1. For this question, use the Franchises data file. The file has data on several variables explained below. We would like to understand what predicts financial growth in 2011. Run the required analysis and articulate the variables that are important. Run the required analysis and articulate the findings.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **offid\_n** | **oyrsact** | **grth2011** | **grth2010** | **distancetoHQ** | **officesincounty** | **Ownmgrexp** |
| **Explanation** | Franchise ID number | How many years has the franchise been a part of the network | Financial growth in revenue from 2010 to 2011 | Financial growth in revenue from 2009 to 2010 | The distance from the franchise location to the headquarters | The number of other offices located in the same county (competition) | The total experience of the franchise’s leadership |

**Model: MODEL1**

**Dependent Variable: grth2011 grth2011**

|  |  |
| --- | --- |
| **Number of Observations Read** | 764 |
| **Number of Observations Used** | 755 |
| **Number of Observations with Missing Values** | 9 |

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 4 | 6.77816 | 1.69454 | 21.83 | <.0001 |
| **Error** | 750 | 58.21569 | 0.07762 |  |  |
| **Corrected Total** | 754 | 64.99385 |  |  |  |

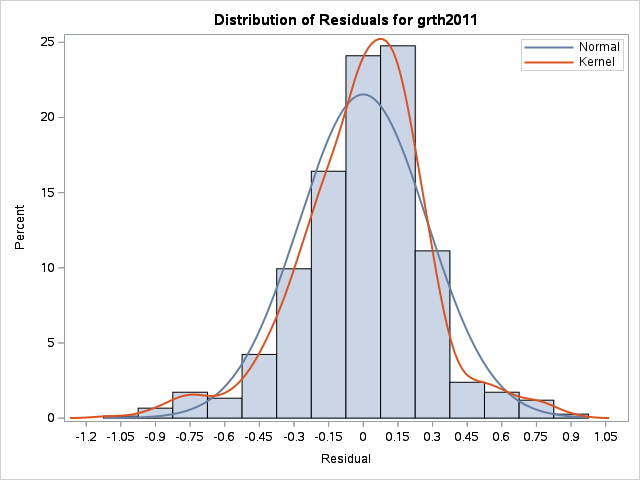
|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 0.27861 | **R-Square** | 0.1043 |
| **Dependent Mean** | 0.04075 | **Adj R-Sq** | 0.0995 |
| **Coeff Var** | 683.73367 |  |  |

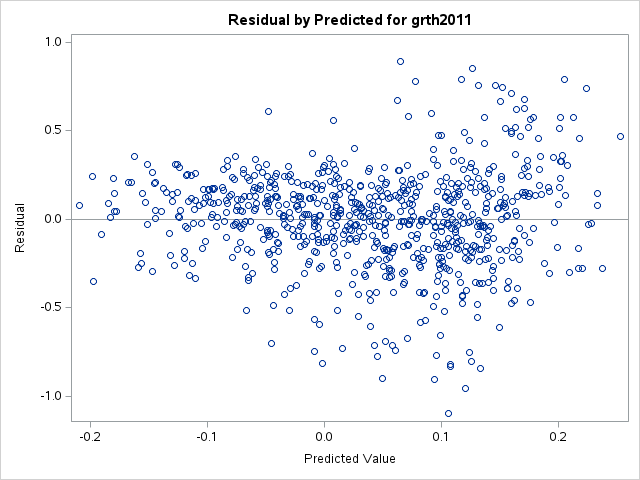
| **Parameter Estimates** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Label** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | **Standardized Estimate** | **Variance Inflation** |
| **Intercept** | Intercept | 1 | 0.30481 | 0.03810 | 8.00 | <.0001 | 0 | 0 |
| **oyrsact** | oyrsact | 1 | -0.00920 | 0.00133 | -6.92 | <.0001 | -0.26408 | 1.22061 |
| **distancetoHQ** | distancetoHQ | 1 | -0.00003647 | 0.00001469 | -2.48 | 0.0133 | -0.08672 | 1.02202 |
| **officesincounty** | officesincounty | 1 | -0.00045199 | 0.00083727 | -0.54 | 0.5895 | -0.01869 | 1.00367 |
| **ownmgrexp** | ownmgrexp | 1 | -0.00331 | 0.00122 | -2.72 | 0.0066 | -0.10312 | 1.20208 |

