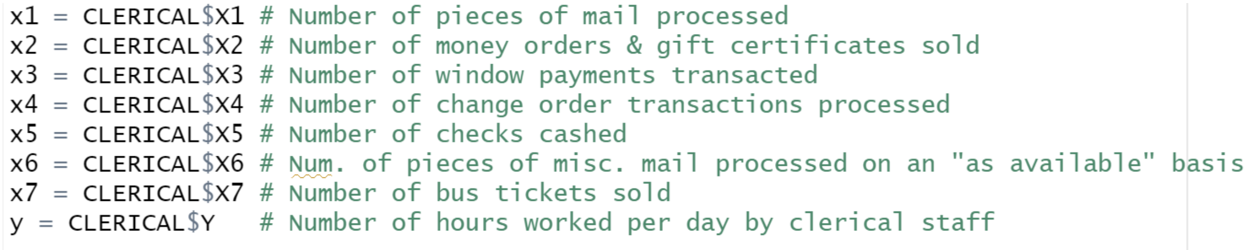
Jathin Banala
801316754

Statistical Analysis of Clerical Work Hours

Introduction

For this project, I have chosen to analyze the total number of hours worked per day by clerical staff members at a department store as a function of several different variables designated x1 through x7, and I have attached an explanation of each variable below.



```
x1 = CLERICAL$X1 # Number of pieces of mail processed
x2 = CLERICAL$X2 # Number of money orders & gift certificates sold
x3 = CLERICAL$X3 # Number of window payments transacted
x4 = CLERICAL$X4 # Number of change order transactions processed
x5 = CLERICAL$X5 # Number of checks cashed
x6 = CLERICAL$X6 # Num. of pieces of misc. mail processed on an "as available" basis
x7 = CLERICAL$X7 # Number of bus tickets sold
y = CLERICAL$Y   # Number of hours worked per day by clerical staff
```

I planned on creating a model utilizing multiple linear regression with 7 independent variables and 1 dependent variable and looking at the residuals to determine whether or not the model is a good fit. In order to carry out my project, I will use the R Studio software and the R programming language.

