thus, the goal in model building (in all model based methods, not just linear regression) is to balance these two aims to obtain a "best" fitting model.
  - Note, "best" doesn't have a universally agreed upon meaning. We'll see different ways we can assess what "best" means throughout today's lecture.

Introduction

- Throughout this semester, it has come up many times that a simpler model (one involving fewer predictors) can be preferable to a more complex model.
  - At least according to Dr B!

- However, it can be shown that a model which is "underfit," that is, one which has less predictors than the true model at the population level, biases the estimates of the predictors retained in the model as well as the estimate of the variance.

- While we can never truly know if we've underfit a model, these real consequences need to be taken into consideration before removing variables from a model.

Introduction

- Now on the other hand, and as mentioned, overfitting a model (including more predictors than is necessary/the true population model contains) also has consequences.

- Multicollinearity is one of the big issues. As discussed in last class, multicollinearity inflates the variance of our regression coefficient estimates and can resultingly decrease statistical power.

- Additionally, the inclusion of unnecessary predictors also takes away degrees of freedom from SSE, which in turn increases our estimate of MSE, which in turn decreases statistical power.

Introduction

- So ultimately we're wanting to build a model which does two things: (1) adheres to the principle of parsimony while (2) not leaving out anything important.

- Note, the procedures we will discuss do not necessarily yield the same results nor should they be solely relied upon when comparing models. We still have to use contextual information to help inform our decision making.

Criteria for Model Selection

- As we're comparing different models, what tools do we have to do that? Well, lots! Some that we know, and some that we may not yet be familiar with.

- First up are our old friends R^2 and adjusted R^2. Recall:

$$R^2 = \frac{SSR}{SST}$$

- and is interpreted as the proportion of variability in the response explained by our predictors. However, remember that as we add predictors to our model, R^2 will increase (because of math reasons).

Criteria for Model Selection

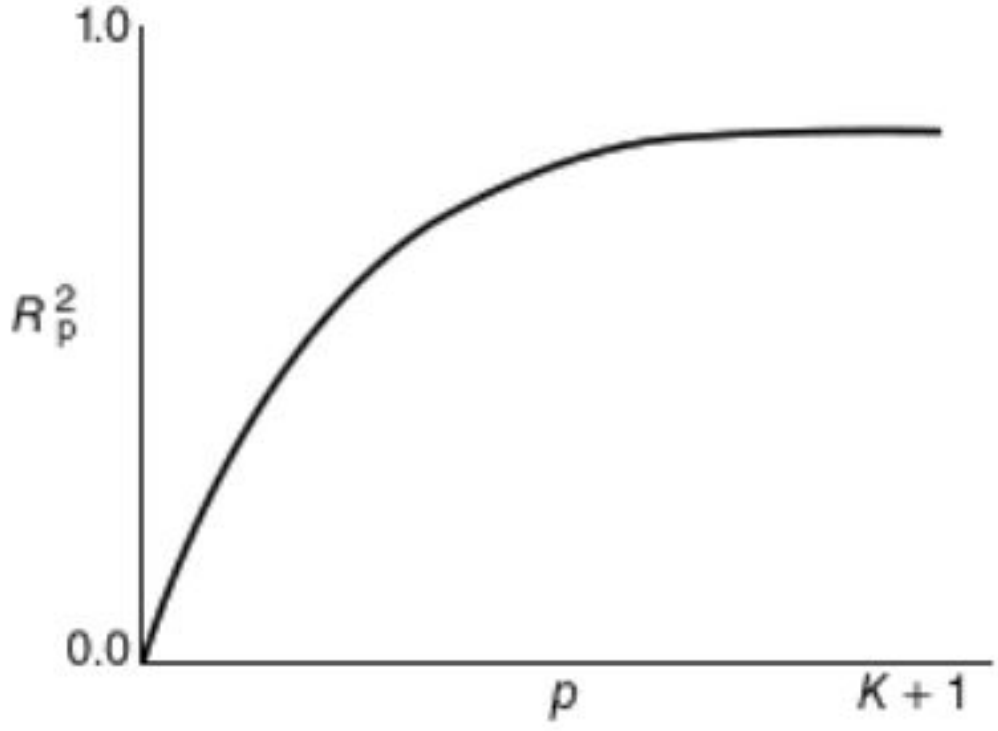

Figure 1: Figure 10.1 from Montgomery text

Criteria for Model Selection

- So typically, we use the adjusted R^2 instead of regular R^2 because it captures the true reduction in SSE. Recall:

$$R^2_{\text{Adj}} = 1 - \frac{MSE}{s_y^2}$$

- We traditionally prefer this measure over regular R^2 because it doesn't necessarily increase when an additional regressor is added to the model.
  - Further, we can think of the rightmost term as the ratio between the marginal and conditional variance estimates, which makes this measure quite informative!

