MSAN 601: Linear Regression Analysis
Homework 3
Due: Wednesday, October 11th at 5:00 PM

Computational Problems

1. **[Subset Selection]** Go through the *Subset Selection Methods* lab in Sections 6.5.1, 6.5.2, and 6.5.3 of the *An Introduction to Statistical Learning* book (available online). For this, please type the code to make sure it works and turn in the code and output as part of your knit file from these problems. Compare the subset of variables that you keep in each strategy, and compare this with which coefficients are non-zero in the LASSO regression fit from class.

2. **[Model Diagnostics]** Using the `sat` dataset from the `faraway` package in R, fit a regression model with the total SAT score as the response and expend, salary, ratio, and takers as predictors. Perform regression diagnostics on this model to answer the following questions. Display any plots that are relevant (but do not include any that you do not need!)

   (a) Write out the fitted model from your regression.

   (b) Check the normality assumption of the model using two different tests and provide a plot that visually inspects this assumption.

   (c) Check for large leverage points and outliers.

   (d) Remove any values that you find to be influential from part (c) and re-fit your regression model. Write out the newly fitted model and comment on any differences between this model and the one from part (a).

   (e) With influential points removed, check the structure of the relationship between the predictors and the response.

3. **[Collinearity]** Using the `divusa` data from the `faraway` package in R, answer the following questions.

   (a) Fit a regression model with divorce as the response and unemployed, femlab, marriage, birth, and military as predictors. Compute the condition numbers and interpret their meaning.

   (b) For the same model as that fit in (a), compute the VIFs. Is there evidence that collinearity causes some predictors not to be significant? Explain.

   (c) Does the removal of insignficant predictors from the model reduce the collinearity?

4. **[Transformations]** Using the `ozone` data in the `faraway` package in R, fit a regression model with O3 as the response and temp, humidity, and ibh as predictors. Check an appropriate plot of the residuals of this fit

to determine whether the relationship between O3 and these predictors is actually linear. Then use the Box-Cox method to determine the best transformation on the response.

Conceptual Problems

1. Consider fitting a model $y = X\beta + \epsilon$ where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

   (a) Let $Z = (X^T X)$. Using the regularization definition of Ridge Regression with tuning parameter λ, show that

   $$\mathbb{E}[\hat{\beta}_{Ridge}] = (I + \lambda Z^{-1})^{-1}\beta$$

   (b) Based on the above result, calculate the bias of $\hat{\beta}_{Ridge}$.

   (c) Show that

   $$\text{Var}(\hat{\beta}_{Ridge}) = \sigma^2(I + \lambda Z^{-1})^{-1}Z^{-1}(I + \lambda Z^{-1})^{-1}$$

   (d) Consider an example where we have a design matrix:

   $$X = \begin{pmatrix} 1 & 0.7 \\ 1 & 0.69 \end{pmatrix}$$

   Using the formula above (and the variance of $\widehat{\hat{\beta}}_{OLS}$), calculate $\text{Var}(\hat{\beta}_{OLS})$ and $\text{Var}(\hat{\beta}_{Ridge})$ for a grid of λ between 0 and 100 when $\text{Var}(\epsilon) = 2$. Why is the variance of the OLS estimators so high?

2. Suppose that we peform best subset, forward stepwise, and backward stepwise selection for linear regression on a single data set. For each approach, we obtain $p+1$ models, containing $0, 1, \ldots, p$ predictors. Explain your answers for each of the following:

   (a) Which of the three methods with k predictors has the smallest *training* MSE?

   (b) Which of the three methods with k predictors has the smallest *test* MSPE?

   (c) (TRUE or FALSE): The predictors in the k-variable model identified by forward stepwise selection are a subset of the predictors in the $(k+1)$-variable model of forward stepwise selection.

   (d) (TRUE or FALSE): The predictors in the k-variable model identified by best subset selection are a subset of the predictors in the $(k+1)$-variable model of best subset selection.

3. In simple linear regression with only a single predictor, the OLS estimators can be written as

   $$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n}(x_i - \overline{x})^2$$
   $$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

Using the above, prove that the point $(\overline{x}, \overline{y})$ is always on the least squares regression line.

4. Suppose we fit a Lasso regression to a data set which has 100 features $(x_1, \ldots, x_{100})$. We now rescale the feature x_1 by multiplying it by 10, and then refit Lasso regression with the same regularization parameter.

   Which of the following option will be correct?

   (a) It is more likely for x_1 to be excluded from the model

   (b) It is more likely for x_1 to be included in the model

   (c) Not possible to say

   (d) None of these

   Explain your reasoning.

5. Suppose that we fit a linear regression model of $\mathbf{y}$ against $\mathbf{x}$ for which we have 100 data points. Call this model 1. Then we duplicate all of the 100 data points and re-fit a linear regression model with these 200 points. Call this model 2. Answer the following:

   (a) Compare the estimated coefficients of model 1 and model 2.

   (b) Compare the MSE of model 1 and model 2.

   (c) Compare the SSE of model 1 and model 2.

   (d) Compare the predictive ability of model 1 and model 2.