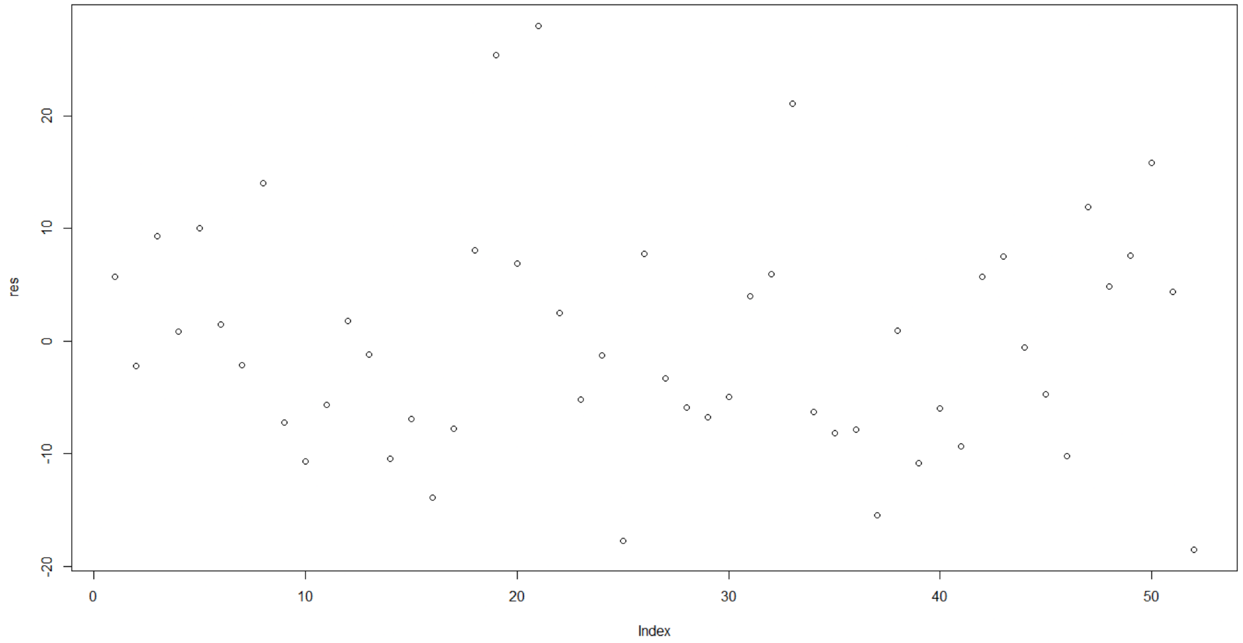
Methodology

In order to create my model, I first had to load the CLERICAL dataset and declare my variables. Then, I elected to create my model, summarize various aspects of the model (coefficients of independent variables, standard error, t values, p values, etc.), and generate an ANOVA (Analysis of Variances) table to illustrate the various components involved in variation and show whether they are a result of error or residual variation. Then, I plotted the residuals, standardized and studentized the residuals, and generated an additional 2 graphs based on the residuals after they have been standardized and after they have been studentized. Finally, I used the qqnorm and qqline functions to produce a normal probability plot of standardized residuals.

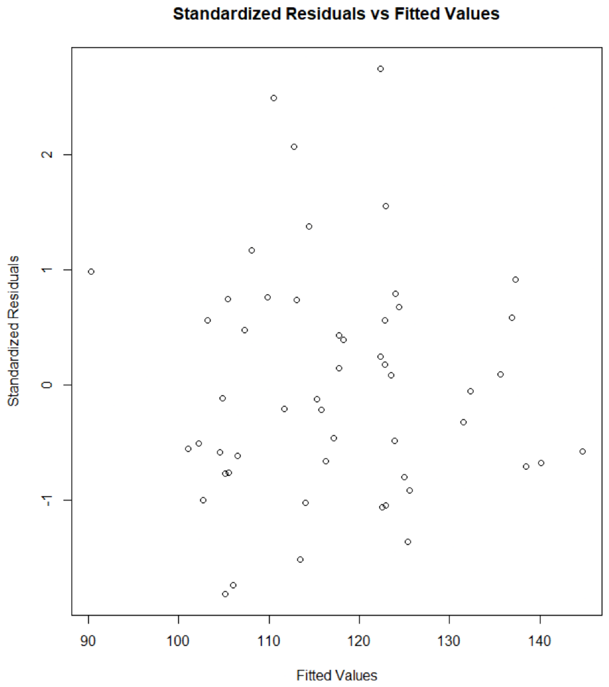
Data Analysis

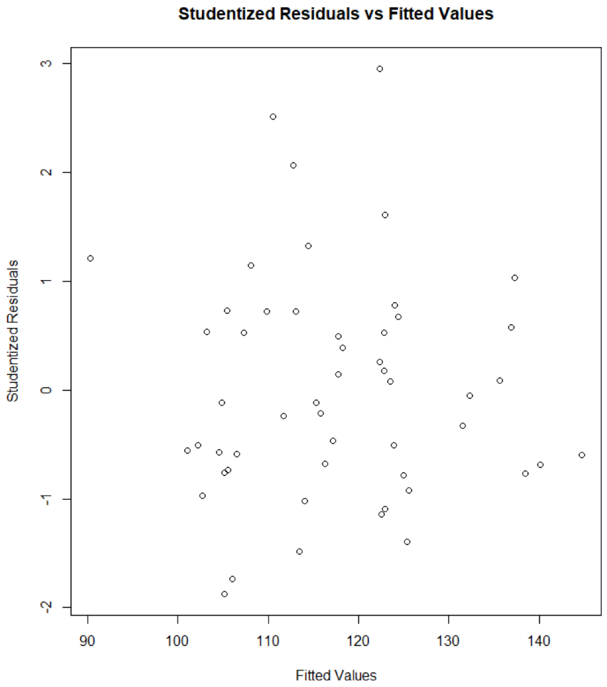
After generating a summary of the model, I was able to come up with the fitted equation E(y) = 60.5537920 + 0.0013496 * x1 + 0.0872715 * x2 + 0.0086879 * x3 - 0.0427781 * x4 + 0.0467902 * x5 + 0.2092130 * x6 + 0.0048192 * x7. The coefficient of determination (R^2) shows that 56.84% of the variance in the dependent variable is explained by the independent variables, and, when adjusted for the number of independent variables in the model, 49.97% of the variance in the dependent variable is explained by the independent variables. The standard error (σ^) = 10.99 with 44 degrees of freedom.