Criteria for Model Selection

- Another commonly used measure for model comparison is MSE. Remember, MSE is:

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p} = \frac{SSE}{n-p}$$

- So when comparing two models, a smaller MSE implies a smaller deviance between y_i and $\hat{y}_i$ which further suggests that the predictors used in that model yield a better model fit.

Criteria for Model Selection

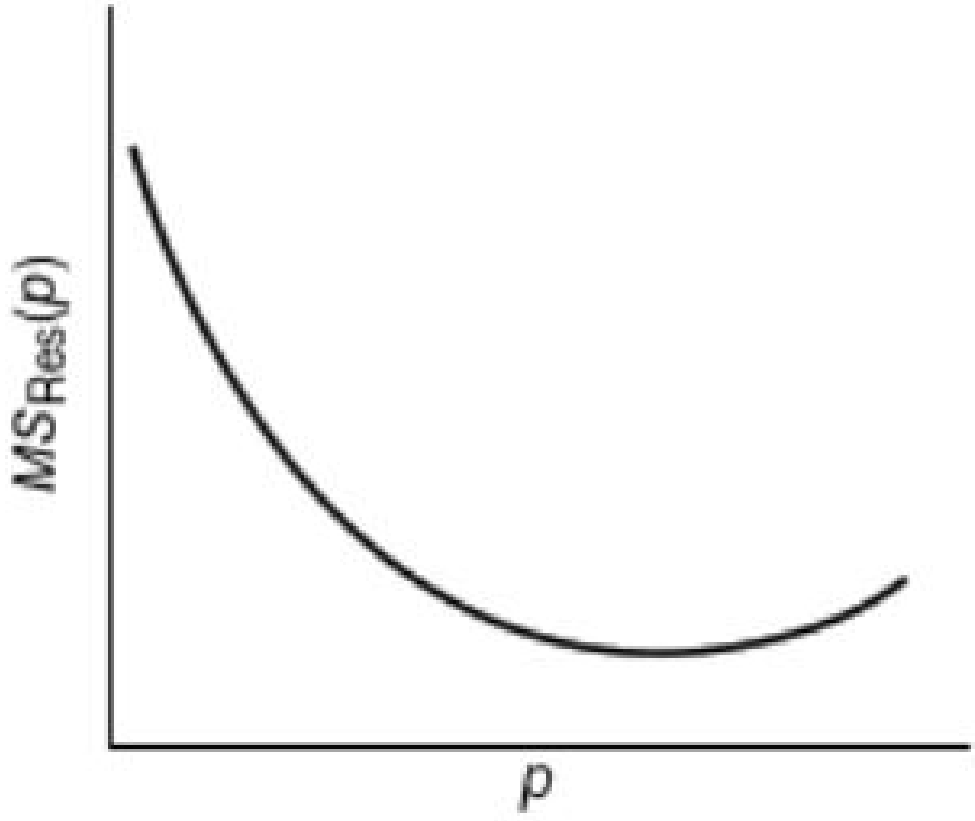

Figure 2: Figure 10.2 from Montgomery text

Criteria for Model Selection

- Another common metric used to compare competing models is called Mallows's C_p. This measure is defined as:

$$C_p = \frac{SSR_{p-1}}{MSE_p} - n + 2p$$

- C_p is basically a measure of bias, comparing the ratio of SSR in a model with $p-1$ β's to the MSE of a model with the full set of p β's, penalizing for the number of predictors. Smaller values are preferable to larger values when comparing two models.

- If we see $C_p > p$ or $C_p < 0$, then this implies the presence of bias, either due to overfitting or possibly underfitting.

Criteria for Model Selection

- Another metric that we've seen before is use of the PRESS residuals. Recall, a PRESS residual is:

$$PRESS_i = y_i - \hat{y}_{(i)}$$

- The much more commonly used PRESS statistic is used to compare models, especially their ability to predict new observations.

$$PRESS_p = \sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2 = \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}} \right)^2$$

Criteria for Model Selection

- Two other popular metrics for comparing models are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).

- As their names imply, they use information (via the likelihood, which is a function of the random sample where all of the information about a sample is contained...you'll learn more about this in math stats).

$$AIC = 2p + n \log\left(\frac{SSR}{n}\right)$$

$$BIC_{\text{Schwartz}} = n \log\left(\frac{SSR}{n}\right) + p \log(n)$$

- Smaller values are preferable to larger values.

