Homework #6

Please submit the solution in the form of R Markdown report, knitted into either of the available formats (HTML, pdf or Word). Provide all relevant code and output. Goal of this homework is to have practice 1) multiple linear regression; 2) quality of fit measures (R^2, RSE); 3) collinearity; 4) variable selection; 4) R coding.

Problem #1

This question involves the use of multiple linear regression on the *Auto* data set of *ISLR* library.

Use the *lm*() function to perform a multiple linear regression with *mpg* as the response and all other variables (except *name*) as the predictors. Use the *summary*() function to print the results.

1. Formulate the H_0 and H_a hypotheses (using parameter notation) for testing whether the overall model is significant. Which part of *summary*() output corresponds to this test? Is the model significant?
2. Which predictors appear to have a statistically significant relationship to the response? Just list them.
3. Interpret the effect of car's weight on its miles per gallon.
4. For the effect from part 3, proceed to report and interpret the 95% confidence interval.
5. Report and interpret both quality-of-fit metrics.

Problem #2

This problem will deal with *cystfibr* data example of *ISwR* package. In particular, we will be building a model to predict *pemax* (patient's maximum respiratory pressure) based on other physical characteristics.

1. Proceed to fit the following multiple linear regression model:

$$pemax \sim .$$

   a. Comment on the 1) overall model significance; 2) significance of any individual predictors. Why do you think this is happening (name the main issue)?
   b. Proceed to address the issue observed in part (a) via studying a correlation matrix of predictors, modifying the model accordingly. Fit the modified model, comment on its 1) overall model significance; 2) significance of any individual predictors.
   c. Proceed to address the issue observed in part (a) via using VIF criteria method. Fit the modified model, comment on its 1) overall model significance; 2) significance of any individual predictors.
   d. In you own words, why does collinearity prevent us from accurately estimating effects of collinear predictors on the response variable?

2. Proceed to fit the following two models:

   - Full model: $pemax \sim sex + weight + height + rv + frc$
   - Reduced model: $pemax \sim sex + height + frc$

   a. Comment on what happens to standard errors for $\hat{\beta}_{height}$ and $\hat{\beta}_{frc}$ coefficients when going from the full to reduced model. Why does this happen?

   b. Proceed to use VIF criteria in order to get from full model down to the reduced model. Which variable is dropped first? Second? Why?

Problem #3

For *Auto* data set from *ISLR* library. Proceed to conduct variable selection via backward AIC approach:

1. Which *R* function allows us to do that? Which variable(s) ended up being dropped from the model?

2. Explain what is meant by "Df", "Sum of Sq", "RSS" and "AIC" in the tables outputted by *step*() function.

3. Explain why the algorithm stopped (!) on that particular subset of variables. *Hint*: What does "<none>" represent? Why is it of interest?)

Problem #4 (BONUS)

This question involves the use of linear regression on the *Advertising* data set. Proceed to calculate RSS and R^2 for the following three models:

- $Sales \sim TV$
- $Sales \sim TV + radio$
- $Sales \sim TV + radio + newspaper$

1. Did the RSS decrease (or at least didn't increase) every time you added an extra variable? Why do you think that is?

2. Did the R^2 increase (or at least didn't decrease) every time you added an extra variable? Why do you think that is? *Hint*: use the definition of R^2 + part (a).