The above multiple regression model displays results based on the predicted dependent variable grth2011 (financial growth in revenue from 2010 to 2011) based on the independent variables of oyrsact (how many years has the franchise been a part of the network), distancetoHQ (the distance from the franchise location to the headquarters), officesincounty (the number of other offices located in the same county (competition)), and ownmgrexp (the total experience of the franchise’s leadership). Offid\_n (franchise ID number) and grth2010 (financial growth in revenue from 2009 to 2010) are not important to predict financial growth in 2011. The F value for the overall model is 21.83 and the p-value is less than 0.0001 which indicates the model is significant. The adjusted R squared value is 0.0995 which means the independent variables explain 9.95% of the variance in grth2011.

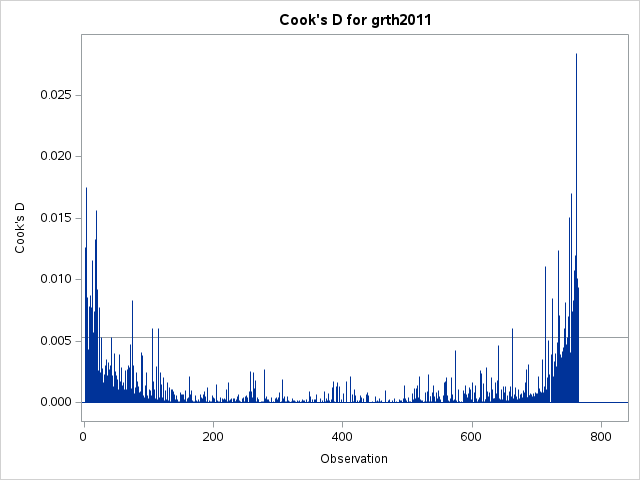
The t-value of the parameter estimate intercept is 8.00 and the p-value is less than 0.0001 deeming the intercept to be significant, however, it does not make sense based on the independent variables to interpret the intercept for grth2011. The t-value of oyrsact is -6.92 and the p-value is less than 0.0001 which means the number of years the franchise been a part of the network has a significant effect on predicting financial growth in 2011.The t-values and p-values for distancetoHQ and ownmgrexp are -2.48, -2.72 and 0.0133, 0.0066 respectively. We can conclude both distancetoHQ and ownmgrexp have a significant effect on predicting financial growth in 2011. On the other hand, the t-value and p-value for officesincounty are -0.54 and 0.5895 meaning officesincounty do not have a significant effect on predicting financial growth in 2011.

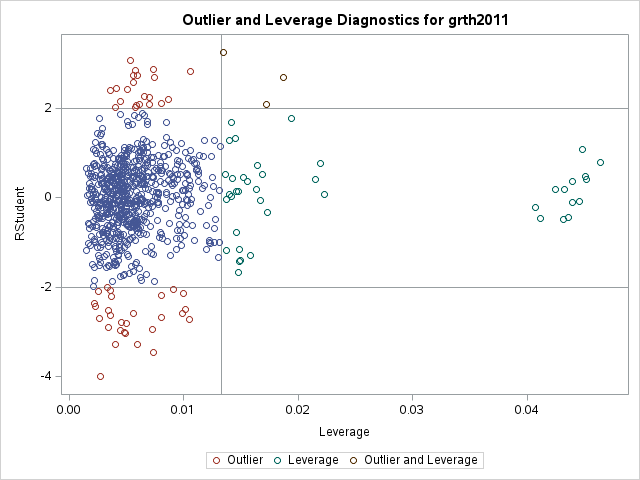
The variance inflation scores are all low and range from 1.00 – 1.22 which indicate the variables are not correlated with each other.





Assumptions: The errors (residuals) follow a normal distribution. Also, there is some concern regarding the possibility of heteroscedastic errors displayed by a potential megaphone effect in the Residual by Predicted for grth2011 plot. Autocorrelation is not of concern since we are not dealing with time series data.





There are multiple observations in Cook’s D that extend past the others. There is one observation on the right-hand side that significantly surpasses other observations and can be categorized as an influential observation. There are also three outlier and leverage points in the Outlier and Leverage Diagnostics for grth2011plot shown above which can be categorized as influential observations. These influential observations may need to be removed from the data and the model would need to be run again. Any modifications to the data need to be documented.

