Eric He

Regression & Multivariate Data Analysis

Homework 5

**Overview of the Data**

An important duty of the New York City Department of Parks and Recreation is the planting and maintaining of trees. Apart from looking nice, the contributions of the city's "urban forest" include intercepting contaminated stormwater that would otherwise flow into rivers and lakes and the Atlantic Ocean, moderating temperatures to be cooler in the summer and warmer in the winter, reducing noise levels, and absorbing and preventing the creation of pollutants and carbon dioxide[1]. According to the Department, these benefits generate a monetary value of $5.60 for every $1 spent on them[2].

To help its decision-making, the Department keeps track of the number, location, species, diameter, and condition of all the trees in the five boroughs of the city. The data gathered is publicly available; this homework uses data from the 1995 Street Tree Census, which can be obtained here[3]. The dataset has 516,968 data points, each corresponding to an individual tree, which I deemed to be too large for this analysis.

To keep the amount of data points low, I decided to look at a specific species of tree called *Robinia pseudoacacia*, commonly known as black locust. The tree is native to the eastern United States and generally grows in younger forests because it is shade-intolerant; it grows quickly and  reproduces using both flowers and basal shoots[4]. Because its flowers can secrete a premium honey and its wood is considered the toughest in the United States, the black locust is a viable product for tree farms.

It may not be the best tree for the urban forest, however. The 1995 census records 385 instances of this tree, a relatively small proportion of the total. This is because the black locust is

---

[1] https://tree-map.nycgovparks.org/learn/benefits
[2] https://www.nycgovparks.org/pagefiles/82/streamlining-tree-selection-in-nyc.pdf
[3] https://data.cityofnewyork.us/Environment/1995-Street-Tree-Census/kyad-zm4j
[4] https://en.wikipedia.org/wiki/Robinia_pseudoacacia

not on the list of street trees approved for planting[5], and thus it is not deliberately planted by the Department. Other species make up a substantial amount of the trees planted; the London planetree, for example, makes up 88,040 of the total because of its ability to grow in urban conditions and provide a thick cover of shade.

Even though the black locust is not systematically planted, it is not deliberately cut down either and the black locust is able to reproduce naturally within the city; a check of the New York City Street Tree Map[6] shows that in 2016 there are 1,903 instances of the tree, almost 1600 more than in 1995.

The dataset records, among other things, the borough (Manhattan, Queens, Brooklyn, the Bronx, and Staten Island) in which the tree was found and the condition the tree was in (unknown, dead, poor, fair, good, excellent). The distribution of trees is as follows.

| Unknown | Dead | Poor | Fair | Good | Excellent | Total |
|---|---|---|---|---|---|---|
| 2 | 3 | 25 | 1 | 256 | 98 | 385 |

| Manhattan | Queens | Brooklyn | Bronx | Staten Island | Total |
|---|---|---|---|---|---|
| 127 | 81 | 31 | 125 | 21 | 385 |

I excluded the unknowns from the dataset. Because there were only 3 dead trees and 1 fair tree, I moved the trees from those categories into Poor and Good, respectively, to prevent the dead and fair categories exercise an undue amount of leverage on the ANOVA. The design is still unbalanced with wildly different numbers of trees across any given category. The following page shows descriptive statistics for each individual subgroup.

[5] https://www.nycgovparks.org/pagefiles/52/Street-Trees-List-For-Permits.pdf
[6] https://tree-map.nycgovparks.org/#speciesId-74464

**Subgroup statistics**

### Descriptive Statistics: Diameter

#### Results for Condition = 2Poor

| Variable | Borough | N | Mean | SE Mean | StDev | Minimum | Q1 | Median | Q3 | Maximum |
|---|---|---|---|---|---|---|---|---|---|---|
| Diameter | Bronx | 17 | 16.88 | 2.86 | 11.78 | 2.00 | 9.00 | 15.00 | 25.00 | 45.00 |
| | Manhattan | 6 | 6.67 | 1.12 | 2.73 | 3.00 | 3.75 | 7.00 | 9.25 | 10.00 |
| | Queens | 5 | 9.40 | 2.96 | 6.62 | 0.00 | 3.50 | 10.00 | 15.00 | 18.00 |

#### Results for Condition = 3Good

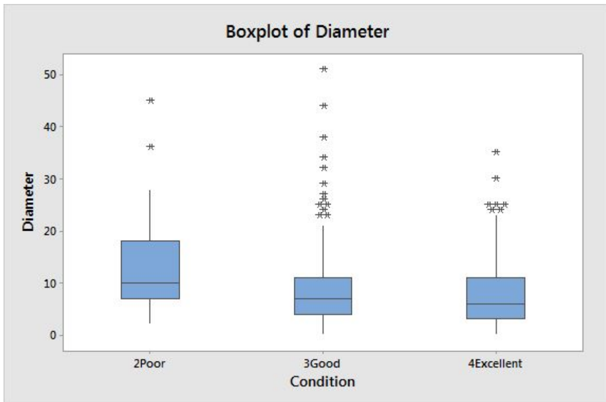
| Variable | Borough | N | Mean | SE Mean | StDev | Minimum | Q1 | Median | Q3 | Maximum |
|---|---|---|---|---|---|---|---|---|---|---|
| Diameter | Bronx | 86 | 8.558 | 0.719 | 6.670 | 2.000 | 3.000 | 6.500 | 13.000 | 32.000 |
| | Brooklyn | 14 | 10.21 | 2.92 | 10.91 | 3.00 | 3.00 | 7.00 | 13.50 | 44.00 |
| | Manhattan | 82 | 6.768 | 0.424 | 3.840 | 0.000 | 4.000 | 6.000 | 10.000 | 19.000 |
| | Queens | 64 | 10.28 | 1.03 | 8.25 | 1.00 | 5.00 | 9.00 | 11.75 | 51.00 |
| | Staten Island | 11 | 14.55 | 2.96 | 9.82 | 4.00 | 8.00 | 11.00 | 19.00 | 38.00 |

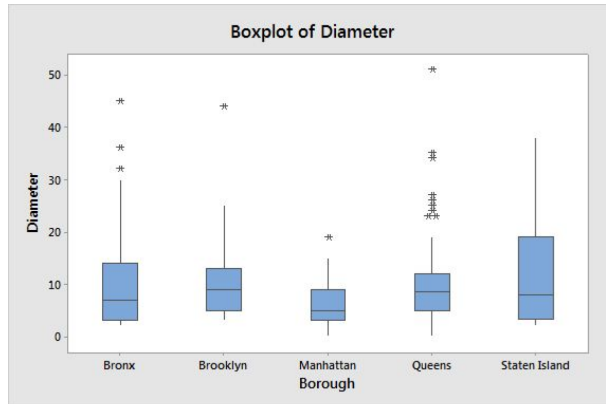
#### Results for Condition = 4Excellent

| Variable | Borough | N | Mean | SE Mean | StDev | Minimum | Q1 | Median | Q3 | Maximum |
|---|---|---|---|---|---|---|---|---|---|---|
| Diameter | Bronx | 21 | 7.43 | 1.57 | 7.19 | 2.00 | 2.50 | 7.00 | 8.00 | 30.00 |
| | Brooklyn | 17 | 10.88 | 1.32 | 5.43 | 3.00 | 6.50 | 11.00 | 13.50 | 25.00 |
| | Manhattan | 39 | 5.564 | 0.554 | 3.463 | 2.000 | 3.000 | 4.000 | 9.000 | 13.000 |
| | Queens | 12 | 11.50 | 3.05 | 10.56 | 0.00 | 4.50 | 7.00 | 20.50 | 35.00 |
| | Staten Island | 9 | 8.00 | 3.09 | 9.26 | 2.00 | 2.00 | 3.00 | 15.50 | 24.00 |

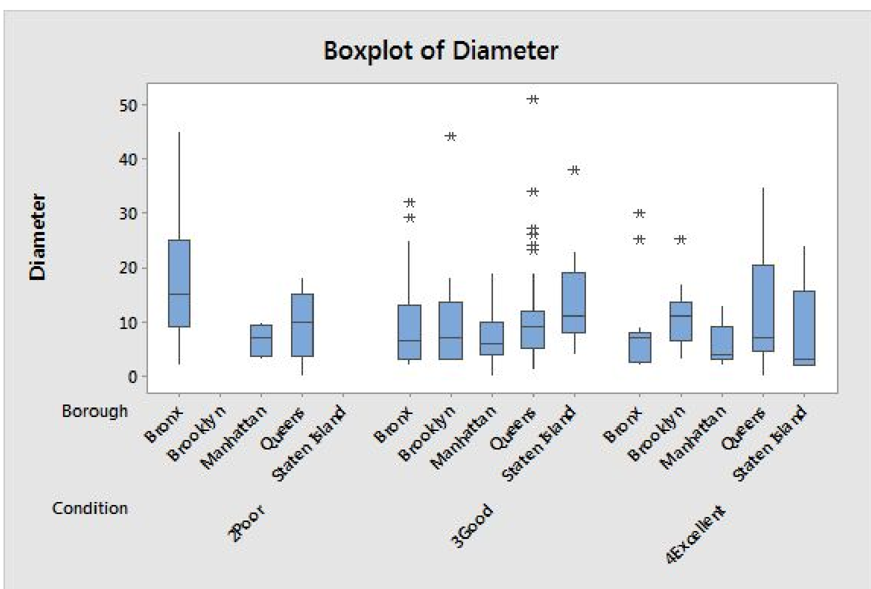
The small amount of black locusts in poor condition results in even smaller subgroups when they are divided by borough. There are only 5 black locusts with a condition of poor in Queens and 6 in Manhattan; there are only 9 black locusts with a condition of excellent in Staten island. These groups will have higher leverage values in the ANOVA than the other groups. Two groups with no data points are Staten Island and Brooklyn black locusts in poor conditions; there are simply no black locusts in poor condition recorded in those areas. This means that it is impossible to fit an interaction effect even if I tried.

Across all three conditions, the borough of Manhattan has consistently smaller average diameters (in inches) of black locust trees. Though this could be due to random chance, such a pattern would not be too surprising given Manhattan accepts less spacious sidewalks for bigger buildings.

