Stat 627/427 (Statistical Machine Learning)

Homework # 7: $\begin{cases} \textbf{Multicollinearity (3.3.3, pp. 99-102).} \\ \textbf{Variable Selection and Shrinkage (6.1, 6.2, 6.5, 6.6)} \end{cases}$

Due Sunday ***March 31*** *by 12 noon on Blackboard. The quiz is on Monday* ***April 1****.*

1. **(Page 125, chap. 3, #14)**. This problem focuses on **multicollinearity**.

   (a) Perform the following commands in R:

   ```
   > set.seed (1)
   > x1 = runif (100)
   > x2 = 0.5*x1 + rnorm(100)/10
   > y = 2 + 2*x1 + 0.3*x2 + rnorm(100)
   ```

   The last line corresponds to creating a linear model in which y is a function of x_1 and x_2. Write out the form of the linear model. What are the regression coefficients?

   (b) What is the correlation between x_1 and x_2? Create a scatterplot displaying the relationship between the variables.

   (c) Using this data, fit a least squares regression to predict y using x_1 and x_2. Describe the results obtained. What are $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$? What are the true β_0, β_1, and β_2? Can you reject the null hypothesis $H_0 : \beta_1 = 0$? How about the null hypothesis $H_0 : \beta_2 = 0$?

   (d) Now fit a least squares regression to predict y using only x_1. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

   (e) Now fit a least squares regression to predict y using only x_2. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_2 = 0$?

   (f) Do the results obtained in (c)–(e) contradict each other? Explain your answer.

   (g) Now suppose we obtain one additional observation, which was unfortunately mismeasured. Use the following R code.

   ```
   > x1=c(x1, 0.1)
   > x2=c(x2, 0.8)
   > y=c(y,6)
   ```

   Re-fit the linear models from (c) to (e) using this new data. What effect does this new observation have on the each of the models? In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers. How do the slopes from all the considered models react on the newly added data point?

   (h) What are standard errors of estimated regression slopes in (a), (d), and (e)? Which models produce more stable and therefore, more reliable estimates?

   (i) Compute both VIF in question (a) and relate them to your answer to question (h).

2. **(Chap. 6, # 2, p.259)** Consider three methods of fitting a linear regression model - (a) lasso, (b) ridge regression, and (c) fitting nonlinear trends. For each method, choose the right answer, comparing it with the least squares regression:

   i. The method is more flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.

   ii. The method is more flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.

iii. The method is less flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.

iv. The method is less flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.

3. (**Chap. 6, ≈# 6, p.261**) Ridge regression minimizes

$$\sum_{i=1}^{n}(Y_i - \beta_0 - X_{i1}\beta_1 - \ldots - X_{ip}\beta_p)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{1}$$

whereas lasso minimizes

$$\sum_{i=1}^{n}(Y_i - \beta_0 - X_{i1}\beta_1 - \ldots - X_{ip}\beta_p)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \,. \tag{2}$$

Consider a "toy" example, where $n = p = 1$, $X = 1$, and the intercept is omitted from the model. Then RSS reduces to $RSS = (Y - \beta)^2$.

(a) Choose some Y and λ, plot (1) and (2) as functions of β, and find their minima on these graphs. Verify that these minima are attained at

$$\hat{\beta}_{ridge} = \frac{Y}{1+\lambda} \quad \text{and} \quad \hat{\beta}_{lasso} = \begin{cases} Y - \lambda/2 & \text{if} \quad Y > \lambda/2 \\ Y + \lambda/2 & \text{if} \quad Y < -\lambda/2 \\ 0 & \text{if} \quad |Y| \leq \lambda/2 \end{cases} \tag{3}$$

(b) Now choose some value of Y and plot ridge regression and lasso solutions (3) on the same axes, as functions of λ. Observe how ridge regression keeps a slope whereas lasso sends the slope to 0 when the penalty term is high.

4. (**Simulation project - Chap. 6, # 8, p.262**)

In this exercise, we will generate simulated data, and will then use this data to perform best subset selection.

(a) Use the **rnorm()** function to generate a predictor X and a noise vector ε of length $n = 100$ (you can refer to our lab "First steps in R" for this command).

(b) Generate a response vector Y according to the model

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon,$$

where β_0, β_1, β_2, and β_3 are constants of your choice.

(c) Use stepwise selection with **step** for variable selection. How does your answer compare to the results in (c)?

(d) Now fit a lasso model with the same predictors. Use cross-validation to select the optimal value of λ. Create plots of the cross-validation error as a function of λ. Report the resulting coefficient estimates, and discuss the results obtained. Which predictors got eliminated by lasso?

(e) Now generate a response vector Y according to the model

$$Y = \beta_0 + \beta_7 X^7 + \varepsilon,$$

and perform best subset selection and the lasso. Discuss the results.