2. Using the world happiness dataset on sas named "happy" and explore if the following factors (Region, Freedom, Ladder, Social Support, Healthy Life Expectancy and GDP per capita) impact happiness. What can you conclude about the relative importance of these factors in influencing happiness? Are the results surprising? Why/Why not?

**Model: MODEL1**

**Dependent Variable: Happiness Happiness**

|  |  |
| --- | --- |
| **Number of Observations Read** | 156 |
| **Number of Observations Used** | 156 |

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 6 | 1.39924 | 0.23321 | 22.76 | <.0001 |
| **Error** | 149 | 1.52689 | 0.01025 |  |  |
| **Corrected Total** | 155 | 2.92613 |  |  |  |

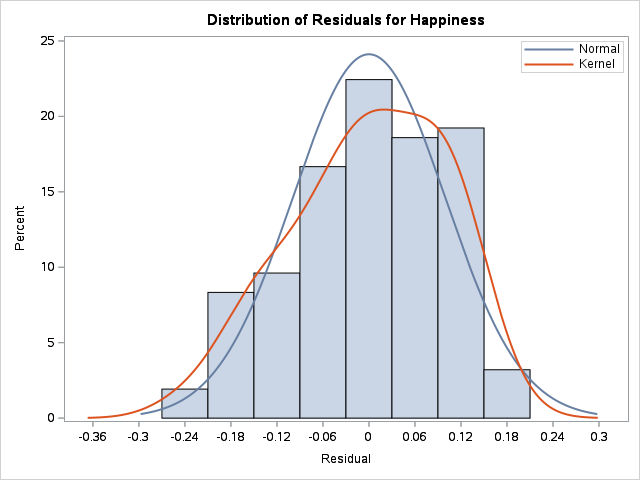
|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 0.10123 | **R-Square** | 0.4782 |
| **Dependent Mean** | 0.70066 | **Adj R-Sq** | 0.4572 |
| **Coeff Var** | 14.44786 |  |  |

| **Parameter Estimates** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Label** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | **Standardized Estimate** | **Variance Inflation** |
| **Intercept** | Intercept | 1 | 0.11643 | 0.09556 | 1.22 | 0.2250 | 0 | 0 |
| **Region** | Region | 1 | 0.00292 | 0.00323 | 0.90 | 0.3677 | 0.07220 | 1.82300 |
| **Ladder** | Ladder | 1 | 0.06425 | 0.01483 | 4.33 | <.0001 | 0.51513 | 4.03517 |
| **Social\_Support** | Social Support | 1 | -0.00615 | 0.09859 | -0.06 | 0.9504 | -0.00543 | 2.16944 |
| **Freedom** | Freedom | 1 | 0.46461 | 0.07250 | 6.41 | <.0001 | 0.47340 | 1.55830 |
| **GDP\_Per\_Capita** | GDP Per Capita | 1 | -0.00000177 | 9.306248E-7 | -1.90 | 0.0588 | -0.18118 | 2.58556 |
| **Healthy\_Life\_Expectancy** | Healthy Life Expectancy | 1 | -0.00146 | 0.00136 | -1.07 | 0.2841 | -0.11919 | 3.51015 |

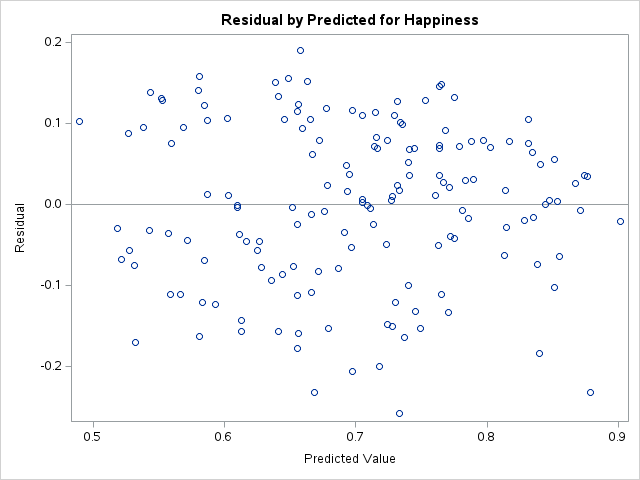
The above multiple regression model displays results based on the predicted dependent variable Happiness based on the independent variables Region, Freedom, Ladder, Social Support, Healthy Life Expectancy and GDP per capita. The F value for the overall model is 22.76 and the p-value is less than 0.0001 which indicates the model is significant. The adjusted R squared value is 0.4572 which means the independent variables explain 45.72% of the variance in Happiness. This is a relatively higher adjusted r squared value than we’ve seen so far from examples and I am not surprised. When we think and discuss about what brings happiness to people, the variables tested in this model are generally a part of that discussion.

