**Midterm Exam**

**Q1)**

1. We are interested in predicting the rating based on sodium content. Construct the appropriate scatter plot.

* Check output of code for answer

1. Based on the scatter plot, is there strong evidence of a linear relationship between the variables.

* **No** strong linear relation found

1. Perform the appropriate regression.

* Check output of code for answer

1. Which cereal is an outlier?

* DFFITS value for **observation 4 is 0.05**. Then corresponding cereal are outlier

1. What is the typical error in predicting rating based on sodium content?

* Typical error in preceding rating may be off by 13 units since **RMSE = 12.95**

1. Interpret the y- intercept.

* Due to the **Y intercept of 53**, if there is no sodium content then rating of cereal is 53.

1. Is there a significant relationship between the two variables? Why?

* Yes, because it has 95% confidence level.

1. What is the meaning of the slope?

* Slope is -0.067 which means that every unit change in sodium content, then rating will decrease by 0.067

1. Construct a 95% confidence interval for the average nutrition rating of all cereals with sodium content of 100.
2. Construct a 95% confidence interval for the nutrition rating for a randomly chosen cereal with sodium content of 100.

**Q2)**

Use the following ANOVA table of a regression model, produced by a software package, to answer the questions below.

|  |  |  |
| --- | --- | --- |
| **Source** | **DF** | **Sum of** |
| **Squares** |
| **Model** | 2 | 253.274 |
| **Error** | 147 | 6897.42 |

1. How many observations are used for this model?

* 147+2= 149
* n = 149+1 = **150**

1. How many parameters are used in this model?

* **2** parameters

1. What is the R-square for this model?

* R-square = 1 – (253.274/7150.694)
* = **0.9645**

1. What is the F value of the model?

* **2.696**

1. Is the F value of the model significant at 5% level of significance? Why?

* **No** because P value is **0.07**

**Q3)**

The forced expiratory volume in one second (FEV1) measurement shows the amount of air a person can forcefully exhale in one second. Using the Lung dataset in CANVAS, develop a regression model for predicting FEV1 of the father using weight and height of the father as predictors.

1. Is this a good model? Why?

* NO because, R square value is 27%.

1. Are the residuals normally distributed? Why?

* Yes it’s normally distributed. we can get it from graph.

1. What are the influential observations? Why?

* 18,31,50,80,84 and 87,106,110 and 130 are blue that all are influential. Because there are dffits and cooked values. These points will change the slope dramatically.

1. What are the high leverage observations? Why?

* Observation 110 is a high leverage observation as the DFFITS value is 0.5933.

1. For predicting FEV1 of the father, is height more important than weight? Why?

* For predicting the FEV1 of the father\_Height is more important,when I ran a regression with just height, the R square was 25% when I ran it with just weight, the R square is les then 1%.

**Q4)**

The Depression data set in CANVAS shows the result of a study on depression. Factors for predicting depression have been grouped into 20 categories (cat\_01 to cat\_20).

1. Normalize the “cat\_01” to “cat\_20” columns using the z transformation with mean=0 and mean=1.

* Check code

1. Perform PCA analysis on the normalized cat\_01 to cat\_20;

* Check code

1. Find the best variables among the top 20 principal components that can predict depression.

* Check code

Datasets: Cereals, Lung, Depression