Homework 9

STAT 625

Due November 8, 2022, 12:30pm

1 Reading

- Read Chapter 11 of the text.

2 Primary Questions (no challenge questions this week):

1. Weisberg problem 10.3 (1 part, counted as 2 parts)
2. Wesiberg problem 10.5 (1 part, counted as 3 parts)
3. Suppose you are interested in looking for gender discrimination in wages. You have data on men, women, and a very small number of people who do not identify as either men or women (non-binary). You do not have enough non-binary people to draw meaningful conclusions about them. First, because gender discrimination is likely to disadvantage both women and non-binary people as compared to men, you decide to combine the women and non-binary respondents as 'non-men.' So you now have a binary variable for gender, with two levels corresponding to 'men' and 'non-men'. You have many variables available for inclusion in the model. Consider each of the following scenarios:
   (a) You use forward stepwise regression. The method selects the predictors 'previous wage' and 'gender'. 'Previous wage' is the wage the person earned 1 year ago. Fitting the model shows that previous wage is highly significant, and the 'gender' term is not significant. Are you satisfied with this selection of model terms? What can you conclude about gender and wages?
   (b) Now suppose you select the model with the best AIC overall. That model includes the terms 'previous wage,' 'years experience,' and 'job title.' 'Gender' is not selected for this model. Does this mean gender is not an important determinant of wages? Explain.
   (c) You use a lasso method to find a reduced set of regressors in the model. The resulting model includes 'years experience' (measured in years). You re-code the 'years experience' variable by dividing by 10 to get 'decades of experience' and re-do the lasso. Now 'decades of experience' is not one of the included model terms. Why would the re-scaled variable behave differently?
   (d) Briefly describe how you would select the regressors to use in this analysis.

3 Extra Credit Questions

(You can get more than full credit for doing this. The others questions are required)

4. (1 part, but counts as 4 parts worth of credit): Design and execute a simple simulation study to show how subset selection over-states significance. You may consider the following steps:
   - Generate 60 X variables of noise (maybe independent N(0,1)). Use a sample size of 100. At the end of this step, you should have a matrix of X variables that is 60 columns by 100 rows.
   - Generate a random Y variable unrelated to X. This should be a vector of length 100.
   - Use a subset selection method to find the best subset of X's.
   - Count the number of X's that seem significant at the nominal $p < .05$ level, and compare this to the number that would be expected by chance if the p-values were valid.
   - Write a function or loop to do the previous steps many times (maybe 100 or 1000 times).
   - Summarize your results using appropriate summary statistics or plots to show that there is a systematic tendency to 'find' more 'significant' variables than you would expect when using a subset selection method.

4 Pre-lecture Check

Complete this week's timed pre-lecture check on **gradescope** by Thursday at 10am.