The t-value of the parameter estimate intercept is 1.22 and the p-value is 0.2250 deeming the intercept to be not significant, however, it does not make sense based on the independent variables to interpret the intercept for Happiness. The t-values for Ladder and Freedom are 4.33 and 6.41 respectively. The p-values for Ladder and Freedom are both less than 0.0001 meaning both Ladder and Freedom have a significant impact on Happiness. All other parameter estimates (Region, Social Support, Healthy Life Expectancy and GDP per capita) have a p-value which is great than an alpha of 0.05 and therefore do not significantly impact Happiness. One independent variable to note is GDP per capita which has a t-value of -1.90 and a p-value of 0.0588 which is very close to being under the alpha value of 0.05.

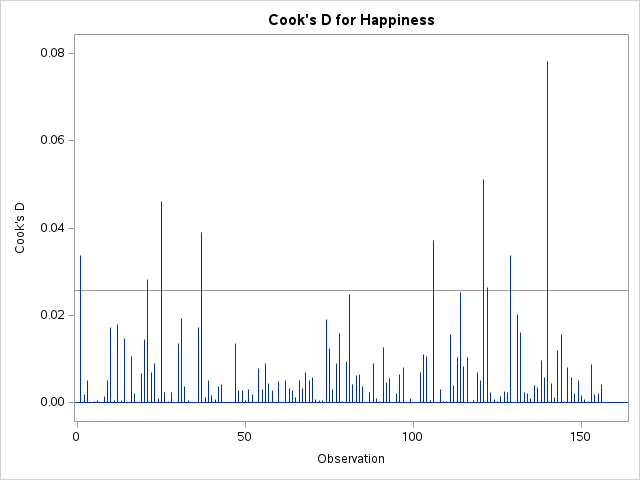
The variance inflation scores are all fairly low and range from 1.56 – 4.04 which indicate the variables are not correlated with each other.



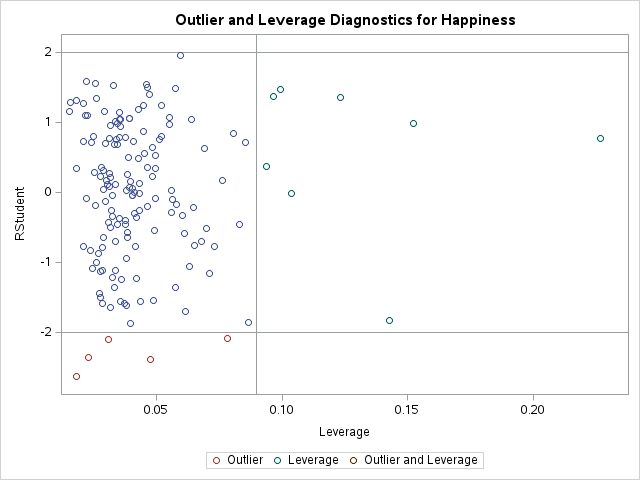
There is some concern here for a normal distribution. The graph is skewed to the left and we increase percent as residual (errors) increase for residual from 0.06 to 0.12 before a big drop off in percent in regards to residual 0.18.



There does not seem to be a megaphone effect. Heterogeneity of errors is not of concern nor is autocorrelation.



Cook’s D displays a few observations that are potential influential observations with one observation to the right being of most importance.



There are a few outliers and a few leverage points, however there are zero points of both outliers and leverage points. Therefore, we cannot remove observations with certainty, but outliers or leverage may be removed and documented to see the impact on the model. The removal of these observations may lead to GDP per capita becoming significant based on a lower p-value.

The results of this model were somewhat surprising. I expected the model to be a significant one, however, I did not expect only two variables to have a significant impact on happiness.

