STA 5207: Homework 7

Due: March, 8th by 11:59 PM

Include your R code in an R chunks as part of your answer. In addition, your written answer to each exercise should be self-contained so that the grader can determine your solution without reading your code or deciphering its output.

Exercise 1 (`longley` Macroeconomic Data) [50 points]

For this exercise we will use the built-in `longley` data set. You can also find the data in `longley.csv` on Canvas. The data set contains macroeconomic data for predicting unemployment. The variables in the model are

- `GNP.deflator`: GNP implicit price deflator (1954 = 100)
- `GNP`: Gross national product.
- `Unemployed`: Number of unemployed.
- `Armed.Forces`: Number of people in the armed forces.
- `Population`: 'noninstituionalized population $\geq$ 14 years of age.
- `Year`: The year.
- `Employed`: Number of people employed.

In the following exercise, we will model the `Employed` variable.

1. (6 points) How many pairs of predictors are highly correlated? Consider "highly" correlated to be a sample correlation above 0.7. What is the largest correlation between any pair of predictors in the data set?
2. (6 points) Fit a model with `Employed` as the response and the remaining variables as predictors. Give the condition number. Does multicollinearity appear to be a problem?
3. (6 points) Calculate and report the variance inflation factor (VIF) for each of the predictors. Which variable has the largest VIF? Do any of the VIFs suggest multicollinearity?
4. (6 points) What proportion of the observed variation in `Population` is explained by the linear relationship with the other predictors? Are there any variables that are nearly orthogonal to the others? Consider a low R_k^2 to be less than 0.3.
5. (6 points) Give the condition indices. How many near linear-dependencies are likely causing most of the problem?
6. (10 points) Fit a new model with `Employed` as the the response and the predictors from the model in part 2 that were significant (use $\alpha = 0.05$). Calculate and report the variance inflation factor for each of the predictors. Do any of the VIFs suggest multicollinearity?
7. (10 points) Use an F-test to compare the models in parts 2 and 6. Report the following:
    - The null hypothesis.
    - The test statistic.
    - The p-value of the test.
    - A statistical decision at $\alpha = 0.05$.
    - Which model do you prefer, the model from part 2 or 6.

Exercise 2 (The `sat` Data Set Revisited) [50 points]

For this exercise we will use the `sat` data set from the `faraway` package, which you analyzed in Homework #3. In the following exercise, we will model the `total` variable as a function of `expend`, `salary`, and `ratio`.

1. (8 points) Among the three predictors `expend`, `salary`, and 'ratio", how many pairs of predictors are are highly correlated? Consider "highly" correlated to be a sample correlation above 0.7.

2. (8 points) Fit a model with `total` as the response and `expend`, `salary`, and `ratio` as the predictors. Give the condition number. Does multicollinearity appear to be a problem?

3. (8 points) Calculate and report the variance inflation factor (VIF) for each of the predictors. Which variable has the largest VIF? Do any of the VIFs suggest multicollinearity?

4. (10 points) Fit a new model with `total` as the response and `ratio` and the sum of `expend` and `salary` – that is `I(expend + salary)` – as the predictors. Note that `expend` and `salary` have the same units (thousands of dollars), so adding them makes sense. Calculate and report the variance inflation factor for each of the two predictors. Do any of the VIFs suggest multicollinearity?

5. (6 points) Conduct a t-test at the 5% significance level for each slope parameter for the model in part 4. Give the test statistic, p-value, and statistical decision for each test.

6. (10 points) Use an F-test to compare the models in parts 2 and 4. Report the following:

   - The null hypothesis (**Hint**: We are testing a linear constraint, see the slides on MLR, page 39).
   - The test statistic.
   - The p-value of the test.
   - A statistical decision at $\alpha = 0.05$.
   - Which model do you prefer, the model from part 2 or part 4.