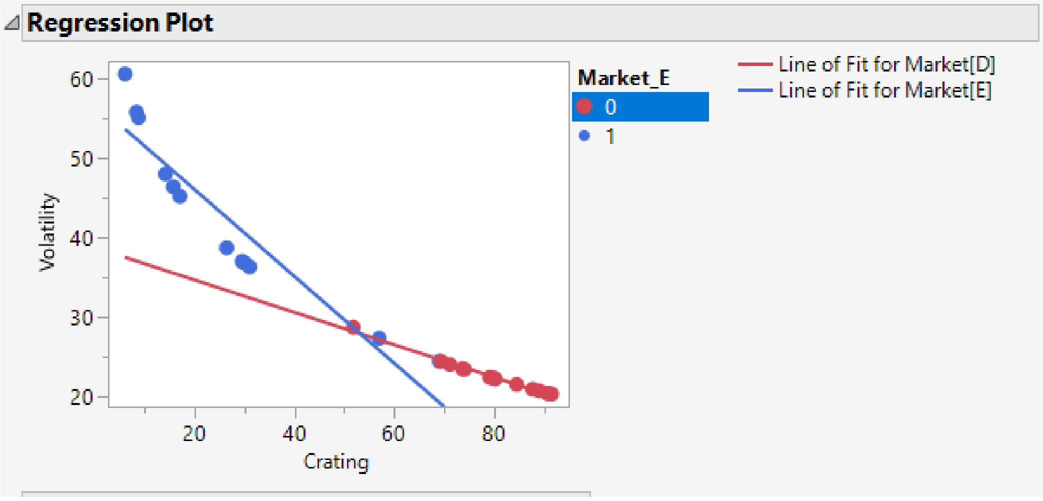
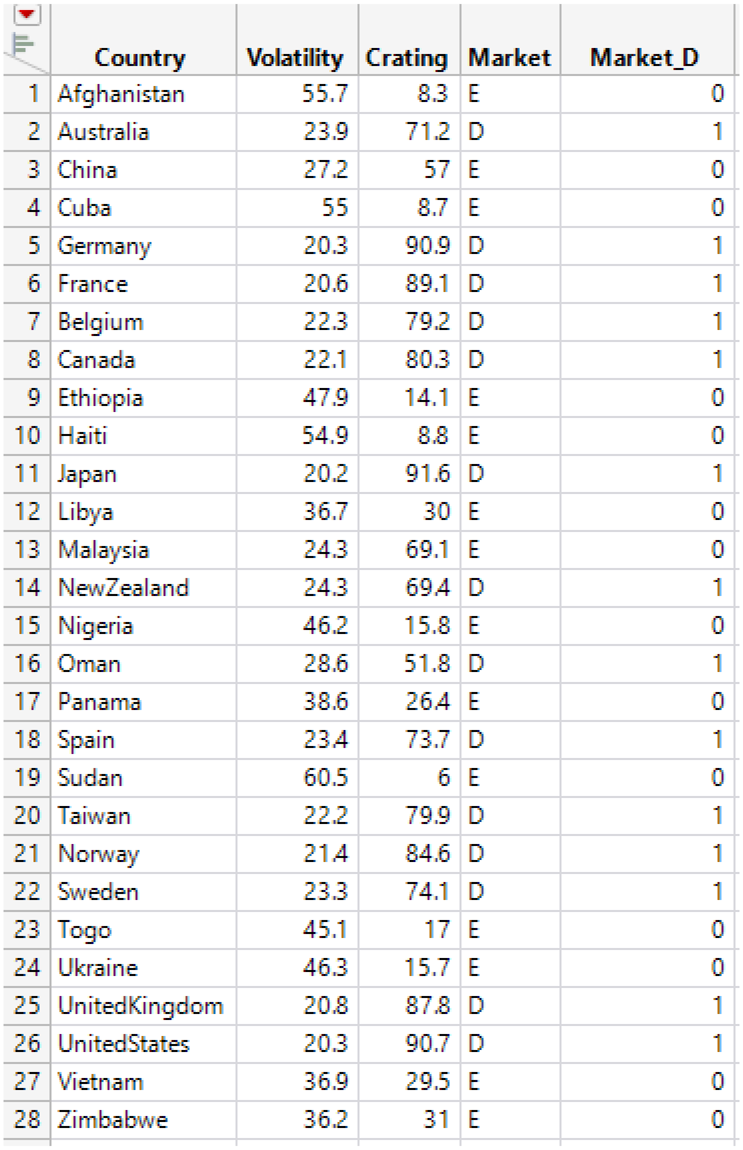
Question 1

a) Write a model that describes the relationship between volatility (y) and credit rating (x1) as two nonparallel lines, one for each type of market. Specify the dummy variable coding scheme that you use.

