SS3859A/9859A Fall 2019

Assignment 3

Due: Nov 13 by 11:59 p.m.

Total: 20 pt (1 pt for each question)

Note: Please use either MS word or PDF format for your submission.

1. Discuss the following statements and explain why they are true or false.

(a) Increasing the number of predictor variables will never decrease the R^2.

(b) Multicollinearity affects the intrepretation of the regression coefficients.

(c) The variance inflation factor of $\hat{\beta}_j$ depends on the R^2 of the regression of the response variable y on the predictor variable x_j.

(d) A high leverage point is always highly influential.

(e) All criteria for the selection of the best regression equation lead to the same set of predictor variables.

2. For question 2, import the data from
`https://raw.githubusercontent.com/hgweon2/ss3859/master/hw3-data.txt`
The imported dataset contains 200 observations with 3 variables: y, x1 and x2.

(a) Plot a scatterplot matrix and briefly discuss the relationships between the variables.

(b) Obtain the fitted model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$. Check the model assumptions using appropriate graphical and testings approaches.

(c) Was there any influential point? Use Cook's distance with threshold $= 4/n$. Report the indices of the influential points.

(d) Among the influential points, how many of them are also considered outliers (whose absolute standardized residuals are greater than 2)?

(e) Suppose that the influential points identified in (c) were simple measurement errors. Remove the influential points from the data and repeat (b) using the new data set. Was the removal of the influential points useful for correcting the model assumptions?

(f) Use the Box-Cox method to determine the best transformation on the response variable y (use all 200 observations). Repeat (b) using the transformed y. (For λ, use a reasonable value near the optimal value). Was this transformation helpful for correcting the model assumptions?

(g) This time, obtain the polynomial model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1^2 + \hat{\beta}_4 x_2^2$. Is this polynomial model preferable to the resulting models in (b) and (f)? Justify your answer.

(h) Add the cubic terms to the model. That is, obtain $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1^2 + \hat{\beta}_4 x_2^2 + \hat{\beta}_5 x_1^3 + \hat{\beta}_6 x_2^3$. Would this cubic model be preferred to the quadratic one in (g)? Justify your answer.

3. For question 3, you will use the mtcars data in R.

(a) Fit a regression model (model_a) using **mpg as the response** and **cyl, disp, hp, wt and drat as predictors** (Do not include and polynomial or interaction terms). Obtain the Variance Inflation Factor (VIF) for each predictor. (You may use the "vif" function in the faraway package.) Does any collinearity exist? Report all predictors whose VIF are higher than 10. Briefly explain how collinearity affects in the regression analysis.

(b) From the result in (a), remove the predictor with the highest VIF value and fit another regression modelusing the rest of the predictors. Obtain the Variance Inflation Factor (VIF) for each predictor used for the model. **This time, do not use any built-in function in R to compute the VIF values. (You can still use the lm function/object.)** Does any collinearity exist? Report all predictors whose VIF are higher than 10.

(c) Considering model_a the full model, find the best subset of predictors to predict mpg (use the forward stepwise selection approach with AIC).

(d) Considering model_a the full model, find the best subset of predictors to predict mpg (use the backward selection approach with BIC). Is the resulting model is significantly different from the model obtained in (c)? Use the significance level 0.05.

4. For question 4, you will use the *prostate* data in the *faraway* package. Consider the following three models.

```
(Model A) Response: lpsa    Predictors: lcavol, lweight, svi

(Model B) Response: lpsa    Predictors: lcavol, lweight, svi, lbph

(Model C) Response: lpsa    Predictors: lcavol, lweight, svi, lbph, lcp, gleason
```

Do not use polynomial or interaction terms.

(a) Find the best model in terms of AIC, BIC and adjusted R^2, respectively.

(b) Find the best model in terms of $PRESS$.

(c) Find the best model using R^2 as the quality criterion. Explain why R^2 is not an appropriate measure for model comparison.