Clarkson University
Department of Mathematics
Due Monday April 20, 2020

STAT 383
Problem Set 7

Name: _____

Instructions: Parts (a)-(d) should be answered for Questions 1-7. Only Questions 2 and 7 will be graded, for a total of 40 points. I would highly recommend that you use some form of software to do all of the analysis for you (fit regression, create ANOVA table, plot residuals, etc.). This assignment is meant to assess your ability to INTERPRET the results. You should turn in both your code AND output - entered nicely into Word or LaTeX and saved/compiled as a PDF.

1. Estimating the costs of drilling oil wells is an important consideration for the oil industry. The Drill1 data set contains the total costs and depths of 16 offshore oil wells located in the Philippines. It is expected that cost is a linear function of the depth.
   (a) Write out the equation of the regression line. Interpret the slope and intercept in the context of this problem. Do they make sense? Include a scatter plot of the data with the correct regression line added (confidence bands would be nice too, but are not required).
   (b) Test the hypothesis that the a linear relationships exists between the predictor and response variable (ANOVA, t-test for β_1, t-test for ρ, or a confidence interval for β_1).
   (c) What is the R^2 for the SLR you have obtained? What does the value mean? Use it to evaluate the linear model.
   (d) Plot the standardized residuals against the independent variable. What can you say about the regression using this graph? (HINT: Are there outliers? Does it seem reasonable to claim the data has a linear fit?)
2. (20 pts) A lab has 12 pairs of power boilers that each have a pre-determined burner area liberation rate (in MBtu/ft^2 per hour). The nitrous oxide (NO_x) emission rates (in ppm) are measured for each. The pollution/emission rate is expected to be a linear function of the burner area liberation rate. Answer all the questions from #1 as they would pertain to the Burner Rates data set.
3. A warehouse manager is interested in the possible improvements to labor efficiency if air conditioning is installed in the warehouse. The time it takes to unload a fully laden truck is expected to be a linear function of the ambient temperature (see Truck Times data set). Answer all the questions from #1 as they would pertain to this data set.
4. A realtor collects data concerning the size of a random sample of newly constructed houses in a certain area together with their tax appraised values (see Tax Appraisal data). It is expected that the appraisal value is a linear function of the home size. Answer all the questions from #1 as they would pertain to this data set.
5. An exercising individual breathes through an apparatus that measures the amount of oxygen in the inhaled air that is used by the individual. The maximum value per unit time of the utilized oxygen is then scaled by the person's body weight to come up with a variable called VO2-max, which is a general indication of the aerobic fitness of the individual. Data is collected for a random sample of 20 male subjects of various ages (see Aerobic1 data set). It is expected that V02-max is a linear function of age. Answer all the questions from #1 as they would pertain to this data set.
6. During the installation of a large computer system, it is useful to know how long specific tasks will take, particularly programming changes. A great deal of effort is spent estimating the amount of time such tasks will take and learning how to effectively use such estimations. Having an accurate idea of the time required for these tasks is crucial for the effective planning and timely completion of the installation. The Install data set relates one expert's time estimates for programming changes to the actual times the tasks took. It is expected that the actual time can be expressed as a linear function of the estimated time. Answer all the questions from #1 as they would pertain to this data set.
7. (20 pts) An engineer is examining the behavior of electrical resistivity (in nΩ·m) of a particular metal as it could be predicted by the temperature (in K). Use the Copper Resistivity data to answer all questions from #1 as they would pertain to this data set.

8. Suppose that a multiple linear regression is done to predict the healthiness rating of a cereal based on its fat, fiber, and sugar content per serving given a data set of 77 cereals. It is calculated that the sum of the squared residuals is 2000.8735 and the total sum of squares is 4932.3139.
   (a) Fill out a complete ANOVA table using the given information (including the appropriate associated hypotheses). What can you conclude?
   (b) Use the ANOVA table you created to determine R^2. Interpret the obtained value.
   (c) Use the table below to determine which coefficients are statistically significant using $\alpha = 0.05$. Interpret the significant coefficients.

Health Rating	Coef.	Std. Err.
Fat	-1.79	0.62
Fiber	-0.49	0.25
Sugar	-1.24	0.14
Constant	53.44	1.34

9. The Drill2 data set includes additional variables: geology, downtime, and rig-index. The geology variable is a score that measures the geological properties of the materials that have to be drilled through (higher scores indicate harder materials) in order to complete the oil well. The downtime variable measures the number of hours that the drilling rig is idle due to factors such as inclement weather or necessary testing. The rig-index variable compares the daily rental costs of the drilling rig to the cost 5 years previously (an index of 2 indicated that the drilling rig is twice as expensive as it had been).
   (a) Fit a multiple linear regression model that tries to predict cost as a linear function of the four other variables.
   (b) Should any variables be omitted from the model? Why or why not?
   (c) What model (give an equation) for cost would you ultimately recommend?
   (d) Plot the standardized residuals against the fitted values. Also plot the residuals against each of the input variables you considered. What can you say about your model from these plots? Do any plots make it clear that you should be excluding certain variables from your model?

10. The Aerobic2 data set includes additional predictor variables heart rate at rest, percentage body fat, and weight.
    (a) Fit a multiple linear regression model that tries to predict V02-max as a linear function of the four other variables.
    (b) Should any variables be omitted from the model? Why or why not?
    (c) What model (give an equation) for V02-max would you ultimately recommend?
    (d) Plot the standardized residuals against the fitted values. Also plot the residuals against each of the input variables you considered. What can you say about your model from these plots? Do any plots make it clear that you should be excluding certain variables from your model?