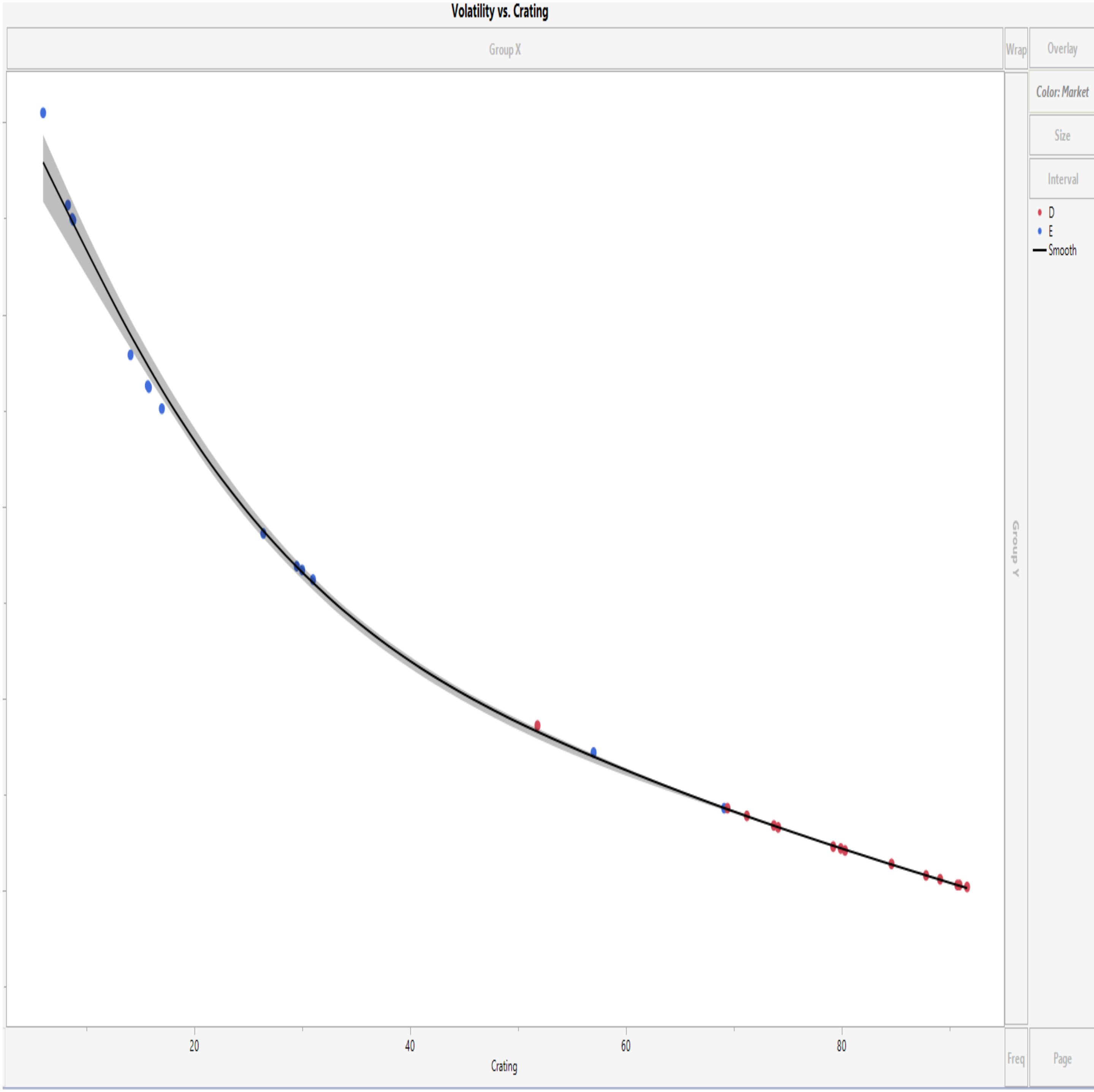


This is the plot of the model with two-non parallel lines, one for each type of the market. Using one-hot encoding (dummy transformation), the Market type variable was transformed into 2 dummy variables - Market_D and Market_E. Market_E was then dropped as a first indicator.