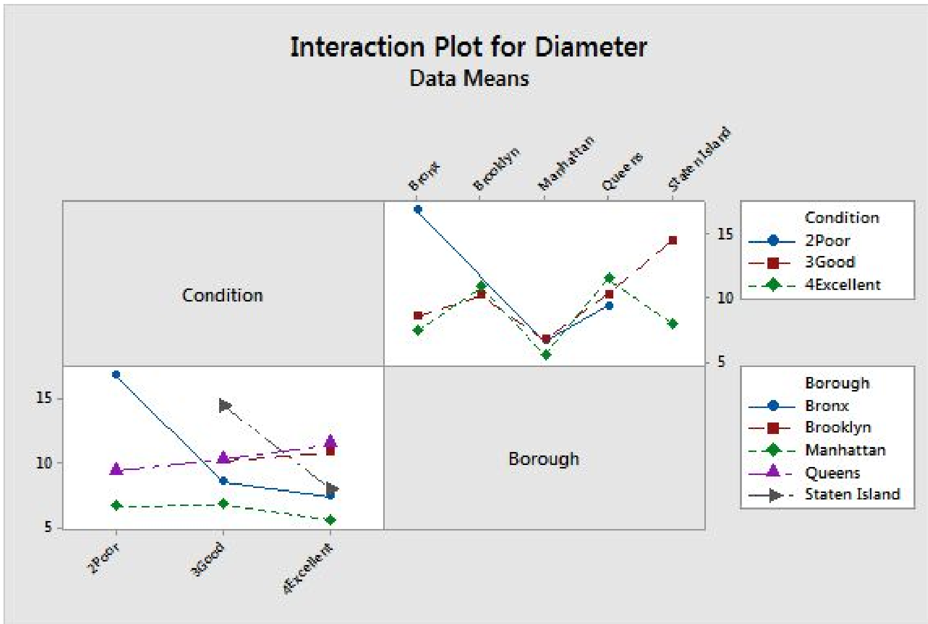



Trees in good condition appear to have the largest diameters, but trees in poor condition have much wider variance than either trees in good or excellent conditions. For the boroughs, Staten Island has a much wider interquartile range than the other four regions and the Bronx has a slightly wider range.




The box plots of tree diameters sorted by condition and borough show varying sizes, implying nonconstant variance, but no conclusion can be made from the box plots alone. There are a number of outliers in the category of trees in good condition with diameters far above the interquartile range.




Though the lines in the interaction plot for condition appear to be different from each other, this could be entirely due to chance. I do not see any obvious reason why there would be an interaction effect between borough and tree condition on diameter. I stuck with effect codings in the following two-way ANOVA because there was no reason to single a group out to be a reference.

**<u>Diameter ~ Condition + Borough</u>**

## General Linear Model: Diameter versus Condition, Borough

Method

Factor coding (-1, 0, +1)

Factor Information

| Factor | Type | Levels | Values |
|---|---|---|---|
| Condition | Fixed | 3 | 2Poor, 3Good, 4Excellent |
| Borough | Fixed | 5 | Bronx, Brooklyn, Manhattan, Queens, Staten Island |

Analysis of Variance

| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Condition | 2 | 654.4 | 327.19 | 6.54 | 0.002 |
| Borough | 4 | 1243.5 | 310.88 | 6.22 | 0.000 |
| Error | 376 | 18802.8 | 50.01 | | |
| Lack-of-Fit | 6 | 713.2 | 118.87 | 2.43 | 0.026 |
| Pure Error | 370 | 18089.6 | 48.89 | | |
| Total | 382 | 20717.7 | | | |

Model Summary

| S | R-sq | R-sq(adj) | R-sq(pred) |
|---|---|---|---|
| 7.07160 | 9.24% | 7.79% | 4.74% |

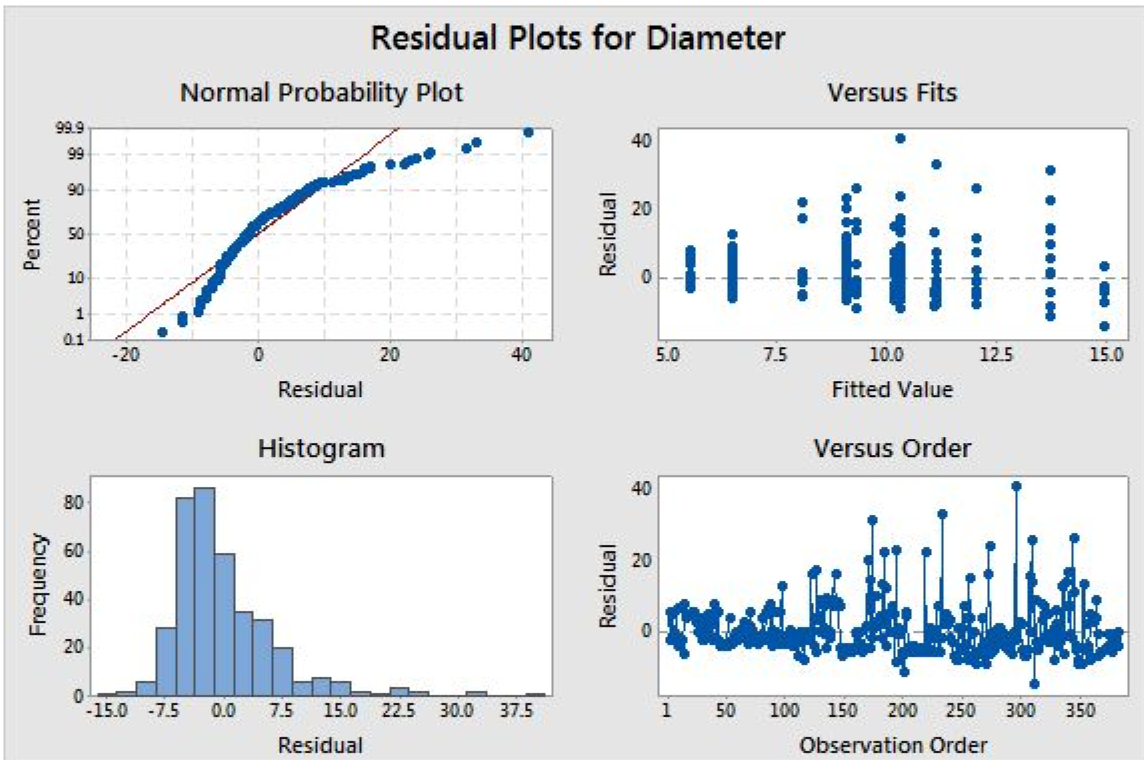
Coefficients

| Term | Coef | SE Coef | T-Value | P-Value | VIF |
|---|---|---|---|---|---|
| Constant | 11.012 | 0.621 | 17.73 | 0.000 | |
| Condition | | | | | |
| 2Poor | 3.423 | 0.956 | 3.58 | 0.000 | 2.07 |
| 3Good | -1.231 | 0.588 | -2.09 | 0.037 | 2.00 |
| Borough | | | | | |
| Bronx | -0.748 | 0.702 | -1.07 | 0.287 | 1.14 |
| Brooklyn | 1.33 | 1.11 | 1.20 | 0.232 | 1.24 |
| Manhattan | -3.312 | 0.676 | -4.90 | 0.000 | 1.07 |
| Queens | 0.481 | 0.784 | 0.61 | 0.540 | 1.12 |

Regression Equation

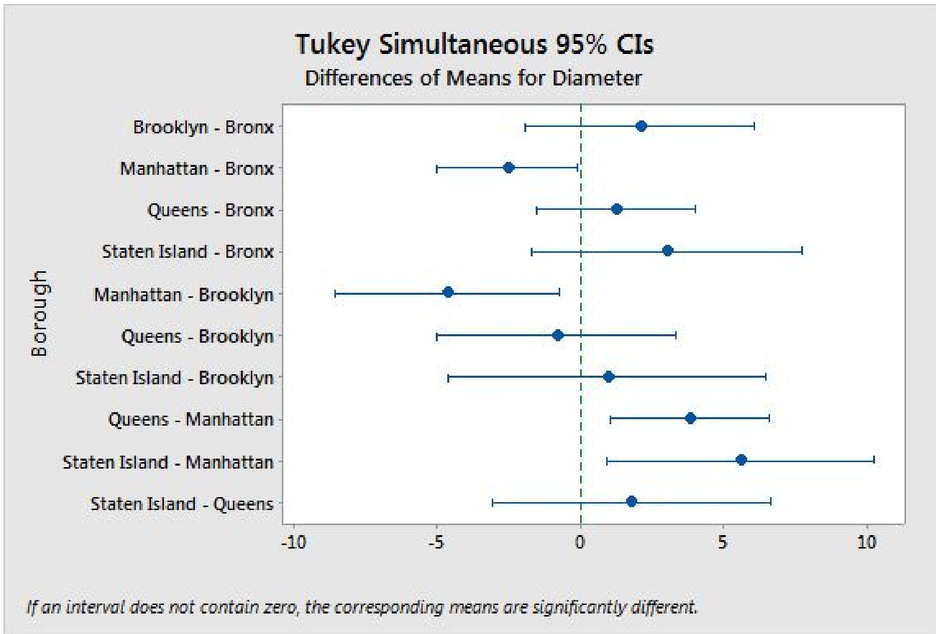
```
Diameter = 11.012 + 3.423 Condition_2Poor - 1.231 Condition_3Good
           - 2.193 Condition_4Excellent - 0.748 Borough_Bronx + 1.33 Borough_Brooklyn
           - 3.312 Borough_Manhattan + 0.481 Borough_Queens + 2.25 Borough_Staten Island
```

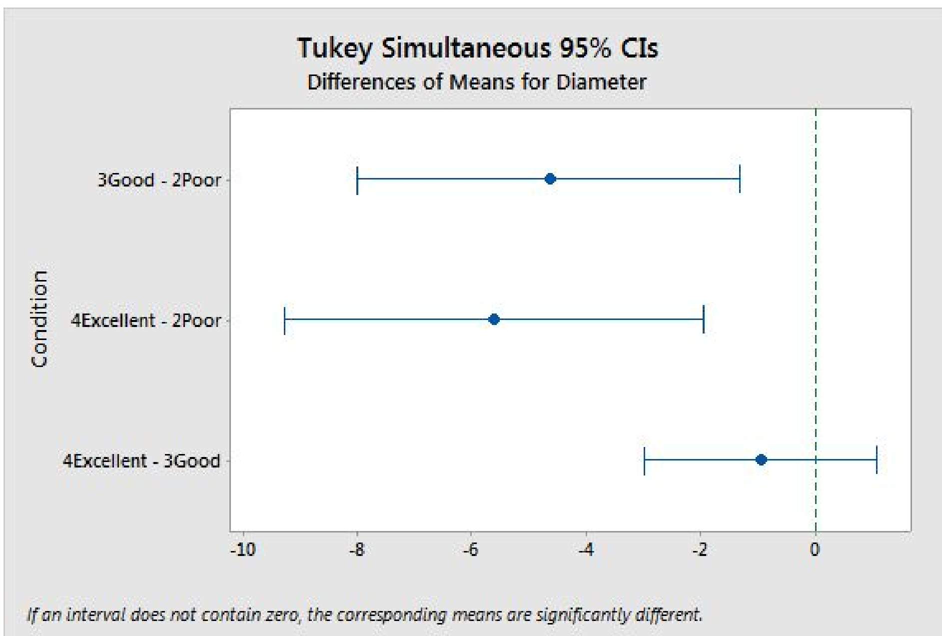
Both categorical variables are strongly significant at a 0.05 level, with tree condition having a p-value of 0.002 and borough having a p-value less than 0.001. The R-squared, however, is only 9.24%, meaning that only 9.24% of the variability of the data can be explained by the regression. The low R-squared value is corroborated by the lack-of-fit test, with a p-value of 0.026, signifies that the model is not capturing everything that is going on. It is possible to improve the model by factoring in any nonconstant variance that might exist across the different groups.