After plotting the residuals, I noticed that there was no discernible pattern and the sum of the residual was equal to 0, indicating that a linear model is appropriate.

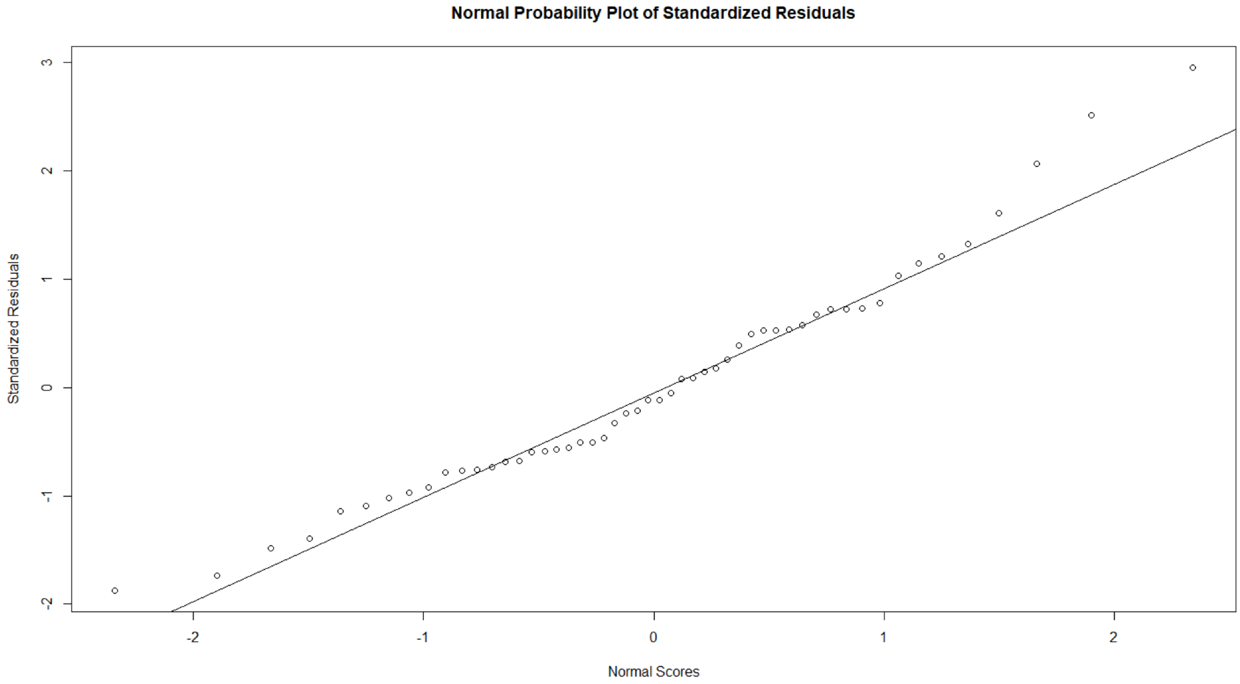


I once again used the plot function in R Studio to generate 2 graphs: the first compared the standardized residuals with the fitted values while the second compared the studentized residuals with the fitted values. In both graphs, there was no discernible pattern to the residuals, once again showing that there is likely a linear relationship between the independent and dependent variables.





Finally, I created a quantile-quantile plot with a line through the first and third quartiles. In this plot, there were no obvious outliers that noticeably skewed the graph, meaning that no data points needed to be omitted from the model or from the data set.



Conclusion

After analyzing the multiple linear regression model based on the prediction equation and the residuals, we can conclude that the model is relatively good and adequate at explaining the variance in the dependent variable. Additionally, the standard error of the model of 10.99 is rather low given that our data primarily consists of numbers in the triple digits, which is another indication that our model is quite accurate.

Citations

Mendenhall, W., & Sincich, T. (2014). *A second course in statistics: Regression Analysis* (7th ed.). Pearson Education.