**Homework 2 (50 pts)**

**Due September 14, 2018**

*A reminder – Questions must be answered in order (including graphs), and all graphs must be fully labeled (main title should include the question number, and all axes should be labeled). Attached the (well labeled) RMD file used for all questions at the end of the homework.*

1. (2pts) 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: Ŷ = 350.7 - .18X

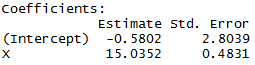
Two-sided P-value for estimated slope: .91

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

2. (4pts) Refer to the problem 5 in the homework1, use R to generate confidence band for

a) mean value prediction, b) single value prediction, c) mean of 3 new values prediction , and d) working-hoteling confidence band. Comment on their difference.

3. (19pts) Tri-City Office Equipment Corporation sells an imported copier on franchise basis and performs preventive maintenance and repair service on this copier. Data have been collected from 45 recent calls on users to perform routine preventive maintenance service; for each call, X is the number of copiers serviced and Y is the total number of minutes spent by the service person. Assume that simple linear regression model is appropriate. The following shows partial result.





a) (3pts) Estimate the change in the mean service time when the number of copiers serviced increases by one. Use a 90 percent confidence interval. Interpret your confidence interval.

b) (4pts) Conduct a t test to determine whether or not there is a linear association between X and Y here (i.e., ; At a significant level of 0.05, state the hypothesis, reject region, estimate the p value, and state the conclusion of your test.

c) (4pts) The manufacturer has suggested that the mean required time should not increase by more 14 minutes for each additional copier that is serviced on a service call. Conduct a test to test whether this standard is being satisfied by Tri-City. At a significant level of 05. State the hypothesis, reject region, estimate the p value, and state the conclusion of your test.

d) (2pts) Does give any relevant information here about the “start-up” time on calls—i.e. about time required before service work is begun on the copiers at a customer location?

e) (4pts) In order to perform the following hypothesis test (,

complete the following ANOVA table for the data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Source** | **degrees of freedom** | **Sum of Squares** | **Mean Square** | **F-value** |
| **Model** | **\_\_\_\_\_** | **76960** | **\_\_\_\_\_** | **\_\_\_\_\_\_** |
| **Error** | **\_\_\_\_\_** | **\_\_\_\_\_** | **\_\_\_\_\_** |  |
| **Corrected Total** | **\_\_\_\_\_** |  |  |  |

According to the F value and degree of freedom, use F tale to estimate the P-value of the test.

f) (2pts) Compare the F test statistic obtained here and demonstrate numerically its equivalence to the T test statistic in b).

4. (17pts) Refer to the problem 2 in homework 1.

a) (6pts) Complete the ANVOA table for the hypothesis test.

ACT and GPA score are associated ACT and GPA score are not associated

Or equivalently,

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Source** | **degrees of freedom** | **Sum of Squares** | **Mean Square** | **F-value** |
| **Model** | **\_\_\_\_\_** | **\_\_\_\_\_** | **\_\_\_\_\_** | **\_\_\_\_\_\_** |
| **Error** | **\_\_\_\_\_** | **\_\_\_\_\_** | **\_\_\_\_\_** |  |
| **Corrected Total** | **\_\_\_\_\_** |  |  |  |

Estimate the p value and state the conclusion

b) (4pts) what is ? Perform a hypothesis test on the correlation, compute the test statistic and estimate p value, then state the conclusion. Use significant level of 0.05.

c) (2pts) Compare 4a) to 3e), which model seems to be a better fit? Discuss the models based on MSE and

d) (3pts) The GPA data used in this problem is actually the first 10 cases of a larger data set, and has a very small , is it possible that for the complete set n>10, will not be zero? Could not be zero for the first 10 cases, yet equal to zero for all 30 cases? If applicable, sketch two scatter plots to demonstrate the two situations.

e) (2pts) Use R to compute a 95% CI for the population coefficient.

5. (Use R to complete this problem)(8pts) Experience with a certain type of plastic indicates that a relation exists between the hardness (measured in Brinell units) of items molded from the plastic, and the elapsed time since termination of the molding process.

Sixteen batches of the plastic were made, and one test item was molded from each batch. Each test item was randomly assigned to one of four predetermined time levels (X=16, 24, 32, or 40 hours), and the hardness (Y) was measured after the assigned elapsed. Data is in plastic.csv

a) (2pts) Obtain the estimated simple linear regression model (SLR or SLM). Plot the estimated regression function and the data. Does a linear regression function appear to be a good fit?

b)(2pts) Plot the residuals against the fitted values to ascertain whether any departures from regression model are evident. State your findings

c)(2pts) Plot a normal probability plot of the residuals. Perform a Shapiro-wilk normality test on the residuals. Does the normality assumption appear to be reasonable here?

d) (2pts) Use the Brown-Forsythe test to determine whether or not the error variance varies with the level of X. Divide the data into two groups, , use . Does your conclusion support your preliminary findings in part b)?