The residuals present several problems. There is evidence of nonconstant variance in the versus fits, with residuals having a wider spread towards the middle and right. The normal probability plot shows the residuals have a long right tail; many problem points can be seen in the top right and bottom left. Again, this may be due to nonconstant variance between groups, but logging the diameter may help as well.

**<u>Comparison of Means: Diameter ~ Condition + Borough</u>**



### Tukey Pairwise Comparisons: Response = Diameter, Term = Condition

Grouping Information Using the Tukey Method and 95% Confidence

| Condition | N | Mean | Grouping |
|---|---|---|---|
| 2Poor | 28 | 14.4352 | A |
| 3Good | 257 | 9.7812 | B |
| 4Excellent | 98 | 8.8192 | B |

Means that do not share a letter are significantly different.

Tukey Simultaneous Tests for Differences of Means

| Difference of Condition Levels | Difference of Means | SE of Difference | Simultaneous 95% CI | T-Value | Adjusted P-Value |
|---|---|---|---|---|---|
| 3Good - 2Poor | -4.65 | 1.42 | ( -7.99, -1.32) | -3.27 | 0.003 |
| 4Excellent - 2Poor | -5.62 | 1.57 | ( -9.28, -1.95) | -3.58 | 0.001 |
| 4Excellent - 3Good | -0.962 | 0.870 | (-2.997, 1.073) | -1.11 | 0.510 |

Individual confidence level = 98.02%

### Tukey Pairwise Comparisons: Response = Diameter, Term = Borough

Grouping Information Using the Tukey Method and 95% Confidence

| Borough | N | Mean | Grouping |
|---|---|---|---|
| Staten Island | 20 | 13.2636 | A |
| Brooklyn | 31 | 12.3389 | A |
| Queens | 81 | 11.4933 | A |
| Bronx | 124 | 10.2636 | A |
| Manhattan | 127 | 7.6999 | B |

Means that do not share a letter are significantly different.

Tukey Simultaneous Tests for Differences of Means

| Difference of Borough Levels | Difference of Means | SE of Difference | Simultaneous 95% CI | T-Value | Adjusted P-Value |
|---|---|---|---|---|---|
| Brooklyn - Bronx | 2.08 | 1.46 | ( -1.92, 6.07) | 1.42 | 0.617 |
| Manhattan - Bronx | -2.564 | 0.908 | (-5.041, -0.087) | -2.82 | 0.038 |
| Queens - Bronx | 1.23 | 1.02 | ( -1.54, 4.00) | 1.21 | 0.746 |
| Staten Island - Bronx | 3.00 | 1.73 | ( -1.72, 7.72) | 1.74 | 0.412 |
| Manhattan - Brooklyn | -4.64 | 1.43 | ( -8.55, -0.73) | -3.24 | 0.011 |
| Queens - Brooklyn | -0.85 | 1.53 | ( -5.03, 3.34) | -0.55 | 0.982 |
| Staten Island - Brooklyn | 0.92 | 2.03 | ( -4.62, 6.47) | 0.46 | 0.991 |
| Queens - Manhattan | 3.79 | 1.01 | ( 1.02, 6.56) | 3.74 | 0.002 |
| Staten Island - Manhattan | 5.56 | 1.71 | ( 0.91, 10.22) | 3.26 | 0.010 |
| Staten Island - Queens | 1.77 | 1.79 | ( -3.10, 6.64) | 0.99 | 0.859 |

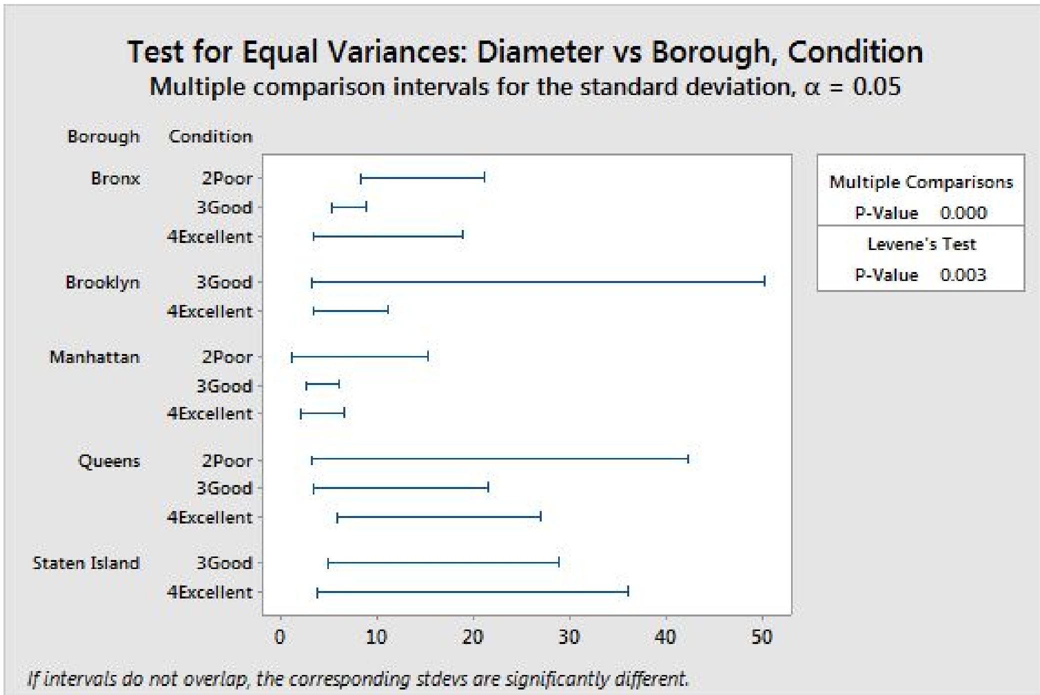
Individual confidence level = 99.34%

The Tukey Honestly Significant Difference Test says that there is no statistically significant difference between the diameters of trees in good condition and the diameters of trees in excellent condition. The diameters of trees in poor condition, on the other hand, have on average much larger diameters than trees in good and excellent condition. This suggests that older black locusts, which should have larger diameters than their younger counterparts, tend to be in much poorer shape.

There are several possible explanations for this. One is that older black locusts tend to be in the same area as other older trees, which tend to be larger, and suffer from having the sunlight they need taken by larger trees. Another explanation is given by the Wikipedia article for black locusts, that "young trees grow quickly and vigorously for a number of years, but soon become stunted and diseased, and rarely live long enough to attain any commercial value" due to locust borer insects eating the wood. Both reasons are also probably reasons why the black locust is not routinely planted by the Department.

When the trees are sorted by borough, the only group that stands out is Manhattan, which as noted earlier has black locusts with much smaller diameters than black locusts in the other four boroughs. The other four boroughs are not significantly different from each other at at a 0.05 level. The previously mentioned reason for Manhattan's uniqueness was the heavier space limitations placed on it, and another possible reason is that Manhattan is much more thorough than the other four boroughs in removing unhealthy trees. Because larger black locusts tend to be in poor health, they are quickly removed in Manhattan, and thus Manhattan's black locusts are typically smaller than those in other boroughs. A counterpoint to this second explanation, however, is that Manhattan has 6 of the 28 black locusts which are in poor health, while Staten Island has no black locusts in poor health and has on average the largest trees.

**Levene's Test: Diameter ~ Condition + Borough**




## General Linear Model: ABS SRES versus Condition, Borough

Method

Factor coding (-1, 0, +1)

Factor Information

| Factor | Type | Levels | Values |
|---|---|---|---|
| Condition | Fixed | 3 | 2Poor, 3Good, 4Excellent |
| Borough | Fixed | 5 | Bronx, Brooklyn, Manhattan, Queens, Staten Island |

Analysis of Variance

| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Condition | 2 | 3.505 | 1.7525 | 3.81 | 0.023 |
| Borough | 4 | 14.423 | 3.6059 | 7.84 | 0.000 |
| Error | 376 | 172.954 | 0.4600 | | |
| Lack-of-Fit | 6 | 3.822 | 0.6370 | 1.39 | 0.216 |
| Pure Error | 370 | 169.132 | 0.4571 | | |
| Total | 382 | 191.271 | | | |

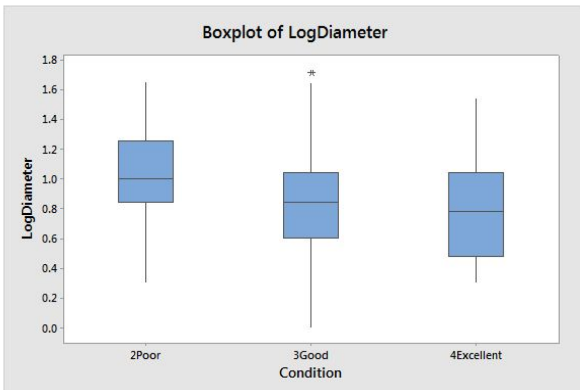
Model Summary

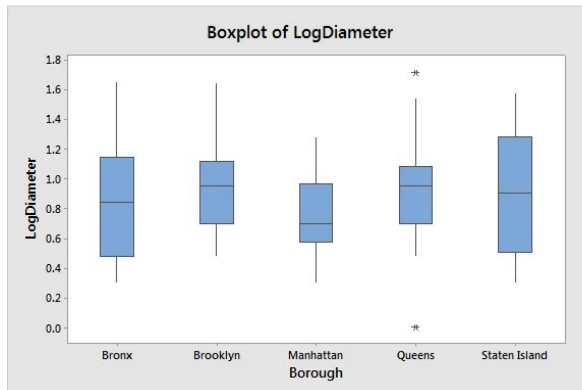
| S | R-sq | R-sq(adj) | R-sq(pred) |
|---|---|---|---|
| 0.678220 | 9.58% | 8.13% | 5.17% |

Levene’s test shows statistically significant f-values for both tree condition and borough, which means that both groups have nonconstant variance in their subgroups. I will first log the diameter, and if nonconstant variance persists, I will use weighted least squares.q

After taking logs, I found that 3 data points had a diameter of 0. They have been removed from the dataset, leaving a total of 380 points.

**Boxplots: Logged Diameter**




The variance for trees in poor condition, which was previously the largest by far, is now around the same size as trees in good condition; trees in excellent condition, on the other hand, have a noticeably wider interquartile range. The boxplots for logged diameters by borough look similar to their unlogged versions, with Staten Island and the Bronx having higher variances. This means that we may still have to use weighted least squares.

