Assessing the Additional Contribution of New Predictor Variables

Dr. Austin R. Brown

Kennesaw State University

Introduction

Some of the materials in today's lecture were adapted from those created by Dr. Taasoobshirazi as well as my former professor, Dr. Khalil Shafie (Thank you Drs. S & T!).

Introduction

- In the last couple of classes, we have learned that we can use more than one predictor variable, either categorical or quantitative predictors in a multiple linear regression model.

- In general, when we include additional predictors in a model, we are doing so under the assumption that the additional predcitors are explaining some non-negligible amount of variability in the response variable.

- But how can I quantify the additional contribution some new set of predictor variables are making beyond what my original set made?
  - We have some metrics to help us out!

R^2 and Adjusted R^2

- We learned in the first class that one method for evaluating a model is through a metric called the coefficient of determination or R^2.

- Recall, R^2 is interpreted as the proportion of variability in the response explained by the predictors in the model.

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

R^2 and Adjusted R^2

- However, an issue with using R^2 as a measure of model adequacy in a MLR model is that R^2 will increase with each added predictor regardless of its relevance or explanatory capability.

- Resultingly, it is recommended to use an adjusted R^2 value which penalizes the original R^2 for each additional predictor in the model. However, it has the same interpretation as regular R^2.

$$R^2_{Adj} = 1 - \frac{SSE/(n-(k+1))}{SST/(n-1)}$$

R^2 and Adjusted R^2

- Note:

$$\frac{SSE}{n-k-1} = MSE = \hat{\sigma}^2$$

- and:

$$\frac{SST}{n-1} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = s_y^2$$

- Thus:

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

Testing the Contribution of Subsets of Predictors

- Now, while we can use R^2_{Adj} to see if adding another variable to our MLR model improves the fit, it can't tell us if that difference is *significant*, necessarily.

- A concept we'll talk about later on in principles of model building is that of parsimony, which here means that a less complicated model is preferable to a more complicated model when there isn't much difference between them, fit-wise.

- So for us to justify the addition of more predictor variables, we need to know that they're doing a better job for us than the simpler model.

Testing the Contribution of Subsets of Predictors

- For example, let's consider the `mtcars` dataset. Suppose we wanted to know if rear axle ratio (drat) was sufficient in predicting miles per gallon compared to a model model which contains drat, quarter mile drag time (qsec), and vehicle weight (wt).
- Conceptually, we fit two models, a full model containing all of the variables, and a reduced model, containing only drat. From here we calculate two SSR's, one for the full and one for the reduced (SSR_{Full} and $SSR_{Reduced}$, respectively).
- For the same reason why we use R^2_{Adj}, $SSR_{Full} > SSR_{Reduced}$. But is the difference great enough for us to use the full model?

Testing the Contribution of Subsets of Predictors

- In effect, we're testing:

$$H_0 : \beta_2 = \beta_3 = 0$$
$$H_1 : \beta_j \neq 0, \quad j = 2, 3$$

- Our test statistic is:

$$F_0 = \frac{(SSR_{Full} - SSR_{Reduced})/(r)}{MSE_{Full}} \sim F(r, n-p)$$

- where r is the number of β's in the reduced model and p is the number of β's in the full model.