STA 5207: Homework 6

Due: Wednesday, October 25th by 11:59 PM

Include your R code as an appendix at the end of your homework. Do not include your code in your answers unless the question explicitly tells you to include your code. Your answers to each exercise should be self-contained without code 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?

–There are a total of six highly correlated pairs of predictors in the dataset:

1. (Population, GNP.deflator)
2. (Population, GNP)
3. (Population, Year)
4. (GNP.deflator, GNP)
5. (GNP.deflator, Year)
6. (GNP, Year)

–The highest correlation between any pair of predictors in the dataset is 0.995, observed between GNP and Year.

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?

–The condition number, 43275, is significantly larger than the recommended threshold of 30, indicating a potential issue with multicollinearity that warrants attention.

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?

```
## GNP.deflator          GNP   Unemployed Armed.Forces   Population
Year
##    135.53244   1788.51348     33.61889      3.58893    399.15102
758.98060
```

–The variable exhibiting the highest VIF is GNP, registering a substantial value of 1788.51348. With the exception of Armed forces, all other predictors possess a VIF value exceeding 5, indicating the presence of 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.

– The provided predictor accounts for 99.75% of the variability in the response.

– No variables demonstrate near orthogonality to one another.

5. (6 points) Give the condition indices. How many near linear-dependencies are likely causing most of the problem?

```
##   Eigenvalue Condition Index
## 1      6.861           1.000
## 2      0.082           9.142
## 3      0.046          12.256
## 4      0.011          25.337
## 5      0.000         230.424
## 6      0.000        1048.080
## 7      0.000       43275.043
```

–The presence of three linear dependencies likely accounts for the majority of the issue, given that the values at indices 5, 6, and 7 exceed 30.

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?

```
##   Unemployed Armed.Forces         Year
##     3.317929     2.223317     3.890861
```

–Since all the variance inflation factors (VIFs) are below 5, it can be inferred that multicollinearity is not an issue for this model.

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.

–The null hypothesis is as follows:

$H_0 : \beta_{GNP.deflator} = \beta_{GNP} = \beta_{Population} = 0$

–The calculated test statistic for this hypothesis is 1.7465, and the corresponding P-value is 0.227.

–Using a significance level (α) of 0.05 for our statistical decision, we find that the obtained P-value (0.227) is greater than 0.05. Therefore, we do not reject the null hypothesis.

–In practical terms, this means that there is not enough statistical evidence to conclude that at least one of the predictors (GNP.deflator, GNP, and Population) has a significant linear relationship with the response variable (Employed), given that the other predictors are included in the model.

–As a result, it may be more appropriate to prefer the model from part 6, which is represented as "Employed ~ Unemployed + Armed.Forces + Year," as it appears to provide a better fit or explanation of the data compared to the model that includes GNP.deflator, GNP, and Population as predictors.

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 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.

–There exists only one pair of predictors showing a strong correlation:

– (Expend, Salary) with a correlation coefficient of 0.87.

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?

–The condition number, 48.122, significantly exceeds 30, suggesting a notable concern regarding multicollinearity.

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?

```
##   expend   salary    ratio
## 9.387552 8.095274 2.285359
```

–The variable with the highest VIF is “expend,” which has a value of 9.387. Additionally, besides the ratio, the other two predictors, “expend” and “salary,” have VIF values greater than 5, indicating the presence of 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?

```
##              ratio I(expend + salary)
##           1.005151           1.005151
```

–The predictors exhibit VIF values below 5, indicating the absence of 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.

For the variable β_{ratio}:

- The computed test statistic is 0.382.
- The corresponding p-value is 0.70399.
- Consequently, at a significance level of $\alpha = 0.05$, the null hypothesis is not rejected.
- This suggests that there is no substantial linear relationship between the variable ‘ratio’ and ‘total,’ considering the presence of other predictors in the model.

Regarding the variable $\beta_{expend+salary}$:

- The computed test statistic is -3.305.
- The corresponding p-value is 0.00182.
- Consequently, at a significance level of $\alpha = 0.05$, the null hypothesis is rejected.
- Therefore, there exists a significant linear relationship between the combined effect of ‘expend’ and ‘salary’ and ‘total,’ taking into account the presence of other predictors in the model.

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.

– Null Hypothesis:

$H_0 : \beta_{expend} = \beta_{salary}$

– Test Statistic Value: 0.911

– P-value of the Test: 0.3448

– Statistical Decision at $\alpha = 0.05$: Since the obtained P-value (0.3448) is greater than 0.05, we fail to reject the null hypothesis.

– Conclusively, it appears that the effects of "expend" and "salary" are approximately equal when considering the other predictors in the model. Consequently, we can favor a model that treats "expend" and "salary" equally, such as the model described in part 4 [total ~ ratio + I(expend + salary)].

Code Appendix

```
## Solution for Ex 1 Q 1

# Loading dataset

library(readr)
data(longley, package='faraway')

library(dplyr)

long_prd = dplyr::select(longley,-Employed)

round(cor(long_prd),3)



## Solution for Ex 1 Q 1

##install.packages("corrplot")

library(corrplot)

corrplot(cor(long_prd),method = 'color', order = 'hclust',  diag = FALSE,
         number.digits = 3, addCoef.col = 'black', tl.pos= 'd', cl.pos ='r')

## Solution for Ex 1 Q 2



library(olsrr)

long_model = lm(Employed~.,data=longley)

round(ols_eigen_cindex(long_model)[,1:2],4)
```

```r
## Solution for Ex 1 Q 3
library(faraway)
long_model = lm(Employed~.,data=longley)
vif(long_model)

## Solution for Ex 1 Q 4
population_model = lm(Population ~ .,data=longley)
summary(population_model)$r.squared
1-1/vif(population_model)

## Solution for Ex 1 Q 5
library(olsrr)
long_model = lm(Employed~.,data=longley)
round(ols_eigen_cindex(long_model)[,1:2],3)

## Solution for Ex 1 Q 6
long_model = lm(Employed~.,data=longley)
summary(long_model)

## Solution for Ex 1 Q 6

# fitting a new model with significant predictors
latest_model = lm(Employed~Unemployed+Armed.Forces+Year,data=longley)
vif(latest_model)
```

```r
## Solution for Ex 1 Q 7

restricted_model=lm(Employed~Unemployed+Armed.Forces+Year,data=longley)
full_model=lm(Employed~GNP.deflator+GNP+Unemployed+Armed.Forces+Population+Ye
ar,
              data=longley)

anova(restricted_model,full_model)


## Solution for Ex 2 Q 1

# Loading dataset

data(sat, package='faraway')

total_model = dplyr::select(sat,expend,salary,ratio)

round(cor(total_model),3)



## Solution for Ex 2 Q 1


corrplot(cor(total_model),method = 'color', order = 'hclust',  diag = FALSE,
         number.digits = 3, addCoef.col = 'black', tl.pos= 'd', cl.pos ='r')

## Solution for Ex 2 Q 2


full_model = lm(total~expend+salary+ratio,data=sat)

round(ols_eigen_cindex(full_model)[,1:2],4)

## Solution for Ex 2 Q 3


full_model = lm(total~expend+salary+ratio,data=sat)
vif(full_model)



## Solution for Ex 2 Q 4

rest_model = lm(total ~ ratio + I(expend+salary),data=sat)
```

```
vif(rest_model)


## Solution for Ex 2 Q 5

summary(rest_model)


## Solution for Ex 2 Q 6
rest_model = lm(total ~ ratio + I(expend+salary),data=sat)
full_model = lm(total~expend+salary+ratio,data=sat)

anova(rest _model,full_model)
```