3. Using the HBAT\_200 dataset posted on blackboard, explore whether the following variables (X3 Firm Size, X6 Product Quality, X8 Tech Support, X9 Complaint Resolution, X17 Price Flexibility, and X18 Delivery Speed) impact X23(Purchase Level).

**Model: MODEL1**

**Dependent Variable: X23\_\_\_Purchase\_Level X23 - Purchase Level**

|  |  |
| --- | --- |
| **Number of Observations Read** | 200 |
| **Number of Observations Used** | 200 |

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 6 | 12230 | 2038.30449 | 104.40 | <.0001 |
| **Error** | 193 | 3768.17307 | 19.52421 |  |  |
| **Corrected Total** | 199 | 15998 |  |  |  |

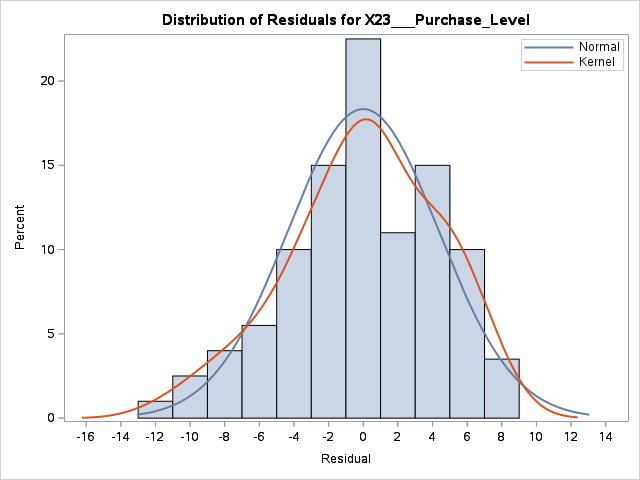
|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 4.41862 | **R-Square** | 0.7645 |
| **Dependent Mean** | 58.20000 | **Adj R-Sq** | 0.7571 |
| **Coeff Var** | 7.59213 |  |  |

| **Parameter Estimates** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Label** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | **Standardized Estimate** | **Variance Inflation** |
| **Intercept** | Intercept | 1 | 3.19472 | 3.07388 | 1.04 | 0.3000 | 0 | 0 |
| **X3\_\_\_Firm\_Size** | X3 - Firm Size | 1 | 1.74186 | 0.67128 | 2.59 | 0.0102 | 0.09736 | 1.15352 |
| **X6\_\_\_Product\_Quality** | X6 - Product Quality | 1 | 3.27481 | 0.28619 | 11.44 | <.0001 | 0.50513 | 1.59679 |
| **X8\_\_\_Technical\_Support** | X8 - Technical Support | 1 | -0.20246 | 0.19276 | -1.05 | 0.2949 | -0.03738 | 1.03756 |
| **X9\_\_\_Complaint\_Resolution** | X9 - Complaint Resolution | 1 | 0.25044 | 0.54441 | 0.46 | 0.6460 | 0.03380 | 4.42251 |
| **X17\_\_\_Price\_Flexibility** | X17 - Price Flexibility | 1 | -0.44252 | 0.40148 | -1.10 | 0.2717 | -0.05886 | 2.33682 |
| **X18\_\_\_Delivery\_Speed** | X18 - Delivery Speed | 1 | 7.85068 | 0.95242 | 8.24 | <.0001 | 0.65614 | 5.19192 |

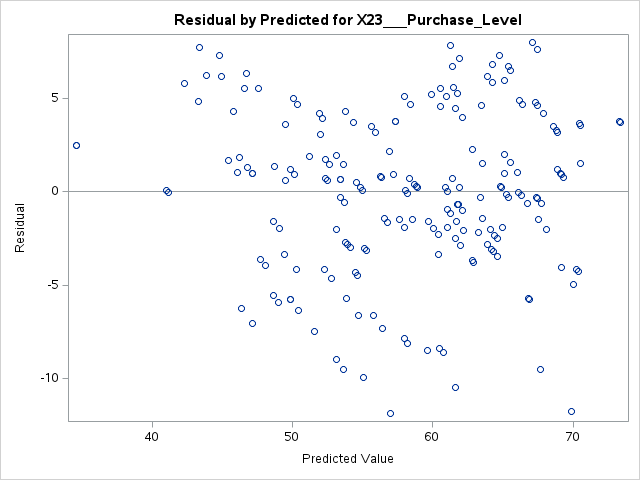
The above multiple regression model displays results based on the predicted dependent variable X23 Purchase Level based on the independent variables X3 Firm Size, X6 Product Quality, X8 Tech Support, X9 Complaint Resolution, X17 Price Flexibility, and X18 Delivery Speed. The F value for the overall model is 104.40 and the p-value is less than 0.0001 which indicates the model is significant. The adjusted R squared value is 0.7571 which means the independent variables explain 75.71% of the variance in Purchase Level.