**LogDiameter ~ Condition + Borough**

Analysis of Variance

| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Condition | 2 | 1.395 | 0.69755 | 7.55 | 0.001 |
| Borough | 4 | 2.278 | 0.56950 | 6.17 | 0.000 |
| Error | 373 | 34.443 | 0.09234 | | |
| Lack-of-Fit | 6 | 1.378 | 0.22968 | 2.55 | 0.020 |
| Pure Error | 367 | 33.065 | 0.09009 | | |
| Total | 379 | 38.048 | | | |

Model Summary

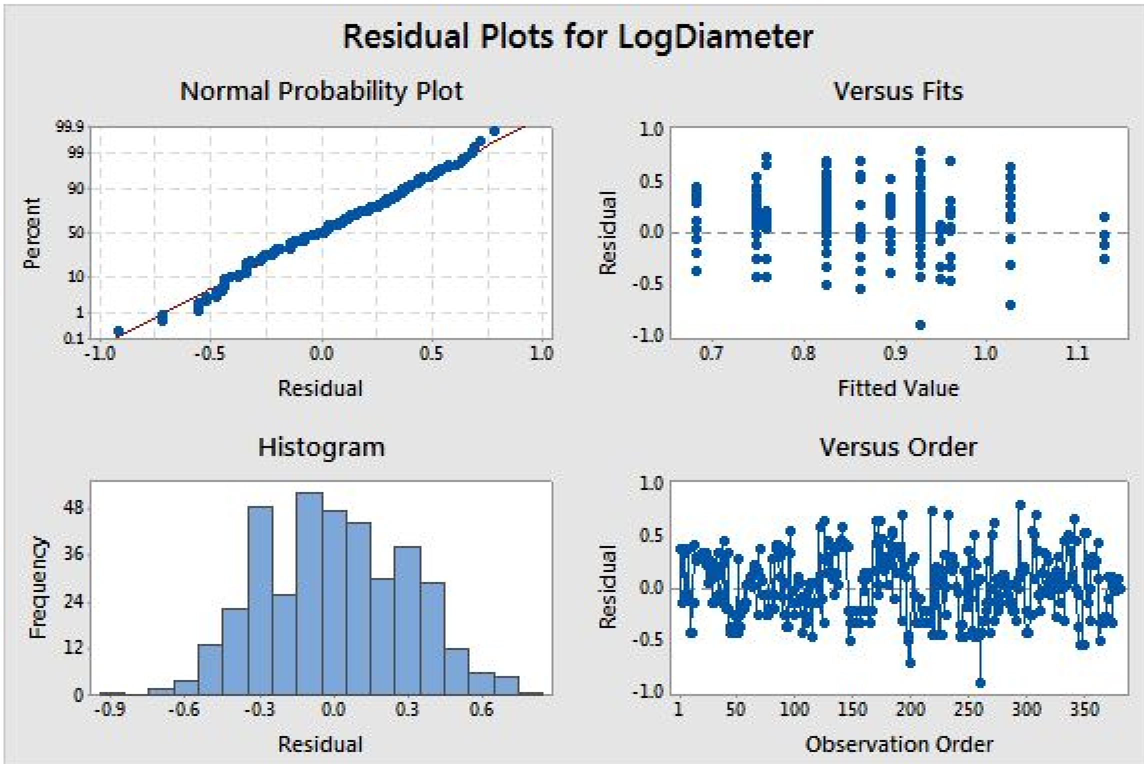
| S | R-sq | R-sq(adj) | R-sq(pred) |
|---|---|---|---|
| 0.303874 | 9.48% | 8.02% | 5.60% |

Coefficients

| Term | Coef | SE Coef | T-Value | P-Value | VIF |
|---|---|---|---|---|---|
| Constant | 0.9216 | 0.0270 | 34.11 | 0.000 | |
| Condition | | | | | |
| 2Poor | 0.1558 | 0.0418 | 3.73 | 0.000 | 2.10 |
| 3Good | -0.0456 | 0.0256 | -1.78 | 0.076 | 2.04 |
| Borough | | | | | |
| Bronx | -0.0528 | 0.0302 | -1.75 | 0.081 | 1.14 |
| Brooklyn | 0.0823 | 0.0477 | 1.73 | 0.085 | 1.25 |
| Manhattan | -0.1292 | 0.0291 | -4.43 | 0.000 | 1.07 |
| Queens | 0.0498 | 0.0340 | 1.46 | 0.145 | 1.13 |

Regression Equation

```
LogDiameter = 0.9216 + 0.1558 Condition_2Poor - 0.0456 Condition_3Good
              - 0.1101 Condition_4Excellent - 0.0528 Borough_Bronx + 0.0823 Borough_Brooklyn
              - 0.1292 Borough_Manhattan + 0.0498 Borough_Queens + 0.0499 Borough_Staten
              Island
```




The non-normality of the residuals mostly disappeared after taking the log of diameter; the residuals now have a bit of a short tail, with a few points scattered to the bottom left. The versus fits graphs shows much more similar variances than before, although the greater variance in the middle has not entirely disappeared.

The outlier towards the bottom of the middle of the versus fits graph and the lowest left point in the normal probability plot correspond to row 260. The next lowest outlier is actually two points, which correspond to next two lowest points in the normal probability plot; these are rows 199 and 200. The points in the top right, from highest to lowest, are rows 218, 232, 308, and 294. I'll rerun Levene's test to check again for nonconstant variance and see if I should use weighted least squares before taking a close look at the outliers because they may disappear after using weighted least squares.

**<u>Levene's Test: LogDiameter ~ Condition + Borough</u>**

**General Linear Model: ABS LOG SRES versus Condition, Borough**

Method

| | |
|---|---|
| Factor coding | (-1, 0, +1) |
| Rows unused | 2 |

Factor Information

| Factor | Type | Levels | Values |
|---|---|---|---|
| Condition | Fixed | 3 | 2Poor, 3Good, 4Excellent |
| Borough | Fixed | 5 | Bronx, Brooklyn, Manhattan, Queens, Staten Island |

Analysis of Variance

| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Condition | 2 | 0.004 | 0.00181 | 0.01 | 0.994 |
| Borough | 4 | 4.628 | 1.15702 | 3.72 | 0.006 |
| Error | 373 | 116.053 | 0.31113 | | |
| Lack-of-Fit | 6 | 3.139 | 0.52324 | 1.70 | 0.120 |
| Pure Error | 367 | 112.914 | 0.30767 | | |
| Total | 379 | 120.727 | | | |

Model Summary

| S | R-sq | R-sq(adj) | R-sq(pred) |
|---|---|---|---|
| 0.557795 | 3.87% | 2.32% | 0.00% |

With logged diameters, the previously strongly significant nonconstant variance across different tree conditions has now disappeared. With an f-value of 0.01, it is not anywhere near statistically significant. Trees sorted by borough is strongly significant with an f-value of 3.72, so I will still use weighted least squares.

**Weighted Least Squares: Logged Diameter ~ Borough + Condition**

### Test for Equal Variances: SRES_1 versus Borough, Condition

Method

| | |
|---|---|
| Null hypothesis | All variances are equal |
| Alternative hypothesis | At least one variance is different |
| Significance level | $\alpha = 0.05$ |

95% Bonferroni Confidence Intervals for Standard Deviations

| Borough | Condition | N | StDev | CI |
|---|---|---|---|---|
| Bronx | 2Poor | 17 | 1.29133 | (0.684258, 2.9362) |
| Bronx | 3Good | 86 | 1.05291 | (0.912782, 1.2568) |
| Bronx | 4Excellent | 21 | 1.14076 | (0.728966, 2.0701) |
| Brooklyn | 3Good | 14 | 1.21827 | (0.652596, 2.8660) |
| Brooklyn | 4Excellent | 17 | 0.78062 | (0.467665, 1.5699) |
| Manhattan | 2Poor | 6 | 0.69087 | (0.130589, 7.0525) |
| Manhattan | 3Good | 81 | 0.84672 | (0.727034, 1.0226) |
| Manhattan | 4Excellent | 39 | 0.86267 | (0.696610, 1.1538) |
| Queens | 2Poor | 4 | 0.57619 | (0.023935, 50.0070) |
| Queens | 3Good | 64 | 0.98042 | (0.730064, 1.3789) |
| Queens | 4Excellent | 11 | 1.08746 | (0.588680, 2.7249) |
| Staten Island | 3Good | 11 | 0.95806 | (0.513359, 2.4253) |
| Staten Island | 4Excellent | 9 | 1.52906 | (0.529354, 6.5064) |

Individual confidence level = 99.6154%

Weights were obtained by using the reciprocal of the square of the standard deviation for each group.

### General Linear Model: Sorted LogDiameter versus Sorted Condition, Sorted Borough

Method

| | |
|---|---|
| Factor coding | (-1, 0, +1) |
| Weights | wt |

Factor Information

| Factor | Type | Levels | Values |
|---|---|---|---|
| Sorted Condition | Fixed | 3 | 2Poor, 3Good, 4Excellent |
| Sorted Borough | Fixed | 5 | Bronx, Brooklyn, Manhattan, Queens, Staten Island |

Analysis of Variance

| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Sorted Condition | 2 | 0.9556 | 0.47781 | 5.18 | 0.006 |
| Sorted Borough | 4 | 3.4347 | 0.85866 | 9.30 | 0.000 |
| Error | 373 | 34.4234 | 0.09229 | | |
| Lack-of-Fit | 6 | 1.1435 | 0.19058 | 2.10 | 0.052 |
| Pure Error | 367 | 33.2800 | 0.09068 | | |
| Total | 379 | 38.7545 | | | |

Model Summary

| S | R-sq | R-sq(adj) | R-sq(pred) |
|---|---|---|---|
| 0.303789 | 11.18% | 9.75% | 7.78% |

```
Coefficients

Term                 Coef  SE Coef  T-Value  P-Value   VIF
Constant           0.9227   0.0256    35.97    0.000
Sorted Condition
  2Poor            0.1117   0.0369     3.03    0.003  1.94
  3Good           -0.0252   0.0235    -1.07    0.283  1.91
Sorted Borough
  Bronx           -0.0715   0.0316    -2.26    0.024  1.09
  Brooklyn         0.0975   0.0466     2.09    0.037  1.23
  Manhattan       -0.1507   0.0272    -5.53    0.000  1.06
  Queens           0.0265   0.0335     0.79    0.430  1.12


Regression Equation

Sorted LogDiameter = 0.9227 + 0.1117 Sorted Condition_2Poor - 0.0252 Sorted Condition_3Good
                     - 0.0865 Sorted Condition_4Excellent - 0.0715 Sorted Borough_Bronx
                     + 0.0975 Sorted Borough_Brooklyn - 0.1507 Sorted Borough_Manhattan
                     + 0.0265 Sorted Borough_Queens + 0.0982 Sorted Borough_Staten Island
```

The f-value for tree condition has dropped; this may be because the weighted least squares still used tree condition in addition to borough to divide the points into subgroups. I will recalculate the weights based only on borough.

