STAT 435
SPRING QUARTER 2018

Homework # 5
Due Friday, May 11, 2018 at 12:00 PM (Noon)
Online Submission Via Canvas

Instructions: You may discuss the homework problems in small groups, but you must write up the final solutions and code yourself. Please turn in your code for the problems that involve coding. However, for the problems that involve coding, you must also provide written answers: you will receive no credit if you submit code without written answers. You might want to use Rmarkdown to prepare your assignment.

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

   (a) Use the `rnorm()` function to generate a predictor X of length $n = 100$, and a noise vector ϵ of length $n = 100$.

   (b) Generate a response vector Y of length $n = 100$ according to the model

   $$Y = 3 - 2X + X^2 + \epsilon.$$

   (c) Use the `regsubsets()` function to perform best subset selection, considering $X, X^2, \ldots, X^7$ as candidate predictors. Make a plot like Figure 6.2 in the textbook. What is the overall best model according to C_p, BIC, and adjusted R^2? Report the coefficients of the best model obtained. Comment on your results.

   (d) Repeat (c) using forward stepwise selection instead of best subset selection.

   (e) Repeat (c) using backward stepwise selection instead of best subset selection.

   Hint: You may need to use the `data.frame()` *function to create a single data set containing both X and Y.*

2. In class, we discussed the fact that if you choose a model using stepwise selection on a data set, and then fit the selected model using least squares on the same data set, then the resulting p-values output by `R` are highly misleading. We'll now see this through simulation.

(a) Use the `rnorm()` function to generate vectors $X_1, X_2, \ldots, X_{100}$ and ϵ, each of length $n = 1000$. *(Hint: use the* `matrix()` *function to create a* 1000×100 *data matrix.)*

(b) Generate data according to

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{100} X_{100} + \epsilon,$$

where $\beta_1 = \ldots = \beta_{100} = 0$.

(c) Fit a least squares regression model to predict Y using $X_1, \ldots, X_p$. Make a histogram of the p-values associated with the null hypotheses $H_{0j} : \beta_j = 0$ for $j = 1, \ldots, 100$.

Hint: You can easily access these p-values using the command `(summary(lm(y~X)))$coef[,4]`.

(d) Recall that under $H_{0j} : \beta_j = 0$, we expect the p-values to have a Unif$[0, 1]$ distribution. In light of this fact, comment on your results in (c). Do any of the features appear to be significantly associated with the response?

(e) Perform forward stepwise selection in order to identify $\mathcal{M}_2$, the best two-variable model. (For this problem, there is no need to calculate the best model $\mathcal{M}_k$ for $k \neq 2$.) Then fit a least squares regression model to the data, using just the features in $\mathcal{M}_2$. Comment on the p-values obtained for the coefficients.

(f) Now generate another 1000 observations by repeating the procedure in (a) and (b). Using the new observations, fit a least squares linear model to predict Y using just the features in $\mathcal{M}_2$ calculated in (e). *(Do not perform forward stepwise selection again using the new observations! Instead, take the* $\mathcal{M}_2$ *obtained earlier in this problem.)* Comment on the p-values for the coefficients. How do they compare to the p-values in (e)?

(g) Are the features in $\mathcal{M}_2$ significantly associated with the response? Justify your answer.

THE BOTTOM LINE: If you showed a friend the p-values obtained in (e), without explaining that you obtained $\mathcal{M}_2$ *by performing forward stepwise selection* ***on this same data****, then he or she might* ***incorrectly*** *conclude that the features in* $\mathcal{M}_2$ *are highly associated with the response.*

3. Let's consider doing least squares and ridge regression under a very simple setting, in which $p = 1$, and $\sum_{i=1}^n y_i = \sum_{i=1}^n x_i = 0$. We consider regression without an intercept. (It's usually a bad idea to do regression without an intercept, but if our feature and response each have mean zero, then it is okay to do this!)

(a) The least squares solution is the value of $\beta \in \mathbb{R}$ that minimizes

$$\sum_{i=1}^{n} (y_i - \beta x_i)^2.$$

Write out an analytical (closed-form) expression for this least squares solution. Your answer should be a function of $x_1, \dots, x_n$ and $y_1, \dots, y_n$. *Hint: Calculus!!*

(b) For a given value of λ, the ridge regression solution minimizes

$$\sum_{i=1}^{n} (y_i - \beta x_i)^2 + \lambda \beta^2.$$

Write out an analytical (closed-form) expression for the ridge regression solution, in terms of $x_1, \dots, x_n$ and $y_1, \dots, y_n$ and λ.

(c) Suppose that the true data-generating model is

$$Y = 3X + \epsilon,$$

where ϵ has mean zero, and X is fixed (non-random). What is the expectation of the least squares estimator from (a)? Is it biased or unbiased?

(d) Suppose again that the true data-generating model is $Y = 3X + \epsilon$, where ϵ has mean zero, and X is fixed (non-random). What is the expectation of the ridge regression estimator from (b)? Is it biased or unbiased? Explain how the bias changes as a function of λ.

(e) Suppose that the true data-generating model is $Y = 3X + \epsilon$, where ϵ has mean zero and variance σ^2, and X is fixed (non-random), and also $\text{Cov}(\epsilon_i, \epsilon_{i'})= 0$ for all $i \neq i'$. What is the variance of the least squares estimator from (a)?

(f) Suppose that the true data-generating model is $Y = 3X + \epsilon$, where ϵ has mean zero and variance σ^2, and X is fixed (non-random), and also $\text{Cov}(\epsilon_i, \epsilon_{i'})= 0$ for all $i \neq i'$. What is the variance of the ridge estimator from (b)? How does the variance change as a function of λ?

