Name: Brad Hall
ISDS 7024
Homework 5

Analyze `gasoline.xlsx` and use the results to answer these questions.

1) Review the correlations among the independent variables (IVs). Choose the 3 IVs that are the least affected by collinearity.

X_1	X_7
X_2	X_8
X_3	X_9
X_4	X_{10}
X_5	X_{11}
X_6	

Solution: $\boldsymbol{X_4, X_6, X_{11}}$

2) Use all of the IVs to predict the dependent variable (DV). Select the 4 IVs with the largest VIFs.

Solution: $\boldsymbol{X_1, X_2, X_3, X_{10}}$

3) Conduct a Principal Component Analysis using all of the IVs.

(a) What are the 3 smallest eigenvalues? (Round to 3 decimal places.)

Solution: **0.003, 0.008, 0.033**

(b) What is the condition number?

Solution: $\boldsymbol{k = 46.93}$.

$$k = \sqrt{\frac{\lambda_1}{\lambda_{11}}} = \sqrt{\frac{7.702575}{0.003497}} = 46.93$$

(c) Is multicollinearity an issue? Why or why not?

Solution: **Yes, multicollinearity is an issue.** This is because the condition number is well above even the conservative threshold of 30.

(d) Examine the four eigenvectors corresponding to the four lowest eigenvalues and determine which 4 IVs are contributing most to the issue of multicollinearity.

Solution: $\boldsymbol{X_1, X_2, X_3, X_{10}}$

(e) How does this compare to your answer in **(2)**.

Solution: They are the same.

4) Create the principal components for all IVs. Correlate the components with Y. Choose the 3 components that have correlations with the largest absolute values.

PC_1	PC_7
PC_2	PC_8
PC_3	PC_9
PC_4	PC_{10}
PC_5	PC_{11}
PC_6	

Solution: $\mathbf{PC_1}$, $\mathbf{PC_9}$, $\mathbf{PC_7}$

5) Run the linear regression using all 11 PCs to predict Y.

(a) What is R-square? (Round to 4 decimal places.)

Solution: $\boldsymbol{R^2 = 0.8353}$

(b) Following the rule that (in general) non-significant predictors should not be retained in a regression model, choose all of the significant PCs.

Solution: $\mathbf{PC_1}$

(c) Is multicollineaity an issue? Why or why not?

Solution: **No. Multicollinearity is not an issue.** The condition number for the principal components is $k = 1$, below the described threshold.

$$k = \sqrt{\frac{\lambda_1}{\lambda_{11}}} = \sqrt{\frac{1}{1}} = 1$$

6) One of the ways to reduce collinearity is to combine variables that are highly correlated. Examine the signs of collinearity and choose the 4 variables you would combine to reduce the overall collinearity in the model.

Solution: $\boldsymbol{X_1, X_2, X_3, X_{10}}$

7) Another approach to eliminate multicollinearity is to delete the variables that contribute to the issue. Using your answer in **(3d)**, delete the 4 variables you determined are contributing most to multicollinearity, and run the linear regression model to predict Y using the remaining variables.

(a) Which variables are significant?

Solution: $\boldsymbol{X_6, X_9}$

Keeping only those variables that are significant, rerun the model.

- Is multicollinearity still a problem?

  Solution: **No. Multicollinearity is not a problem.**

- What is the value of R^2? (Round to 4 decimal places.)

  Solution: $\boldsymbol{R^2 = 0.6462}$

8) Based upon your answer in 6, would you use the PCs to predict Y or would you simply delete the variable contributing most to multicollinearity?

Solution: We would use the PCs to predict Y since we get a significantly larger R-square value.

9) Which of the following is true concerning the VIF?

(a) If you were to calculate VIFs for a set of Principal Components created from the same data set, they would all equal 0.

Solution: **False.** They would all equal 1.

(b) An average VIF of 20 would indicate that the squared error of the OLS estimators is 20 times larger than it would be if the predictors were orthogonal.

Solution: **True.**