DASC 5420: Exercise 3–Regression

Due Date: March 10, 2024

Make an **R Markdown** file and produce a **pdf** file by solving the following problems. In the title of your file make sure you include your name and student ID number.

1. Generate two vectors, x_1 and x_2, containing 100 observations drawn from $N(1.1, 3.5)$ and $N(5, 10)$, respectively. Plot the data in side by side histogram with different colors and check normality. Make sure to use set.seed(5420) prior to starting to ensure consistent results. [2.5]

2. Generate a line plot that shows the values of three functions for the inputs 0 to 100. Use different colors for three functions and label the three functions with proper legend. [2.5]

   Hints: example functions include $x, 1/x, x^2, x\log x, 3x^3, 20x$ etc.

3. (a) Make a set of data set with 4 variables. For 2 of the variables, the data should be partially correlated and 1 variable should not be correlated and 1 variable will be a categorical variable with 3 groups (say, 1,2 and 3).

   Generate a pairwise scatter plot of this data by categorical variable. Use different colors and points/shape for each groups of categorical variable. Give a proper title of the plot. Make sure to use set.seed(5420) prior to starting to ensure consistent results. [2.5]

   (b) Standardize (or normalize) the data from part 3(a) and draw another plot with the scaled data. Remember we only standardize the quantitative variables. [2.5]

4. In this exercise you will create some simulated data and will fit simple linear regression models to it. Make sure to use set.seed(542024) prior to starting part (a) to ensure consistent results. [10]

   (a) Using the 'rnorm()' function, create a vector, x, containing 100 observations drawn from a $N(0, 1.5)$ distribution. This represents a feature, X.

   (b) Using the 'rnorm()' function, create a vector, e, containing 100 observations drawn from a normal distribution with mean zero and variance 0.20.

   (c) Using x and e, generate a vector y according to the model

   $$Y = -2.2 + 0.8X + e \quad (1)$$

   (d) What is the length of the vector y? What are the values of β_0 and β_1 in this linear model? Comment on the correlation between y and x based on this model.

   (e) Create a scatter plot displaying the relationship between x and y. Comment on what you observe.

   (f) Fit a least squares linear model to predict y using x. Comment on the model obtained. How do $\hat{\beta}_0$ and $\hat{\beta}_1$ compare to β_0 and β_1?

   (g) Display the least squares line on the scatter plot obtained in (d). Draw the population regression line on the plot, in a different color. Use the 'legend()' command to create an appropriate legend.

   (h) Is there evidence of non-linear association between any of the x and y? To answer this question, fit a model that predicts y using x and x^2. That is, fit a model of the form

   $$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + e.$$

   Provide interpretation of your results.

(i) Repeat (a)–(f) after modifying the data generation process in such a way that there is more noise in the data. The model (1) should remain the same. You can do this by increasing the variance of the normal distribution used to generate the error term e in (b). Interpret your results.

(j) What are the confidence intervals for β_0 and β_1 based on the original data set and the noisier data set? Compare the results.

(k) Prediction: what is the predicted Y associated with a X of 10? What are the associated 95% confidence and prediction intervals? Comment on the results. Use the original model (1) to do this.

Upload the final **pdf** file in Exercise 3: Regression in **Moodle**. Make sure you have included all the **R** codes in the output file.

Good Luck!