The t-value of the parameter estimate intercept is 1.04 and the p-value is 0.3000 deeming the intercept to be not significant, however, it does not make sense based on the independent variables to interpret the intercept for purchase level. The t-value and p-value for firm size is 2.59 and 0.0102 resulting in firm size to significantly impact purchase level. The t-value and p-value for product quality is 11.44 and less than 0.0001 resulting in product quality to also significantly impact purchase level. The t-value and p-value for technical support is -1.05 and 0.2949 resulting in technical support not significantly impacting purchase level. Complaint resolution has a t-value of 0.46 and p-value of 0.6460 resulting in complaint resolution not significantly impacting purchase level. The t-value and p-value of price flexibility is -1.10 and 0.2717 resulting in price flexibility not significantly impacting purchase level. Lastly is delivery speed with a t-value of 8.24 and a p-value of less than 0.0001 resulting in delivery speed significantly impacting purchase level.

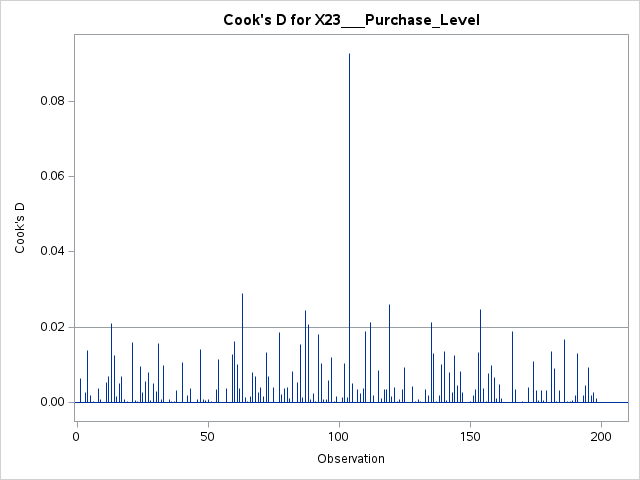
The variance inflation scores are for two of the variables: higher complaint resolution (4.42) and delivery speed (5.19). These variables may be correlated with each other. All other variables have a variance inflation score of 2.34 or less resulting in these variables to not be correlated with each other.

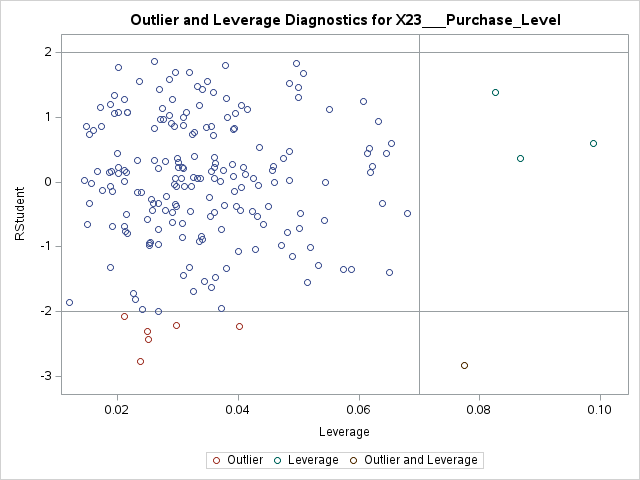


The graph shows a relatively normal distribution. There is a noticeable shift in percent between residual 2 and 4. The errors appear to be the opposite of what a normal distribution should be. All other residual/percents align with a normal distribution.



There is no megaphone effect. Heterogeneity is not of concern nor is autocorrelation.





Cook’s D and the Outlier and Leverage Diagnostics plot both display a potential influential observation #104. This observation should be removed and documented to see the impact it has on the model. This one observation has the potential to significantly change our model.

Also, complaint resolution or delivery should be removed from the model to test if there is an impact or not on the model.