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?

```r
data <- longley
preds <- dplyr::select(data, -Employed)
round(cor(preds), 3)
```

```
##              GNP.deflator   GNP Unemployed Armed.Forces Population  Year
## GNP.deflator        1.000 0.992      0.621        0.465      0.979 0.991
## GNP                 0.992 1.000      0.604        0.446      0.991 0.995
## Unemployed          0.621 0.604      1.000       -0.177      0.687 0.668
## Armed.Forces        0.465 0.446     -0.177        1.000      0.364 0.417
## Population          0.979 0.991      0.687        0.364      1.000 0.994
## Year                0.991 0.995      0.668        0.417      0.994 1.000
```

   Answer: Highly correlated pairs of predictors (predictors with a sample correlation above 0.7) are GNP and GNP.deflator, Population and GNP.deflator, Year and GNP.deflator, Population and GNP, Year and GNP, and Year and Population. The largest correlation between any pair of predictors in the dataset is the correlation between Year and GNP with a correlation of .995.

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

```r
library(olsrr)
```

```
##
## Attaching package: 'olsrr'

## The following object is masked from 'package:datasets':
##
##     rivers
```

```r
model <- lm(Employed ~., data=data)
summary(model)
```

```
##
## Call:
## lm(formula = Employed ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41011 -0.15767 -0.02816  0.10155  0.45539
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.482e+03  8.904e+02  -3.911 0.003560 **
## GNP.deflator  1.506e-02  8.492e-02   0.177 0.863141
## GNP          -3.582e-02  3.349e-02  -1.070 0.312681
## Unemployed   -2.020e-02  4.884e-03  -4.136 0.002535 **
## Armed.Forces -1.033e-02  2.143e-03  -4.822 0.000944 ***
## Population   -5.110e-02  2.261e-01  -0.226 0.826212
## Year          1.829e+00  4.555e-01   4.016 0.003037 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3049 on 9 degrees of freedom
## Multiple R-squared:  0.9955, Adjusted R-squared:  0.9925
## F-statistic: 330.3 on 6 and 9 DF,  p-value: 4.984e-10
```

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

```
##   Eigenvalue Condition Index
## 1     6.8614          1.0000
## 2     0.0821          9.1417
## 3     0.0457         12.2557
## 4     0.0107         25.3366
## 5     0.0001        230.4239
## 6     0.0000       1048.0803
## 7     0.0000      43275.0435
```

Answer: The condition number is 43275.04. This is much greater than 30, so we say that multicollinearity appears 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?

```r
library(faraway)
```

```
##
## Attaching package: 'faraway'

## The following object is masked from 'package:olsrr':
##
##     hsb
```

```r
vif(model)
```

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

Answer: The VIFs for each predictor are as follows: GNP.deflator = 135.5, GNP = 1788.5, Unemployed = 33.6, Armed.Forces = 3.6, Population = 399.2, Year = 759. The variable with the largest VIF is GNP with a VIF of 1788.5. A VIF greater than 5 suggests colinearity, so all predictor's except for Armed.Force's VIF values 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.

```r
summary(lm(Population~GNP.deflator+GNP+Unemployed+Armed.Forces+Year, data=data))
```

```
##
## Call:
## lm(formula = Population ~ GNP.deflator + GNP + Unemployed + Armed.Forces +
##     Year, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.57524 -0.18536  0.07539  0.24615  0.58666
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.618e+03  1.136e+03   1.424 0.184790
## GNP.deflator -2.476e-01  8.932e-02  -2.772 0.019720 *
## GNP           1.234e-01  2.591e-02   4.765 0.000763 ***
## Unemployed    1.638e-02  4.454e-03   3.678 0.004261 **
## Armed.Forces  1.791e-03  2.943e-03   0.608 0.556517
## Year         -7.820e-01  5.872e-01  -1.332 0.212452
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4264 on 10 degrees of freedom
## Multiple R-squared:  0.9975, Adjusted R-squared:  0.9962
## F-statistic: 796.3 on 5 and 10 DF,  p-value: 1.154e-12
```

```r
1-1/vif(model)
```

```
## GNP.deflator          GNP   Unemployed Armed.Forces   Population         Year
##    0.9926217    0.9994409    0.9702548    0.7213654    0.9974947    0.9986824
```

Answer: The proportion of the observed variation in Population that is explained by the linear relationship with the other predictors is 99.75%. There are no variables that are nearly orthogonal to the others.

