MATH 8050 HW2

Due Monday, September 12

Show all work. You may prepare either hand-written or typed solutions, but please make sure that they are complete and legible. Incomplete solutions or answers that cannot be read will be given no credit. Please type up the R results, including copying the figures into a document. Only the relevant R output needs to be shown in the body of your homework paper. However, be sure to include all of your R code as an appendix.

1. On Blackboard you will find a dataset named `SaltConc.xlsx`. This spreadsheet contains data that were recorded on the salt concentration (in mg/L) found in surface streams in a particular watershed and the corresponding percentage of watershed area consisting of paved roads. Read the dataset into `R` and do the following:

   (a) Find the mean, median, standard deviation, maximum, and minimum values for both the variables in the dataset.

   (b) Create side-by-side boxplots for the two variables in the dataset.

   (c) Create a histogram of the observed Areas. Label the x-axis as "Area" and title it "Area Distribution".

   (d) Create a scatter plot of Salt (y) against Area (x). Label the axes as "Salt Concentration" and "Roadway Area".

   (e) Write a simple function that takes a dataframe as its argument and returns the sum of all values in the second observation of that dataframe. Then pass the salt concentration dataset into that function to make sure that it is performing as desired.

   (f) The `R` function `t.test` can be used to carry out t-tests (as you might guess from its name!). The syntax is

   ```
   t.test( x = <data vector 1>, alternative = <alternatve>,
           mu = <H0 value>, conf.level = <confidence level>)
   ```

   where `alternative` indicates whether it is a right-tailed (`"greater"`), left-tailed (`"less"`), or two-tailed (`"two.sided"`) test. (There are other options in the function that we won't worry about here - type `?t.test` from the command line to see further documentation.) Use this function to test $H_0: \mu_{\text{Salt}} = 20$ versus $H_1: \mu_{\text{Salt}} \neq 20$ at the $\alpha = 0.10$ significance level, where μ_{Salt} is the population mean of the salt concentrations.

2. Derive the least-squares estimates of the unknown parameters in the usual linear regression model, $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.

3. When asked to state the simple linear regression model, a student wrote it as $E(Y_i) = \beta_0 + \beta_1 x_i + \epsilon_i$. Do you agree? Explain.

4. A student working on a summer internship in the economic research department of a large corporation studied the relation between sales of a product (Y, in millions of dollars) and a population (x, in millions of persons) in the firm's 50 marketing districts. The normal error regression model was employed. The student first wished to test whether or not a linear association between Y and x existed. The student accessed a simple linear regression program and obtained the following information on the regression coefficients:

		95 Percent	
Parameter	Estimated Value	Confidence Limits	
Intercept	7.43119	-1.18518	16.0476
Slope	0.755048	0.452886	1.05721

   (a) The student concluded from these results that there is a linear association between Y and x. Is the conclusion warranted? What is the implied level of significance?

   (b) Someone questioned the negative lower confidence limit for the intercept, pointing out that dollar sales cannot be negative even if the population in a district is zero. Discuss.

5. A member of a student team playing an interactive marketing game received the following computer output when studying the relation between advertising expenditures (x) and sales (Y) for one of the team's products:

   Estimated regression equation: $\hat{Y} = 350.7 - 0.18x$
   Two-sided p-value for estimated slope: 0.91

   The student stated, "The message I get here is that the more we spend on advertising this product, the fewer units we sell!" Comment.