Computational Techniques for Variable Selection

- Typically, when we have a large pool of predictors to select from, it is quite inefficient for us to manually build a bunch of different models and compare all of the metrics by hand (computer programming 101: less code is preferable to more code or more colloquially it's better to be lazy haha).

- There are a couple of common techniques we can use in order to make the computer build models for us, calculate all of the comparative metrics, and then give us some output that we can evaluate.

- The first is called "all possible regressions," and does exactly what it sounds like it does. If we have K predictors (assuming β_0 is included in the model), then there are 2^K possible regression equations.

Computational Techniques for Variable Selection

- Let's use the Cement data to see how this works.

Computational Techniques for Variable Selection

- All possible regressions is a pretty useful tool for model selection. It's primary limitation, however, is computational efficiency (even a relatively small number of variables, like 30, would generate over one billion regressions).

- So we need a more computationally efficient way of including variables in the model that doesn't involve fitting every single regression model possible.

- This is where, forward, backward, stepwise and lasso variable selection come into play.

Computational Techniques for Variable Selection: Forward Selection

- The way forward selection works is by starting with a model containing 0 predictors (sometimes referred to as the *null* model).
- Then, predictors are iteratively added, one at a time, until all of the predictors are in the model.
- The variable to be added at each step is that which gives the greatest *additional* improvement to the fit of the model.
- We can look at this algorithmically, as taken from ISLR Chapter 6:

Computational Techniques for Variable Selection: Forward Selection

1. Start with the null model, denoted M_0
2. Fit a SLR using each predictor and calculate AIC or R^2. 2a. Choose the model with the lowest AIC or highest R^2.
3. Repeat steps 2 and 2a now fitting a MLR.
4. Stop when all variables have been added or when improvements to metric of choice drop below a specified threshold.

Computational Techniques for Variable Selection: Forward Selection

- We can see that forward selection is a gigantic improvement in terms of computational efficiency over all-possible-regressions.
- Considering the cement data, for all possible regressions, we had $2^4 = 16$ models.
- For forward selection, we have $1 = p(p+1)/2$ possible models. So for the cement data, this adds up to 11 models.
- For more variables, obviously the improvement gains increase substantially.

Computational Techniques for Variable Selection: Backward Selection

- Backward selection works opposite of forward selection. We now start with a model containing all p predictors (sometimes referred to as the *full* model).
- Then, predictors are iteratively removed, one at a time, until we arrive at the null model.
- So our algorithm for backward selection is basically just the opposite of forward selection:

Computational Techniques for Variable Selection: Backward Selection

1. Start with the full model, denoted M_p
2. Fit a MLR using all but one predictor (for all predictors) and calculate AIC or R^2. 2a. Choose the model with the lowest AIC or highest R^2.
3. Repeat steps 2 and 2a with the model obtained from the original step 2a.
4. Stop when all variables have been removed or when improvements to metric of choice drop below a specified threshold.

Compuational Techniques for Variable Selection: Stepwise Selection

- In general, backward and forward selection won't arrive at the same model. They might, but this isn't always the case.
- Forward selection tends to select a more parsimonious model whereas backward selection tends to select a more verbose model.
- A reasonable thought might be: "Why not take both approaches??"
  - This is the rationale behind *stepwise* selection, which begins with the same steps as forward selection (starting from null model)

Compuational Techniques for Variable Selection: Lasso Regression

- Remember the main benefit of ridge regression was that we were able to shrink the variance of the coefficient estimates by sacrificing a little bit of bias
  - The bias-variance tradeoff
- However, in ridge regression, we fit a full regression model despite the fact that not all variables may necessarily be relevant or important to include in the model.
- This is where *lasso* regression comes into play:

Computational Techniques for Variable Selection: Lasso Regression

- The idea with lasso regression is that we calculate our β coefficients similarly to regular least-squares, but we have another biasing/tuning parameter:

$$\hat{\beta}_{\text{Lasso}} = \min \left(\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right)$$

- Where:

$$\sum_{j=1}^{p} |\beta_j| \leq t$$

Computational Techniques for Variable Selection: Lasso Regression

- How is this different than ridge? In ridge, we shrunk our coefficient estimates toward zero through minimizing variance.
- With lasso, our tuning parameter can force some coefficients to be exactly zero, thus taking the benefits of both ridge and variable selection in a single model.
- Thus, it is said the lasso results in a *sparse* model compared to ridge.