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

```r
library(olsrr)
round(ols_eigen_cindex(model), 3)
```

```
##   Eigenvalue Condition Index intercept GNP.deflator   GNP Unemployed
## 1      6.861           1.000         0        0.000 0.000      0.000
## 2      0.082           9.142         0        0.000 0.000      0.014
## 3      0.046          12.256         0        0.000 0.000      0.001
## 4      0.011          25.337         0        0.000 0.001      0.065
## 5      0.000         230.424         0        0.457 0.016      0.006
## 6      0.000        1048.080         0        0.505 0.328      0.225
## 7      0.000       43275.043         1        0.038 0.655      0.689
##   Armed.Forces Population Year
## 1        0.000      0.000    0
## 2        0.092      0.000    0
## 3        0.064      0.000    0
## 4        0.427      0.000    0
## 5        0.115      0.010    0
## 6        0.000      0.831    0
## 7        0.302      0.160    1
```

Answer: There are 3 indexes in which the condition index is greater than 30. Therefore we can conclude that there are three linear-dependencies that 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?

```r
new_model <- lm(Employed ~ Unemployed + Armed.Forces + Year, data=data)
vif(new_model)
```

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

Answer: We choose Unemplyed, Armed.Forces, and Year as our predictors because they all had a p-value of less than 0.05 in the linear model from part 2. All VIFs are less than 5, so we conclude that collinearity is not a problem 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.

```r
anova(new_model, model)
```

```
## Analysis of Variance Table
##
## Model 1: Employed ~ Unemployed + Armed.Forces + Year
## Model 2: Employed ~ GNP.deflator + GNP + Unemployed + Armed.Forces + Population +
##     Year
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1     12 1.32336
## 2      9 0.83642  3   0.48694 1.7465  0.227
```

Answer: The null hypothesis is that the full model does not provide a significantly better fit to the data than the restricted model. The test-statistic for this test is 1.7465, and the p-value is 0.227. Since the p-value is not less than our significance level of 0.05, we do not reject the null hypothesis and thus conclude that the full model does not provide a significantly better fit to the data than the restricted model. Therefore, we would prefer the model from part 6 over the model from part 2.

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.

```r
library(faraway)
data <- dplyr::select(sat, total, expend, salary, ratio)
preds <- dplyr::select(data, -total)
round(cor(preds), 3)
```

```
##        expend salary  ratio
## expend  1.000  0.870 -0.371
## salary  0.870  1.000 -0.001
## ratio  -0.371 -0.001  1.000
```

Answer: Using 0.7 as highly correlated, we see that the only pair of predictors which are highly correlated are salary and expend with a correlation score of 0.870.

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?

```r
model <- lm(total ~ ., data=data)
round(ols_eigen_cindex(model)[,1:2],4)
```

```
##   Eigenvalue Condition Index
## 1     3.9393          1.0000
## 2     0.0516          8.7394
## 3     0.0074         23.1080
## 4     0.0017         48.1229
```

Answer: The condition number is 48.1229. Since the condition number is greater than 30, we should be concerned about collinearity.

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?

```r
vif(model)
```

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

Answer: The VIF values are as follows: expend = 9.39, salary = 8.10, ratio = 2.29. The variable with the largest VIF is expend. Expend and Salary both have VIF values greater than 5 which suggests collinearity.

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?

```r
new_model <- lm(total ~ ratio + I(expend + salary), data=data)
vif(new_model)
```

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

Answer: After adding expend and salary together, both of the sum of these predictors and ratio have a VIF value of 1.005151. Since this value is less than 5 we can say that non 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.

```r
summary(new_model)
```

```
## 
## Call:
## lm(formula = total ~ ratio + I(expend + salary), data = data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -146.82  -43.88    5.57   39.93  126.55 
## 
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)    
## (Intercept)        1122.749     95.620  11.742  1.4e-15 ***
## ratio                 1.657      4.335   0.382  0.70399    
## I(expend + salary)   -4.536      1.372  -3.305  0.00182 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 68.59 on 47 degrees of freedom
## Multiple R-squared:  0.194,  Adjusted R-squared:  0.1597 
## F-statistic: 5.655 on 2 and 47 DF,  p-value: 0.006302
```

Answer: For the slope parameter corresponding to ratio, the test statistic is 0.382, the p-value is 0.704, and a statistical decision at the 5% significance level is to not reject the null hypothesis and say that there is no significant linear relationship between ratio and total with the other predictor present in the model. For the slope parameter corresponding to the sum of expend and salary, the test statistic is -3.305, the p-value is 0.00182, and a statistical decision at the 5% significance level is to reject the null hypothesis and say that there is a significant linear relationship between the sum of expend and salary and the response, total, with the other predictor present 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.

```r
anova(new_model, model)
```

```
## Analysis of Variance Table
##
## Model 1: total ~ ratio + I(expend + salary)
## Model 2: total ~ expend + salary + ratio
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1     47 221106
## 2     46 216812  1    4293.7 0.911 0.3448
```

Answer: The null hypothesis of this F-test is that there is no sufficient improvement in the linear model to predict "total" when the predictors expend and salary are combined into one predictor. The test statistic is 0.911 and the p-value is 0.3448. A statistical decision at $\alpha = 0.05$ is to reject the null hypothesis at a p-value of less than 0.05. Since our p-value is greater than 0.05, we do not reject the null and thus conclude that there is no sufficient improvement in the linear model to predict "total" when the predictors expend and salary are combined into one predictor. Therefore, we would prefer the model from part 2.