**Weighted Least Squares on Borough: Logged Diameter ~ Borough + Condition**

### Test for Equal Variances: SRES_1 versus Borough

```
Method

Null hypothesis         All variances are equal
Alternative hypothesis  At least one variance is different
Significance level      α = 0.05


95% Bonferroni Confidence Intervals for Standard Deviations

      Borough    N    StDev           CI
        Bronx  124  1.09797  (0.97415, 1.26379)
     Brooklyn   31  1.03873  (0.80214, 1.46698)
    Manhattan  126  0.84941  (0.76402, 0.96405)
       Queens   79  0.98416  (0.78193, 1.28042)
Staten Island   20  1.34692  (1.03334, 2.01522)

Individual confidence level = 99%
```

## General Linear Model: LogDiameter versus Borough, Condition

Method

| | |
|---|---|
| Factor coding | (-1, 0, +1) |
| Weights | wt |

Factor Information

| Factor | Type | Levels | Values |
|---|---|---|---|
| Borough | Fixed | 5 | Bronx, Brooklyn, Manhattan, Queens, Staten Island |
| Condition | Fixed | 3 | 2Poor, 3Good, 4Excellent |

Analysis of Variance

| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Borough | 4 | 2.508 | 0.62697 | 6.89 | 0.000 |
| Condition | 2 | 1.116 | 0.55811 | 6.13 | 0.002 |
| Error | 373 | 33.963 | 0.09105 | | |
| Lack-of-Fit | 6 | 1.140 | 0.19008 | 2.13 | 0.050 |
| Pure Error | 367 | 32.822 | 0.08943 | | |
| Total | 379 | 37.641 | | | |

Model Summary

| S | R-sq | R-sq(adj) | R-sq(pred) |
|---|---|---|---|
| 0.301750 | 9.77% | 8.32% | 6.32% |

Coefficients

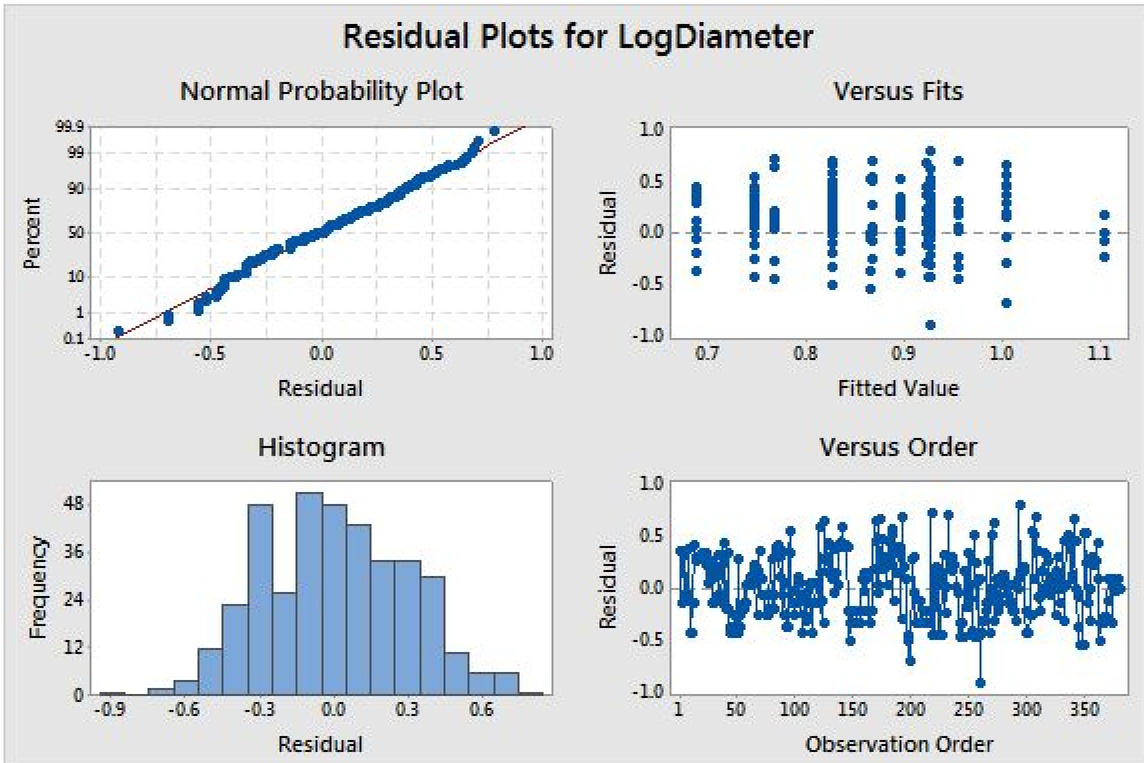
| Term | Coef | SE Coef | T-Value | P-Value | VIF |
|---|---|---|---|---|---|
| Constant | 0.9149 | 0.0296 | 30.90 | 0.000 | |
| Borough | | | | | |
| Bronx | -0.0497 | 0.0338 | -1.47 | 0.142 | 1.16 |
| Brooklyn | 0.0798 | 0.0503 | 1.59 | 0.113 | 1.08 |
| Manhattan | -0.1291 | 0.0295 | -4.37 | 0.000 | 1.13 |
| Queens | 0.0510 | 0.0357 | 1.43 | 0.154 | 1.12 |
| Condition | | | | | |
| 2Poor | 0.1379 | 0.0416 | 3.32 | 0.001 | 2.19 |
| 3Good | -0.0397 | 0.0253 | -1.57 | 0.117 | 2.14 |

Regression Equation

```
LogDiameter = 0.9149 - 0.0497 Borough_Bronx + 0.0798 Borough_Brooklyn
              - 0.1291 Borough_Manhattan + 0.0510 Borough_Queens + 0.0480 Borough_Staten
              Island + 0.1379 Condition_2Poor - 0.0397 Condition_3Good
              - 0.0982 Condition_4Excellent
```

The f-value for borough dropped back from 9.30 to 6.89 using the weightings stratified across different boroughs, while the f-value for tree condition went back up from 5.18 to 6.13. The logged diameters without using weightings has f-values of 7.55 for condition and 6.17 for borough. All three models have extremely significant predictors and the coefficients are all very similar to each other.

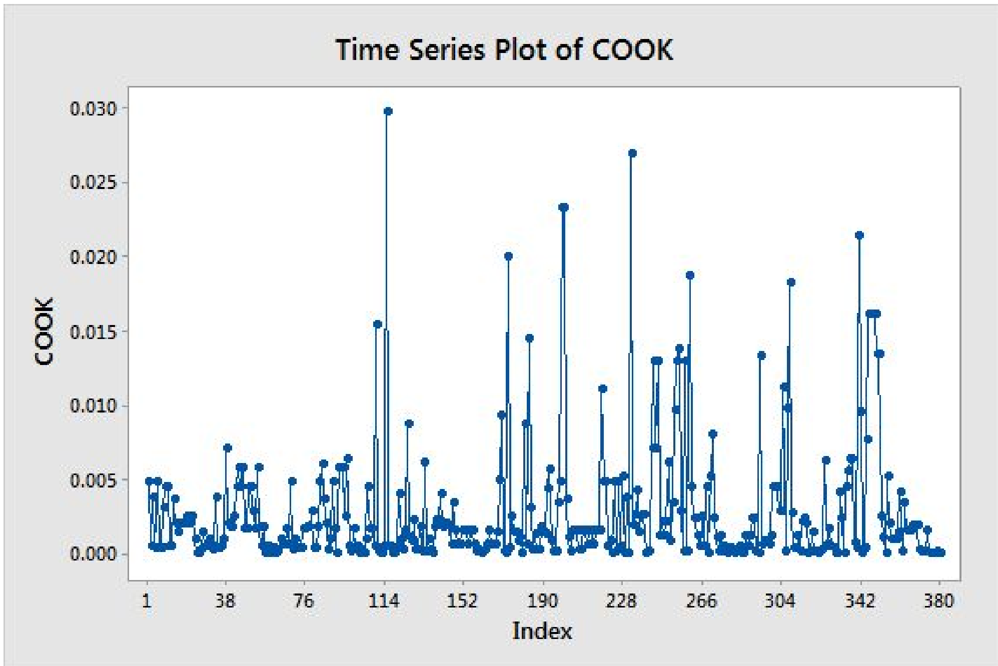
**Analysis of Residuals of WLS on Borough**




The short tail of the residuals looks almost exactly the same, and the previously mentioned notable points are all the same: 260, 199, and 200 in the lower left and 218, 232, 308, and 294 in the top right.

| | | | |
|---|---|---|---|
| 199 | Poor | Bronx | -2.42880 |
| 200 | Poor | Bronx | -2.42880 |
| 260 | Good | Queens | -3.06665 |
| 218 | Excellent | Bronx | 2.38666 |
| 232 | Good | Brooklyn | 2.23761 |
| 294 | Good | Queens | 2.64890 |
| 308 | Excellent | Queens | 2.30472 |

There are not any discernible patterns in the outliers, with trees in all conditions and boroughs.

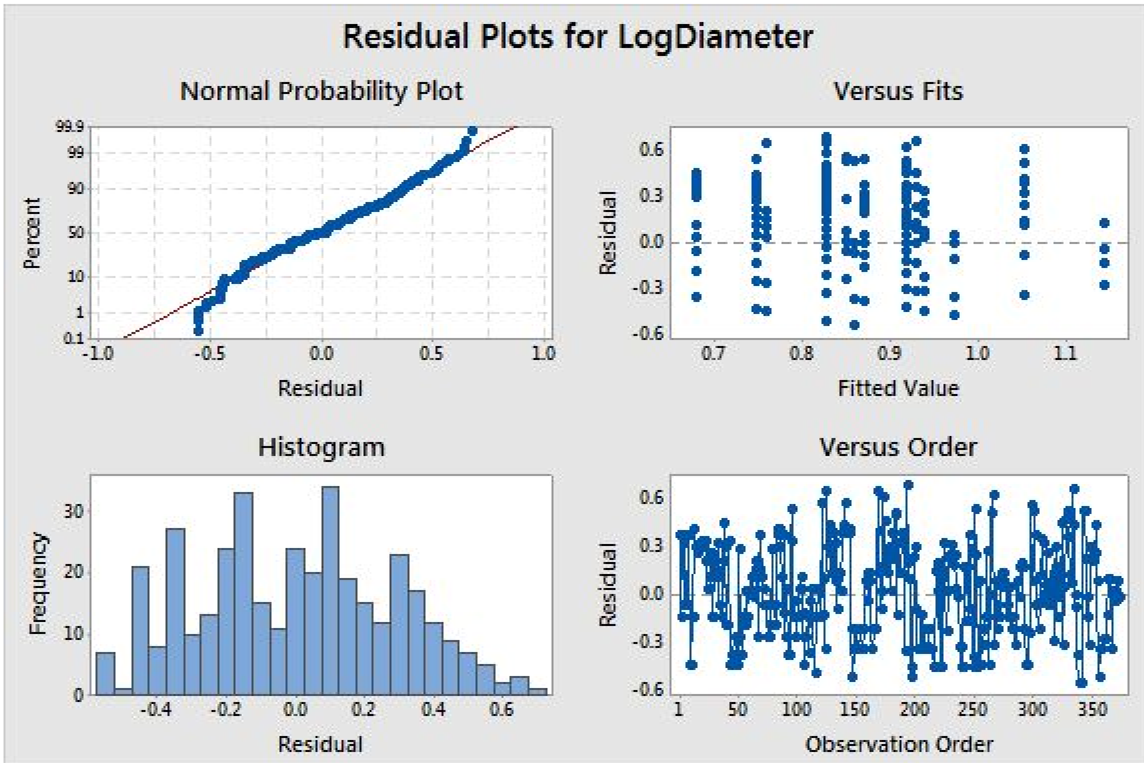



Points 115, 232, 199, 200, 341, and 173 are influential points according to Cook's distance function. Including points 199 and 200, 4 of the 6 points are trees in poor condition. This is because there are not many trees in poor condition and thus the trees in poor condition tend to have high leverage and thus also higher influence.

