Elijah McLaughlin

October 6, 2023

STA 3064

Regression Case Study

**1. Data Exploration:**

* 1. Outside of the further data step, no further data manipulation required.
  2. Scatterplot and correlation plot:  
      A chart with blue dots

     Description automatically generated  
     A screenshot of a data sheet

     Description automatically generated

1. **Simple Regression Model Fitting and Analysis:**
   1. A graph of a line graph

      Description automatically generated with medium confidence   
      Linear equation: Phenol = 44.85626 + 202.74772\*DScavAct
   2. R2=0.6920, explains 69.2% of variability and suggests good, positive correlation between Phenol and DScavAct.
   3. A chart of data analysis

      Description automatically generated with medium confidence  
      Data appears skewed near 0 on residual plot. This suggests at least one outlier, which follows with the linear regression graph in (a).
   4. After exploring data, I removed Beer 34 as it is an outlier. Performing a simple regression analysis, R2=0.7913. This is the better model for further analysis.A graph of a line graph

      Description automatically generated with medium confidence
   5. The 95% confidence interval is [190.53147, 269.18252]. 95% of predicted data for DScavAct will be between roughly 190 to 269.
   6. Normal theory CI: [190.53147,269.18252] and BootCI: [200.188,260.178]. The BootCI has a higher lower bound and lower higher bound, narrowing the 95% confidence interval from (e). Since the sample size is relatively small, the normal theory confidence interval may not be accurate. Performing a bootstrap confidence interval would be more accurate since it performs random sampling with replacement helps with accuracy for smaller datasets such as this one. Using the bootstrap confidence interval will be more effective for this analysis.
   7. F Value = 9.67, and Pr > F = 0.0001, which is strong evidence to reject the null hypothesis and the model is statistically significant.
   8. The prediction interval is [190.53147, 269.18252], and the predicted x\* value is 237.4375. It seems that the new data did not affect the 95% confidence interval. In fact, the predicted value for x\* is between the 95% confidence interval, which logically follows the confidence intervals.
2. **Multiple Regression Model Fitting and Analysis:**
   1. The model may be more accurate since it will have more data, however it could lead to too much data and have the model become inaccurate. Finding variables with the statistically significant correlation to Phenol will be ideal.
   2. I used the same step from (1. b). DScavAct, AScavAct, ReducePower, MetalChelate all have correlation >0.5 and appear statistically significant.
   3. R2=0.9071, indicating the chosen variables explain 90.71% of variance in Phenol, and are statistically significant.
   4. A diagram of a function

      Description automatically generated with medium confidence  
      Data is skewed right in the residual plot. Since the dataset is small, producing a bootstrap distribution may be better for analysis. F Value = 83.00 and Pr > <0.0001, suggesting the data is statistically significant and therefore fits. In the parameter estimates, AScavAct appears to not be statistically significant, and a reduced model may be better.
   5. A diagram of a function

      Description automatically generated with medium confidence  
      The only difference between the two models is the F Value, which is 113.86 in the reduced model of interest. Because of Pr > <0.0001, both models are statistically significant and suggests the rejection of the null hypothesis in both full model and reduced model of interest.
   6. A graph with blue dots and red circles

      Description automatically generated  
      A graph with red dots and blue dots

      Description automatically generated  
      A graph with red dots and blue dots

      Description automatically generated  
      The areas circled in red appear to be outliers and may be thrown out. The outliers could be skewing the data as well as distorting the slope and intercept of regression analysis. The outliers could also lead to statistical significance when there shouldn’t be. However, with a small dataset, it may impact the regression analysis more by removing the outliers entirely. Since it may be more harmful to remove the outliers, I will not remove any of the data points.
   7. From the above analysis, I believe the model from (d) is the best. The F Value for the full model is lower than the reduced model, while remaining statistically significant.