Assignment 3 — Due 11:59pm CDT October 9, 2022

LATE submission will only be accepted under special circumstances. You must show your work and organize it to get full credit. Your submission should be neat, clearly legible (typed or written legibly), present the problems in the order assigned, and use complete sentences in proper English.

*Questions that require the use of a statistical package should be answered with **edited formatted output**, i.e., unedited computer output will not be considered as a valid answer. You should include only the numbers, tables, or plots needed to answer the questions. Your commands/program script without output MUST be included as an appendix at the end of your document. **Plots are not unedited computer output.** Make sure to insert the plots requested and those that justify your answers.*

"Bonus Points" problems are not required, if submitted they will be graded and could increase your score, but not lower it. "Practice Problems" are not required and should not be submitted. They provide an opportunity to practice your skills. Feel free to ask questions about them, however I will not grade them.

1. **National Track Women Records** (`T1-9.dat`). The national track records for women in 54 countries for 7 running events is presented in Table 1.9 page 44. Let X_1 = 100m (s); X_2 = 200m (s); X_3 = 400m (s); X_4 = 800m (min); X_5 = 1500m (min); X_6 = 3000m (min); and X_7 = Marathon (min).

   Define the following linear combinations

$$\begin{aligned} V_1 &= \frac{1}{3}X_1 + \frac{1}{6}X_2 + \frac{1}{12}X_3, \\ V_2 &= \frac{1}{2.4}X_4 + \frac{1}{4.5}X_5 + \frac{1}{9}X_6, \\ V_3 &= \frac{1}{60}X_7 \end{aligned}$$

   (a) Compute $\bar{\boldsymbol{x}}, \boldsymbol{S}$ and $\boldsymbol{R}$ for the original 7 variables, i.e. for $X_1, \ldots, X7$.

   (b) Calculate the observed values of V_1, V_2, and V_3 for every country. Report only the values of V_1, V_2, and V_3 for Dominican Republic, Ireland, Guatemala, and Denmark. This short list is only for this question, use all the countries to answer everything else.

   (c) Calculate $\bar{\boldsymbol{v}}$, and $\boldsymbol{S}_v$, the sample mean vector and the variance-covariance matrix of V_1, V_2, and V_3 using the values that you computed on 1b.

   (d) Now, recalculate $\bar{\boldsymbol{v}}$, and $\boldsymbol{S}_v$, using only the values of $\bar{\boldsymbol{x}}$, and $\boldsymbol{S}$ obtained in 1a. Formulas can be found in textbook's **Result 3.5**, page 141 or **Result 3.6**, page 144.[1]
   *Hint: Identify the values that form the vectors **b** and **c**, and the matrix **A**.*

   (e) Compare your results from parts 1c and 1d.

   [*Bonus Points*] What interpretation can you give to V_1, V_2, and V_3 in the context of the data?

2. **Seoul Bikes 1** (`SeoulBikes_F12022.csv`): For this problem, we will use the Seoul bikes data again. Perform a regression analysis using Y = RentedBikeCount as the response, and the other variables as predictors. For simplicity, we will treat Y = RentedBikeCount as a continuous variable and we will exclude X_1 = Day, X_2 = Month, and X_3 = Year from the analysis.

   Hint: You don't need to include all the interactions but we discovered an interaction in previous assignments that could be important in this model.

   (a) Find the most appropriate linear regression model. What is your estimated linear regression model? What are the error associated to your estimates?

   (b) Using the estimated residuals, check that your linear model satisfies the assumptions of a traditional linear regression model. Include the plots that you used and the relevant interpretations.

[1] Other countries editions of the textbook have them listed as **Result 2.5**, and **Result 2.6**.

(c) What is the 95% prediction interval for the rented bike count of the Evening of a Non Holiday day in Spring under functional hours, with Temperature = 12.1, Humidity = 29, WindSpeed = 2.3, Visibility = 1734, DewPointTemp = -5.4, SolarRadiation = 2.26, without rain or snow?

3. **Seoul Bikes 2** (`SeoulBikesVer2_Fl2022.csv`): For this problem, we will use a New version of the Seoul bikes data. `SeoulBikesVer2_Fl2022.csv` has only the morning rows, i.e., $X_{16} =$"Morning", and includes three variables associated with rented bikes: Y_1 = RentedBikeCount, the original variable we analyzed in the previous problem; Y_2 = BikeCountPlus2, and Y_3 = BikeCountEvening, the rented bike counts on one-hour periods that start two and ten hours after the original count, respectively. Y_3 are the same evening counts we analyzed as different observations. All three bike counts, Y_k's are potential response variables, not predictors.

   Perform a multivariate regression analysis with the responses Y_1 = RentedBikeCount, Y_2 = BikeCountPlus2, and Y_3 = BikeCountEvening, and the predictors: X_5 = Temperature, X_6 = Humidity, X_7 = Wind speed, X_8 = Visibility, X_9 = Solar radiation, and X_{13} = Seasons.

   (a) Find the most appropriate multivariate linear regression model by retaining only those predictor variables that are significant. What is your estimated linear regression model? What are the errors associated to your estimates? As before, testing all the interactions is not required, but one interaction might be important.

   (b) Using the estimated residuals, check that your linear model satisfies the assumptions of a traditional multivariate linear regression model. Include the plots that you used and the relevant interpretations.

   (c) What is the 95% prediction ellipse for all three hours of rented bike counts for a Non Holiday day in Spring with Temperature = 12.1, Humidity = 29, WindSpeed = 2.3, Visibility = 1734, DewPointTemp = -5.4, SolarRadiation = 2.26? Compare this ellipse with the prediction interval from the univariate linear model above. Comment.

Bonus Points 2: Comment on the decision of excluding Day and Month. What would be the implications of including them? For instance, should they be treated as categorical or continuous? How the interpretation changes if you choose categorical rather than continuous? Compare the consequences vs. the benefits of including them as categorical, as continuous, and of excluding them?

Readings: Textbook, Section 7.5 and 7.6.

Practice Problem 1: Bulls Using the data on the characteristics of bulls sold at auction in Table 1.10:

1. Perform a regression analysis using the response Y_1 = SalePr and predictor variables Breed, YrHgt, FtFrBody, PrctFFB, Frame, BkFat, SaleHt, and SaleWt.

   (a) Find the "best" linear regression equation by retaining only those predictor variables that are individually significant.

   (b) Using the best fitting model, construct a 95% prediction interval for the selling price for a bull with predictor variables (in the order listed above) 5, 48.7, 990, 74.0, 7, 0.18, 54.2 and 1450.

   (c) Examine the residuals from the best fitting model.

2. Repeat the previous analysis, using the natural logarithm of the sales price as the response variable. That is, set $Y_1 = \ln(\text{SalePr})$. Which analysis do you prefer? Why?

Practice Problem 2: Let $\boldsymbol{A} = \begin{bmatrix} 13 & -4 & 2 \\ -4 & 13 & -2 \\ 2 & -2 & 10 \end{bmatrix}$.

(a) Is $\boldsymbol{A}$ symmetric?

(b) Show that $\boldsymbol{A}$ is positive definite.

(c) Determine the eigenvalues and eigenvectors of $\boldsymbol{A}$.

(d) Write the spectral decomposition of $\boldsymbol{A}$.

(e) Find $\boldsymbol{A}^{-1}$.

(f) Find the eigenvalues and eigenvectors of $\boldsymbol{A}^{-1}$.