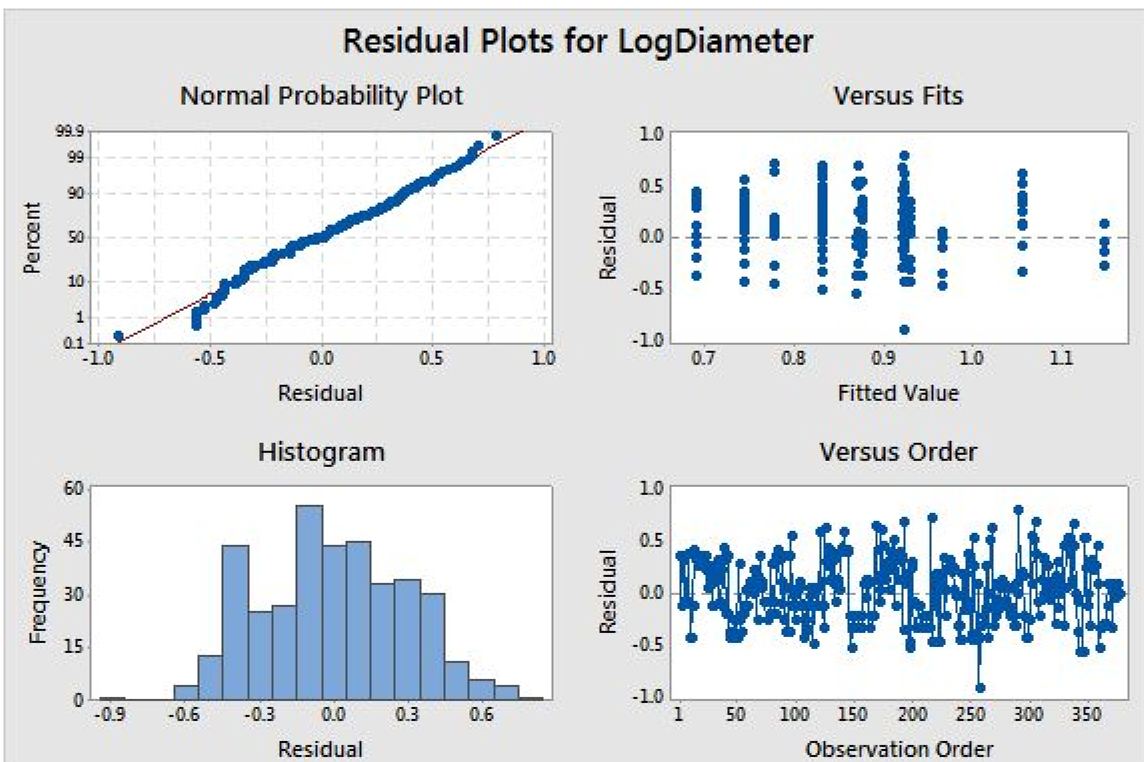
| | | |
|---|---|---|
| 115 | Poor | Manhattan |
| 173 | Poor | Bronx |
| 341 | Good | Staten Island. |

I will remove the top four influential points and the aforementioned outliers (232, 199 and 200 are both influential and outliers in this case) and rerun the ANOVA.

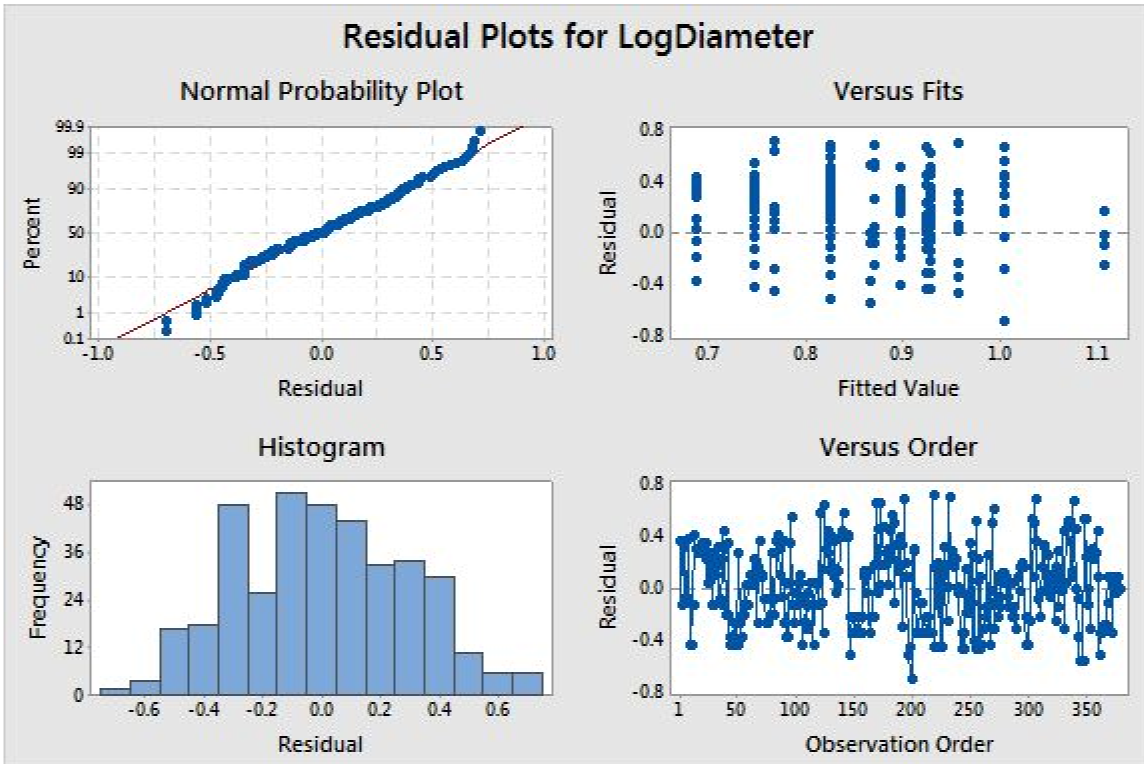
**Models with Outliers and Influential Points Removed**




The short tail has grown markedly worse when the points are removed.




Residuals are similar to what they were before when only points that were both outliers and influential points were removed.




Removing just the most bottom left and top right points only serves to make the short tail more pronounced in the normal probability plot.

Coefficients for the three models with removed outliers are shown below, and the fourth set is the coefficients for the weighted least squares with no points removed again. They are highly similar to each other and the f-values continue to be extremely significant for all of the models.

Coefficients

| Term | Coef | SE Coef | T-Value | P-Value | VIF |
|---|---|---|---|---|---|
| Constant | 0.9235 | 0.0287 | 32.23 | 0.000 | |
| Borough | | | | | |
| Bronx | -0.0445 | 0.0324 | -1.37 | 0.171 | 1.16 |
| Brooklyn | 0.0662 | 0.0486 | 1.36 | 0.174 | 1.09 |
| Manhattan | -0.1245 | 0.0282 | -4.41 | 0.000 | 1.13 |
| Queens | 0.0465 | 0.0345 | 1.35 | 0.178 | 1.12 |
| Condition | | | | | |
| 2Poor | 0.1726 | 0.0407 | 4.25 | 0.000 | 2.23 |
| 3Good | -0.0521 | 0.0246 | -2.12 | 0.035 | 2.19 |

Coefficients

| Term | Coef | SE Coef | T-Value | P-Value | VIF |
|---|---|---|---|---|---|
| Constant | 0.9260 | 0.0296 | 31.29 | 0.000 | |
| Borough | | | | | |
| Bronx | -0.0387 | 0.0334 | -1.16 | 0.248 | 1.16 |
| Brooklyn | 0.0598 | 0.0502 | 1.19 | 0.234 | 1.09 |
| Manhattan | -0.1268 | 0.0291 | -4.35 | 0.000 | 1.14 |
| Queens | 0.0540 | 0.0352 | 1.54 | 0.126 | 1.12 |
| Condition | | | | | |
| 2Poor | 0.1668 | 0.0420 | 3.97 | 0.000 | 2.26 |
| 3Good | -0.0571 | 0.0254 | -2.25 | 0.025 | 2.21 |

Coefficients

| Term | Coef | SE Coef | T-Value | P-Value | VIF |
|---|---|---|---|---|---|
| Constant | 0.9151 | 0.0290 | 31.53 | 0.000 | |
| Borough | | | | | |
| Bronx | -0.0501 | 0.0332 | -1.51 | 0.132 | 1.16 |
| Brooklyn | 0.0795 | 0.0493 | 1.61 | 0.108 | 1.08 |
| Manhattan | -0.1295 | 0.0289 | -4.47 | 0.000 | 1.13 |
| Queens | 0.0524 | 0.0352 | 1.49 | 0.137 | 1.11 |
| Condition | | | | | |
| 2Poor | 0.1377 | 0.0407 | 3.38 | 0.001 | 2.18 |
| 3Good | -0.0395 | 0.0248 | -1.59 | 0.112 | 2.13 |

Current Model Coefficients

Coefficients

| Term | Coef | SE Coef | T-Value | P-Value | VIF |
|---|---|---|---|---|---|
| Constant | 0.9149 | 0.0296 | 30.90 | 0.000 | |
| Borough | | | | | |
| Bronx | -0.0497 | 0.0338 | -1.47 | 0.142 | 1.16 |
| Brooklyn | 0.0798 | 0.0503 | 1.59 | 0.113 | 1.08 |
| Manhattan | -0.1291 | 0.0295 | -4.37 | 0.000 | 1.13 |
| Queens | 0.0510 | 0.0357 | 1.43 | 0.154 | 1.12 |
| Condition | | | | | |
| 2Poor | 0.1379 | 0.0416 | 3.32 | 0.001 | 2.19 |
| 3Good | -0.0397 | 0.0253 | -1.57 | 0.117 | 2.14 |

I will just stick with this model because all the models are extremely similar to each other and removing outliers has little effect on the f-scores and the residuals are more normally distributed in this model.

