Assignment - 1

BSAN 450 (Spring 2023)

**This assignment is due on January 26, 2023 at 9:00 PM Central. The total points possible are 100 and there are three (3) problems, each carrying equal points. You can form groups of 3 or less to attempt these problems. Each group will submit one copy of the assignment on Canvas, either in word or pdf, and the assignment should clearly include the names of the group members**.

1. The iron content of crushed blast furnace slag needs to be determined. Two methods are available. One involves a chemical analysis in the laboratory, which is time consuming and expensive. The other is a much cheaper and quicker magnetic test that can be carried out on site. Both measurements on 53 slags are available, the data are in the file ironcontent.csv. Investigate the extent to which the results of the chemical tests of iron content can be predicted from the magnetic test of iron content.

1. Chart, scatter chart

   Description automatically generatedPlot the chemical results versus the magnetic results.
2. Estimate a simple linear regression model with the chemical results as the response (*Y*) variable and the iron results as the input (*X*) variable.

Chem = 8.9565 + .58664(Magn)

1. Chart, histogram

   Description automatically generatedChart, scatter chart

   Description automatically generatedPerform the diagnostic checks for the regression model. Are there any problems with the model you estimated?

The residuals vs fitted plot looks random, the qq-plot looks to follow the 45-degree line, and there does not appear to be any significant outliers in the residuals vs leverage plot. Additionally, the histogram of the residuals appears to be roughly ~ normal. The Shapiro-Wilks test backs this up (p-value: 0.3516).

1. Test the hypothesis that the slope of the regression line equal 0 versus the alternative hypothesis that the slope of the regression line is not equal to 0. What can you conclude from this test?

From the summary output, we can see that the p-value for Magn is 4.38e-10 (t-value derived from estimate/tstd. error), so we reject the null that the Magn = 0. Therefore, Magn is a good predictor of Chem. Similarly, the F-stat returns a p-value of 4.375e-10, which suggests that the model should be more complex than just an intercept-only model.

1. What is the value of R-squared for this example? What does this value tell you about this model?

The R-squared value, 0.5372, suggests that 54% of the variation in Chem can be explained by Magn.

1. Find a 95% confidence interval for the parameters *β*0 and *β*1.

|  |  |  |
| --- | --- | --- |
|  | 2.5% | 97.5% |
| Intercept | 5.6392798 | 12.2737223 |
| Magn | 0.4335801 | 0.7397025 |

1. Use your model to predict the chemical result for a magnetic result of 20. Include 95% prediction limits.

|  |  |  |  |
| --- | --- | --- | --- |
|  | fit | lwr | upr |
| 1 | 20.68933 | 13.66964 | 27.70901 |

2. In this example there are 2 variables: Salary = the salary of the managers in a company and Experience = the years of experience for the manager. The objective is to fit a model for the Salary in terms of the Experience. The data is in a file named ManSalary.csv.

After you set the working directory in which this file is stored on your computer, read the data into R Studio.

1. Plot a scatter plot of the Salary on the vertical axis and Experience on the horizontal axis. Comment on this plot. Is there a relationship between these two variables? How would you describe this relationship?

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

There looks to be a moderately-strong, positive, somewhat linear relationship between experience and salary. However, the second plot shows that a linear relationship may not be the best fit.

1. Estimate a simple linear model with the *Y* variable equal to Salary and the *X* variable equal to Experience. Print out a summary of this model.

Salary = 11369.2 + 2141.3(Experience)

1. Chart, histogram

   Description automatically generatedDiagram

   Description automatically generatedPerform the diagnostic checks of the residuals for the model you estimated in part (b). For each plot is there any indication of a problem with the model?

The residuals vs fitted plot looks random, the qq-plot looks to follow the 45-degree line, and there does not appear to be any significant outliers in the residuals vs leverage plot. Additionally, the histogram of the residuals appears to be roughly ~ normal. The Shapiro-Wilks test backs this up (p-value: 0.3219).

1. If you find problem(s) with the diagnostic checks then suggest an appropriate transformation for Salary and re-analyze the data after you make the appropriate transformation.

Chart

Description automatically generated

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedLambda = 0.06060606

1. Why is the model that you estimated in part (d) above better than the one you estimated in part (b)?

Chart, diagram

Description automatically generatedChart, histogram

Description automatically generated

The diagnostic plots look slightly better all around. Most notably, the residuals vs fitted follows mean 0 better. Additionally, the p-value from the Shapiro-Wilk test was higher (0.4381).

3. In an industrial laboratory, under uniform conditions, batches of electrical insulating fluid were subjected to constant voltages until the insulating property of the fluids broke down. Seven different voltage levels were studied. The measured responses were the times, in minutes, until breakdown. The data are in the file named breakdown.csv and the variables in the file are Time and Voltage. Analyze this data to find a model to predict Time based on Voltage.

Chart, scatter chart

Description automatically generated 1) Plot data

From this data, it is obvious that a linear model will not fit the data well.

Additionally, the Shapiro-Wilks test returns a p-value < .05, therefore the residuals are not normally distributed. 2) From