(g) In light of your answers to parts (d) and (f), argue that λ in ridge regression allows us to control model complexity by trading off bias for variance.

Hint: For this problem, you might want to brush up on some basic properties of means and variances! For instance, if $Cov(Z, W) = 0$, then $Var(Z + W) = Var(Z) + Var(W)$. And if a is a constant, then $Var(aW) = a^2 Var(W)$, and $Var(a + W) = Var(W)$.

4. Suppose that you collect data to predict Y (height in inches) using X (weight in pounds). You fit a least squares model to the data, and you get

$$\hat{Y} = 3.1 + 0.57X.$$

(a) Suppose you decide that you want to measure weight in ounces instead of pounds. Write out the least squares model for predicting Y using $\tilde{X}$ (weight in ounces). (You should calculate the coefficient estimates explicitly.) *Hint: there are 16 ounces in a pound!*

(b) Consider fitting a least squares model to predict Y using X and $\tilde{X}$. Let β denote the coefficient for X in the least squares model, and let $\tilde{\beta}$ denote the coefficient for $\tilde{X}$. Argue that any equation of the form

$$\hat{Y} = 3.1 + \beta X + \tilde{\beta}\tilde{X},$$

where $\beta + 16\tilde{\beta} = 0.57$, is a valid least squares model.

(c) Suppose that you use ridge regression to predict Y using X, using some value of λ, and obtain the fitted model

$$\hat{Y} = 3.1 + 0.4X.$$

Now consider fitting a ridge regression model to predict Y using $\tilde{X}$, again using that same value of λ. Will the coefficient of $\tilde{X}$ be equal to 0.4/16, greater than 0.4/16, or less than 0.4/16? Explain your answer.

(d) For the same value of λ considered in (c), suppose you perform ridge regression to predict Y using X, and separately you perform ridge regression to predict Y using $\tilde{X}$. Which fitted model will have smaller residual sum of squares (on the training set)? Explain your answer.

(e) Finally, suppose you use ridge regression to predict Y using X and $\tilde{X}$, using some value of λ (not necessarily the same value of λ used in (d)), and obtain the fitted model

$$\hat{Y} = 3.17 + 0.03X + 0.03\tilde{X}.$$

Is the following claim true or false? Explain your answer.
Claim: Any equation of the form

$$\hat{Y} = 3.17 + \beta X + \tilde{\beta}\tilde{X},$$

where $\beta + 16\tilde{\beta} = 0.03 + 16 \times 0.03 = 0.51$, *is a valid ridge regression solution for that value of* λ.

(f) Argue that your answers to the previous sub-problems support the following claim:
Claim: least squares is scale-invariant, but ridge regression is not.

5. Suppose we wish to fit a linear regression model using least squares. Let $\mathcal{M}_k^{BSS}, \mathcal{M}_k^{FWD}, \mathcal{M}_k^{BWD}$ denote the best k-feature models in the best subset, forward stepwise, and backward stepwise selection procedures. (For notational details, see Algorithms 6.1, 6.2, and 6.3 of the textbook.)

Recall that the training set residual sum of squares (or RSS for short) is defined as $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

For each claim, fill in the blank with one of the following: "less than", "less than or equal to", "greater than", "greater than or equal to", "equal to". Say "not enough information to tell" if it is not possible to complete the sentence as given. Explain each of your answers.

(a) *Claim: The RSS of* $\mathcal{M}_1^{FWD}$ *is* ________________ *the RSS of* $\mathcal{M}_1^{BWD}$.

(b) *Claim: The RSS of* $\mathcal{M}_0^{FWD}$ *is* ________________ *the RSS of* $\mathcal{M}_0^{BWD}$.

(c) *Claim: The RSS of* $\mathcal{M}_1^{FWD}$ *is* ________________ *the RSS of* $\mathcal{M}_1^{BSS}$.

(d) *Claim: The RSS of* $\mathcal{M}_2^{FWD}$ *is* ________________ *the RSS of* $\mathcal{M}_1^{BSS}$.

(e) *Claim: The RSS of* $\mathcal{M}_1^{BWD}$ *is* ________________ *the RSS of* $\mathcal{M}_1^{BSS}$.

(f) *Claim: The RSS of* $\mathcal{M}_p^{BWD}$ *is* ________________ *the RSS of* $\mathcal{M}_p^{BSS}$.

(g) *Claim: The RSS of* $\mathcal{M}_{p-1}^{BWD}$ *is* ________________ *the RSS of* $\mathcal{M}_{p-1}^{BSS}$.

(h) *Claim: The RSS of* $\mathcal{M}_4^{BWD}$ *is* ________________ *the RSS of* $\mathcal{M}_4^{BSS}$.

(i) *Claim: The RSS of* $\mathcal{M}_4^{BWD}$ *is* ________________ *the RSS of* $\mathcal{M}_4^{FWD}$.

(j) *Claim: The RSS of* $\mathcal{M}_4^{BWD}$ *is* ________________ *the RSS of* $\mathcal{M}_3^{BWD}$.

6. *This problem is extra credit!!!!* Let $\mathbf{y}$ denote an n-vector of response values, and let $\mathbf{X}$ denote an $n \times p$ design matrix. We can write the ridge regression problem as

$$\text{minimize}_{\beta \in \mathbb{R}^p} \left\{ \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \|\beta\|^2 \right\},$$

where we are omitting the intercept for convenience. Derive an analytical (closed-form) expression for the ridge regression estimator. Your answer should be a function of $\mathbf{X}$, $\mathbf{y}$, and λ.