**ANOVA Output Interpretation**

```
Regression Equation

LogDiameter = 0.9149 - 0.0497 Borough_Bronx + 0.0798 Borough_Brooklyn
              - 0.1291 Borough_Manhattan + 0.0510 Borough_Queens + 0.0480 Borough_Staten
              Island + 0.1379 Condition_2Poor - 0.0397 Condition_3Good
              - 0.0982 Condition_4Excellent
```

LogDiameter is taken using log base 10, so:
-The constant term of .9149 corresponds to a diameter of 10^(.9149)=**8.221**. This diameter, in inches, is the average for all the trees in the dataset.

When looking at black locusts without regard to tree condition:
-being in **the Bronx is associated with a** multiplicative factor of 10^-.0497= 0.8918 on diameter. This corresponds to a **10.82 percent smaller diameter** than the average tree in the dataset.
-being in **Brooklyn is associated with a** multiplicative factor of 10^0.0798= 1.0120711 on diameter. This corresponds to a **1.20711 percent larger diameter** than the average tree in the dataset.
-being in **Manhattan is associated with a** multiplicative factor of 10^-.1291= 0.74828 on diameter. This corresponds to a **25.172 percent smaller diameter** than the average tree in the dataset.
-being in **Queens is associated with a** multiplicative factor of 10^0.0510= 1.124605 on diameter. This corresponds to a **12.4605 percent larger diameter** than the average tree in the dataset.
-being in **Staten Island is associated with a** multiplicative factor of 10^0.048= 1.116863 on diameter. This corresponds to a **11.6863 percent larger diameter** than the average tree in the dataset.

When looking at black locusts without regard to location:
-being in **poor condition is associated with a** multiplicative factor of 10^0.1379=1.3737 on diameter. This corresponds to a **37.37 percent larger diameter** than the average tree in the dataset.
-being in **good condition is associated with a** multiplicative factor of 10^-.0397=0.91264 on diameter. This corresponds to a **8.736 percent smaller diameter** than the average tree in the dataset.
-being in **excellent condition is associated with a** multiplicative factor of 10^-0.0982=0.79762 on diameter. This corresponds to a **20.238 percent smaller diameter** than the average tree in the dataset.

**General Linear Model: LogDiameter versus Borough, Condition**

Method

| | |
|---|---|
| Factor coding | (-1, 0, +1) |
| Weights | wt |

Factor Information

| Factor | Type | Levels | Values |
|---|---|---|---|
| Borough | Fixed | 5 | Bronx, Brooklyn, Manhattan, Queens, Staten Island |
| Condition | Fixed | 3 | 2Poor, 3Good, 4Excellent |

Analysis of Variance

| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Borough | 4 | 2.508 | 0.62697 | 6.89 | 0.000 |
| Condition | 2 | 1.116 | 0.55811 | 6.13 | 0.002 |
| Error | 373 | 33.963 | 0.09105 | | |
| Lack-of-Fit | 6 | 1.140 | 0.19008 | 2.13 | 0.050 |
| Pure Error | 367 | 32.822 | 0.08943 | | |
| Total | 379 | 37.641 | | | |

Model Summary

| S | R-sq | R-sq(adj) | R-sq(pred) |
|---|---|---|---|
| 0.301750 | 9.77% | 8.32% | 6.32% |

Coefficients

| Term | Coef | SE Coef | T-Value | P-Value | VIF |
|---|---|---|---|---|---|
| Constant | 0.9149 | 0.0296 | 30.90 | 0.000 | |
| Borough | | | | | |
| Bronx | -0.0497 | 0.0338 | -1.47 | 0.142 | 1.16 |
| Brooklyn | 0.0798 | 0.0503 | 1.59 | 0.113 | 1.08 |
| Manhattan | -0.1291 | 0.0295 | -4.37 | 0.000 | 1.13 |
| Queens | 0.0510 | 0.0357 | 1.43 | 0.154 | 1.12 |
| Condition | | | | | |
| 2Poor | 0.1379 | 0.0416 | 3.32 | 0.001 | 2.19 |
| 3Good | -0.0397 | 0.0253 | -1.57 | 0.117 | 2.14 |

Regression Equation

```
LogDiameter = 0.9149 - 0.0497 Borough_Bronx + 0.0798 Borough_Brooklyn
              - 0.1291 Borough_Manhattan + 0.0510 Borough_Queens + 0.0480 Borough_Staten
              Island + 0.1379 Condition_2Poor - 0.0397 Condition_3Good
              - 0.0982 Condition_4Excellent
```

The lack-of-fit score has an f-value of 2.13 with a corresponding p-value of 0.05. This means that it is very likely the model is not capturing all the factors affecting diameter. Powerful factors that could better explain tree diameter could be the percent sun the black locust gets over the course of a day and the tree's age.

Because the model uses weighted least squares and a logged response, the R-squared score of 9.77% and the S score of 0.301750 both have no physically interpretation. Every standard error must be weighted according to the different groups. This means that the standard error for trees in Manhattan is 0.3017/sqrt(1.38601)= **.256**, 0.3017/sqrt(.82951)= **.331** for trees in the Bronx, 0.3017/sqrt(.92682)= **.313** for trees in Brooklyn, 0.3017/sqrt(1.03245)= **.297** for trees in Queens and 0.3017/sqrt(.55121)= **.406** for trees in Staten Island.

For Staten Island, this is a 95% prediction interval of (0.154, 6.486), meaning the tree's actual diameter will be **between 15.4% and 649%** of the predicted diameter 95% of the time.

For Manhattan, this is a 95% prediction interval of (0.308, 3.251), meaning the tree's actual diameter will be **between 30.8% and 325%** of the predicted diameter 95% of the time.

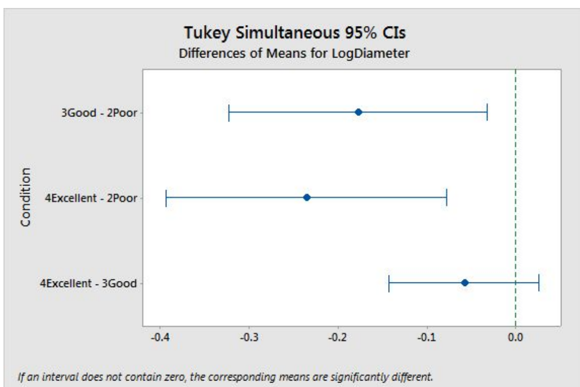
For the Bronx, this is a 95% prediction interval of (0.218, 4.592), meaning the tree's actual diameter will be **between 21.8% and 459%** of the predicted diameter 95% of the time.

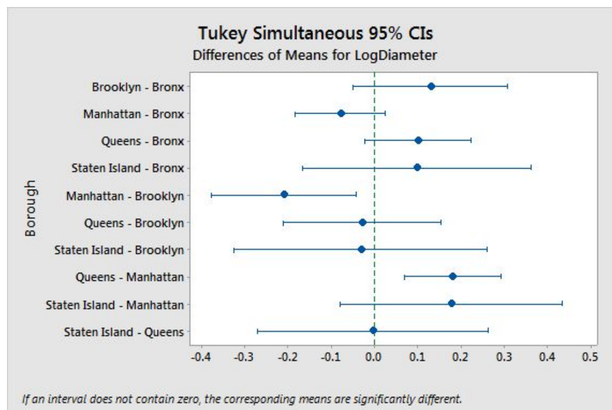
For Brooklyn, this is a 95% prediction interval of (0.237, 4.227), meaning the tree's actual diameter will be **between 23.7% and 423%** of the predicted diameter 95% of the time.

For Queens, this is a 95% prediction interval of (0.255, 3.926), meaning the tree's actual diameter will be **between 25.5% and 393%** of the predicted diameter 95% of the time.

These prediction intervals are extremely wide but that is because the model does not capture most of the factors underlying tree diameters.

## Comparison of Means: WLS Logged Diameter ~ Condition + Borough




### Tukey Pairwise Comparisons: Response = LogDiameter, Term = Condition

Grouping Information Using the Tukey Method and 95% Confidence

| Condition | N | Mean | Grouping | |
|---|---|---|---|---|
| 2Poor | 27 | 1.05278 | A | |
| 3Good | 256 | 0.87514 | | B |
| 4Excellent | 97 | 0.81672 | | B |

Means that do not share a letter are significantly different.

Tukey Simultaneous Tests for Differences of Means

| Difference of Condition Levels | Difference of Means | SE of Difference | Simultaneous 95% CI | T-Value | Adjusted P-Value |
|---|---|---|---|---|---|
| 3Good - 2Poor | -0.1776 | 0.0621 | (-0.3231, -0.0322) | -2.86 | 0.012 |
| 4Excellent - 2Poor | -0.2361 | 0.0675 | (-0.3941, -0.0780) | -3.50 | 0.001 |
| 4Excellent - 3Good | -0.0584 | 0.0360 | (-0.1427, 0.0259) | -1.62 | 0.237 |

Individual confidence level = 98.02%

**Tukey Pairwise Comparisons: Response = LogDiameter, Term = Borough**

Grouping Information Using the Tukey Method and 95% Confidence

| Borough | N | Mean | Grouping | |
|---|---|---|---|---|
| Brooklyn | 31 | 0.994670 | A | |
| Queens | 79 | 0.965840 | A | |
| Staten Island | 20 | 0.962920 | A | B |
| Bronx | 124 | 0.865165 | A | B |
| Manhattan | 126 | 0.785801 | | B |

Means that do not share a letter are significantly different.

Tukey Simultaneous Tests for Differences of Means

| Difference of Borough Levels | Difference of Means | SE of Difference | Simultaneous 95% CI | T-Value | Adjusted P-Value |
|---|---|---|---|---|---|
| Brooklyn - Bronx | 0.1295 | 0.0654 | (-0.0491, 0.3081) | 1.98 | 0.276 |
| Manhattan - Bronx | -0.0794 | 0.0382 | (-0.1835, 0.0248) | -2.08 | 0.229 |
| Queens - Bronx | 0.1007 | 0.0451 | (-0.0224, 0.2237) | 2.23 | 0.168 |
| Staten Island - Bronx | 0.0978 | 0.0964 | (-0.1654, 0.3609) | 1.01 | 0.849 |
| Manhattan - Brooklyn | -0.2089 | 0.0614 | (-0.3764, -0.0414) | -3.40 | 0.006 |
| Queens - Brooklyn | -0.0288 | 0.0671 | (-0.2119, 0.1543) | -0.43 | 0.993 |
| Staten Island - Brooklyn | -0.032 | 0.107 | ( -0.324, 0.260) | -0.30 | 0.998 |
| Queens - Manhattan | 0.1800 | 0.0409 | ( 0.0683, 0.2917) | 4.40 | 0.000 |
| Staten Island - Manhattan | 0.1771 | 0.0939 | (-0.0791, 0.4333) | 1.89 | 0.324 |
| Staten Island - Queens | -0.0029 | 0.0975 | (-0.2690, 0.2631) | -0.03 | 1.000 |

Individual confidence level = 99.34%

The Tukey comparison of means tests still shows that trees in poor condition are strongly significantly different from trees in good and excellent condition. Trees in good and excellent condition are likely not different from each other, with a p-value of 0.237.

Previously Manhattan trees were in their own group, and clearly different from all four other boroughs. Now, however, trees in Manhattan are only significantly different from trees in Brooklyn and Queens, and are not likely to be different from trees in Staten Island or the Bronx.

## Model Verification using Gray Birch trees

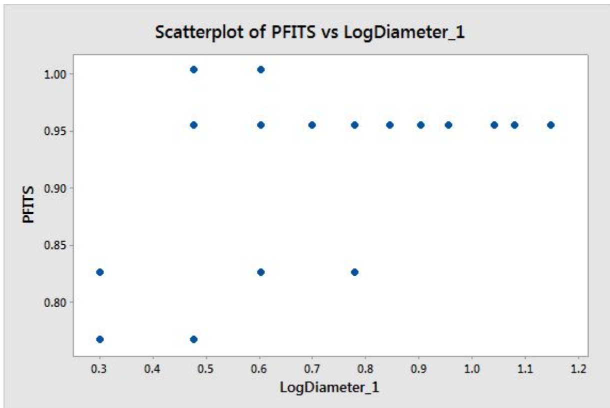
The model will be verified using a tree similar to the black locust. This tree is *Betula populifolia*, commonly known as the gray birch tree. Like the black locust, the gray birch is a shade-intolerant deciduous tree native to the eastern United States, although the gray birch exists on the coast and the black locust is a little bit inland[7]. Both are pioneer species, which means that they both are the first trees to populate a non-forested area. The gray birch "grows quickly to… a 15 inch diameter," however, which is higher than the average diameter of 8 inches for the black locust trees. Hopefully, the general effects of the borough and the tree's condition should remain the same. There are 35 gray birch trees in the 1995 dataset which I can use to verify the model. There are no gray birch trees in Staten Island or Manhattan or Queens, however.

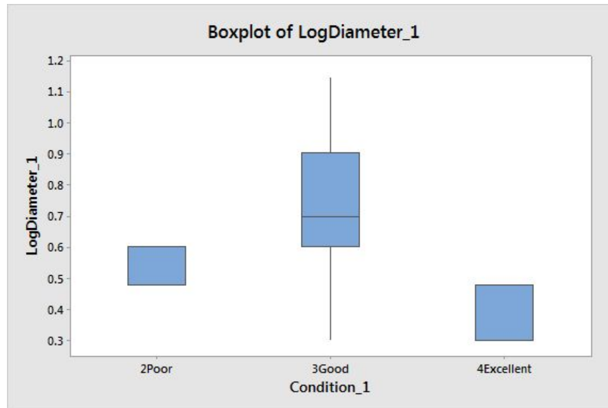
---

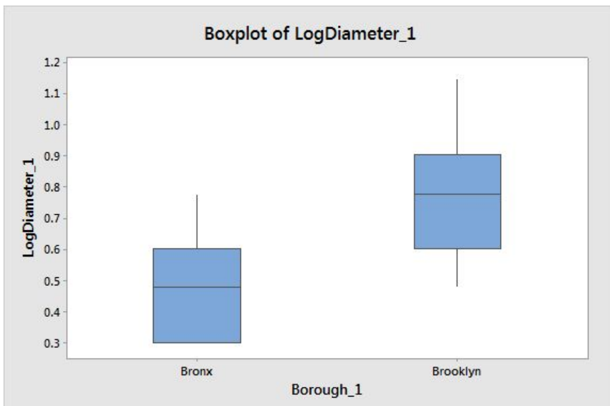
[7] https://en.wikipedia.org/wiki/Betula_populifolia

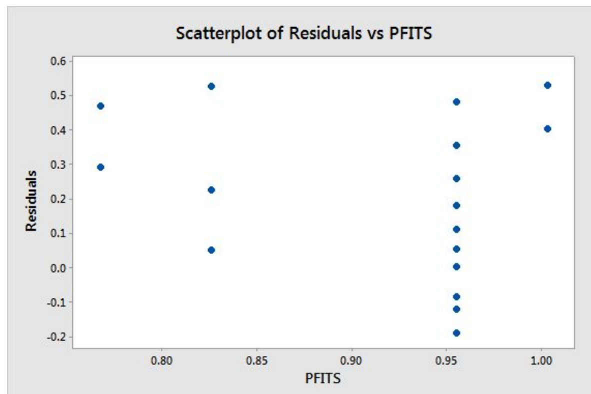
| LogDiameter | PFITS | PLIM | PLIM_1 | Residuals |
|---|---|---|---|---|
| 0.30103 | 0.82542 | 0.170963 | 1.47988 | 0.524393 |
| 0.60206 | 0.82542 | 0.170963 | 1.47988 | 0.223363 |
| 0.60206 | 0.82542 | 0.170963 | 1.47988 | 0.223363 |
| 0.77815 | 0.82542 | 0.170963 | 1.47988 | 0.047272 |
| 0.60206 | 1.00306 | 0.340671 | 1.66546 | 0.401003 |
| 0.60206 | 0.82542 | 0.170963 | 1.47988 | 0.223363 |
| 0.30103 | 0.76701 | 0.110258 | 1.42376 | 0.465978 |
| 0.47712 | 1.00306 | 0.340671 | 1.66546 | 0.525942 |
| 0.47712 | 1.00306 | 0.340671 | 1.66546 | 0.525942 |
| 0.60206 | 0.95493 | 0.327539 | 1.58232 | 0.352869 |
| 0.95424 | 0.95493 | 0.327539 | 1.58232 | 0.000686 |
| 0.60206 | 0.95493 | 0.327539 | 1.58232 | 0.352869 |
| 0.47712 | 0.95493 | 0.327539 | 1.58232 | 0.477807 |
| 0.90309 | 0.95493 | 0.327539 | 1.58232 | 0.051839 |
| 0.47712 | 0.95493 | 0.327539 | 1.58232 | 0.477807 |
| 0.69897 | 0.95493 | 0.327539 | 1.58232 | 0.255959 |
| 0.47712 | 0.95493 | 0.327539 | 1.58232 | 0.477807 |
| 0.84510 | 0.95493 | 0.327539 | 1.58232 | 0.109831 |
| 0.69897 | 0.95493 | 0.327539 | 1.58232 | 0.255959 |
| 0.77815 | 0.95493 | 0.327539 | 1.58232 | 0.176777 |
| 0.90309 | 0.95493 | 0.327539 | 1.58232 | 0.051839 |
| 0.90309 | 0.95493 | 0.327539 | 1.58232 | 0.051839 |
| 0.77815 | 0.95493 | 0.327539 | 1.58232 | 0.176777 |
| 1.04139 | 0.95493 | 0.327539 | 1.58232 | -0.086464 |
| 1.14613 | 0.95493 | 0.327539 | 1.58232 | -0.191199 |
| 1.04139 | 0.95493 | 0.327539 | 1.58232 | -0.086464 |
| 1.07918 | 0.95493 | 0.327539 | 1.58232 | -0.124253 |
| 0.60206 | 0.95493 | 0.327539 | 1.58232 | 0.352869 |
| 0.69897 | 0.95493 | 0.327539 | 1.58232 | 0.255959 |
| 0.90309 | 0.95493 | 0.327539 | 1.58232 | 0.051839 |
| 0.69897 | 0.95493 | 0.327539 | 1.58232 | 0.255959 |
| 0.77815 | 0.95493 | 0.327539 | 1.58232 | 0.176777 |
| 0.60206 | 0.95493 | 0.327539 | 1.58232 | 0.352869 |
| 0.47712 | 0.76701 | 0.110258 | 1.42376 | 0.289887 |
| 0.30103 | 0.76701 | 0.110258 | 1.42376 | 0.465978 |

Not a single point lies outside the prediction interval because it is so wide. More insight can be gained from looking at the plots.

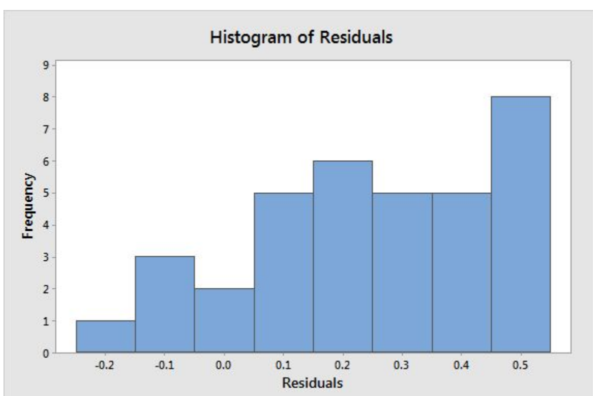


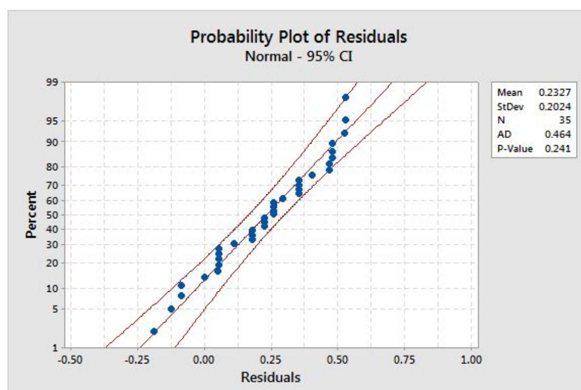




The scatterplot of predicted versus actual diameters shows an extremely weak correlation. The boxplots with both borough and tree condition are significantly different from before, however, with trees in poor condition being the group with the lowest variance.




The residuals have a right skew and roughly follow the line of the normal probability plot. The residuals versus fits shows that the only maybe accurate forecast are the points on the 0.95 line, which correspond to Brooklyn trees in good condition; they are exclusively the points with negative residuals while the other trees have consistently larger diameters than predicted. This possibility was already mentioned, since gray birch trees tend to be larger than black locusts.

Overall, the model does not do a terrible job of predicting, but there is huge variance within tree diameters that the model is not accounting for.