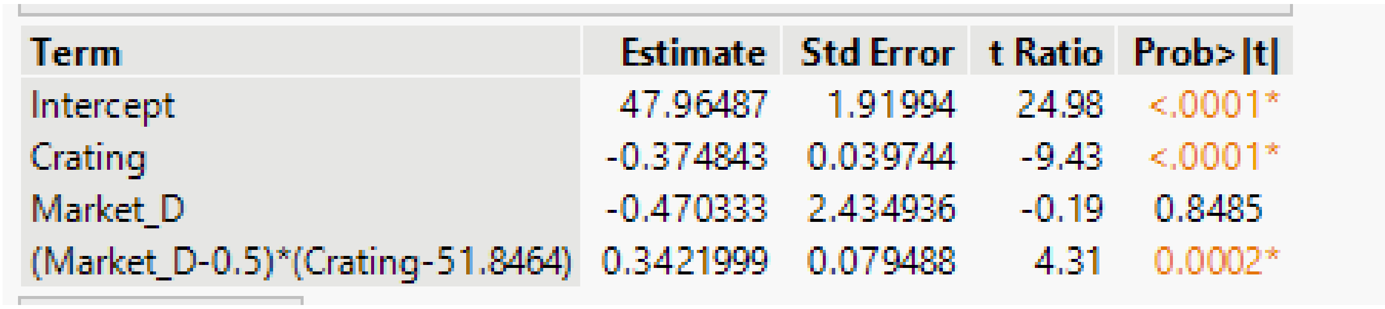
	Country	Volatility	Crating	Market	Market_D
1	Afghanistan	55.7	8.3	E	0
2	Australia	23.9	71.2	D	1
3	China	27.2	57	E	0
4	Cuba	55	8.7	E	0
5	Germany	20.3	90.9	D	1
6	France	20.6	89.1	D	1
7	Belgium	22.3	79.2	D	1
8	Canada	22.1	80.3	D	1
9	Ethiopia	47.9	14.1	E	0
10	Haiti	54.9	8.8	E	0
11	Japan	20.2	91.6	D	1
12	Libya	36.7	30	E	0
13	Malaysia	24.3	69.1	E	0
14	NewZealand	24.3	69.4	D	1
15	Nigeria	46.2	15.8	E	0
16	Oman	28.6	51.8	D	1
17	Panama	38.6	26.4	E	0
18	Spain	23.4	73.7	D	1
19	Sudan	60.5	6	E	0
20	Taiwan	22.2	79.9	D	1
21	Norway	21.4	84.6	D	1
22	Sweden	23.3	74.1	D	1
23	Togo	45.1	17	E	0
24	Ukraine	46.3	15.7	E	0
25	UnitedKingdom	20.8	87.8	D	1
26	UnitedStates	20.3	90.7	D	1
27	Vietnam	36.9	29.5	E	0
28	Zimbabwe	36.2	31	E	0

b) Plot volatility (y) against credit rating (x1) for all the developed markets in the sample. On the same graph, plot y against x1 for all emerging markets in the sample. Does it appear that the model specified in part a is appropriate? Explain.



The plot suggests that the model specified in part **a** might not be fully appropriate, as plot shows non-linear distribution. Model might be more appropriate if it is converted to a polynomial function with interaction by market type.

c. Fit the model from part a to the data. Report the least squares prediction equation for each of the two types of markets. (Market D Dummy indicator used)
These are the parameter estimates:

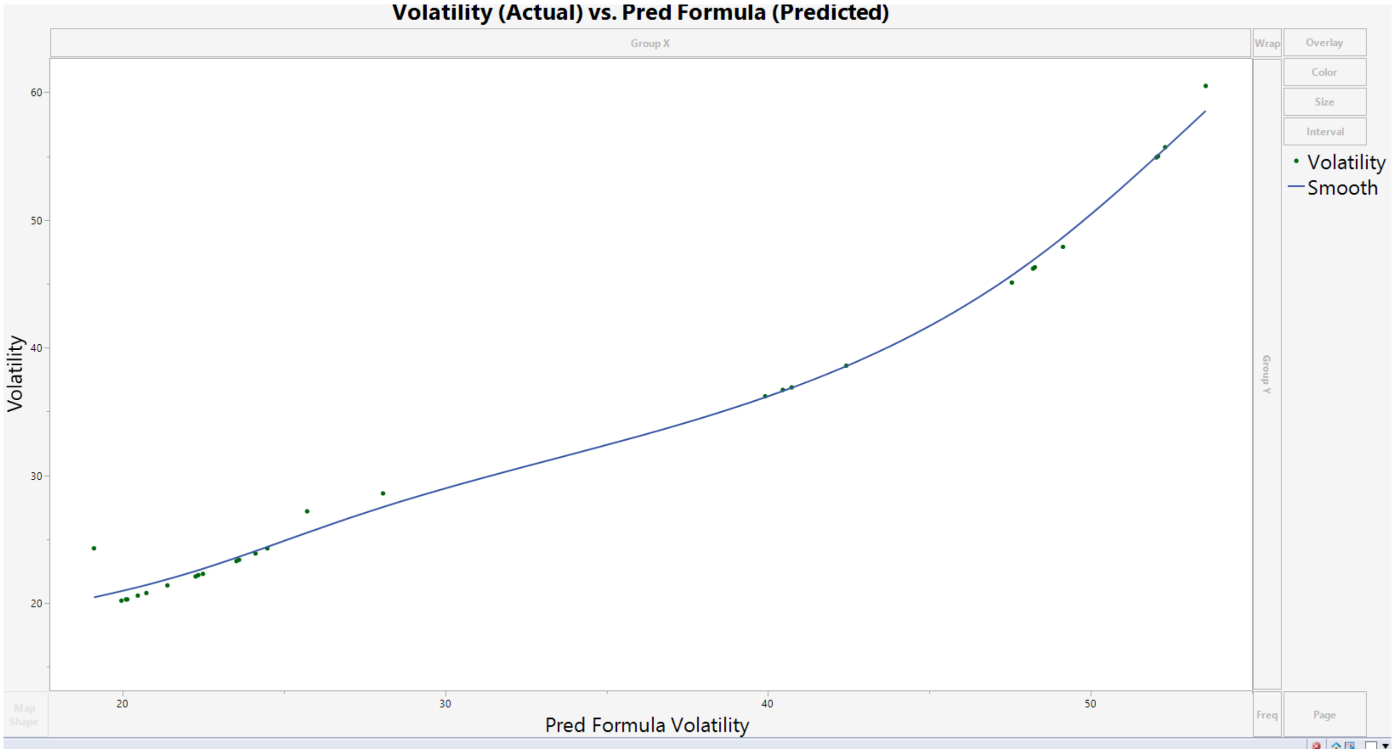


Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	47.96487	1.91994	24.98	<.0001*
Crating	-0.374843	0.039744	-9.43	<.0001*
Market_D	-0.470333	2.434936	-0.19	0.8485
(Market_D-0.5)*(Crating-51.8464)	0.3421999	0.079488	4.31	0.0002*

Equation for Developed Markets:
Y(Volatility) = 47.96487 - 0.374843 * (Crating) -0.470333(Market D (1)) + 0.3421999(0.5 * (Crating - 51.8464))

Equation for Emerging Markets:
Y(Volatility) = 47.96487 - 0.374843 * (Crating) - 0.3421999(-0.5 * (Crating - 51.8464))

d. Plot the two prediction equations of part c on a scatterplot of the data.

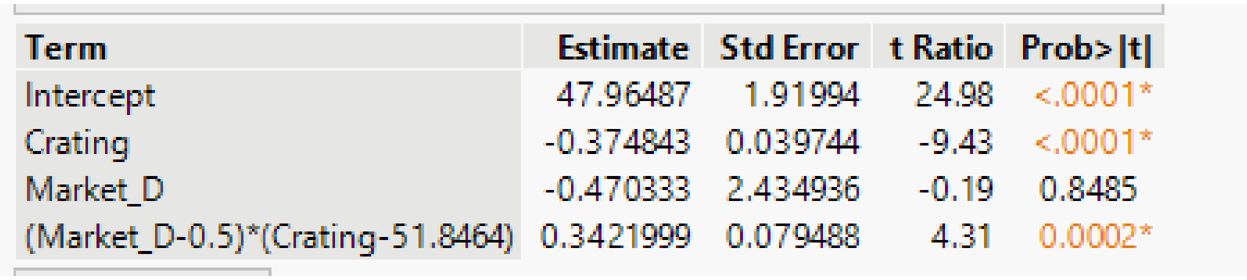


e. Is there evidence to conclude that the slope of the linear relationship between volatility y and credit rating x1 depends on market type? Test using alpha= 0.01.

Hypothesis Testing: T-Test

Null Hypothesis: The slope of the relationship between y and x1 does not depend on market type. This implies that the interaction term's coefficient is zero.

Alternative Hypothesis: The slope of the relationship between y and x1 does depend on market type. This implies that the interaction term's coefficient is not zero.

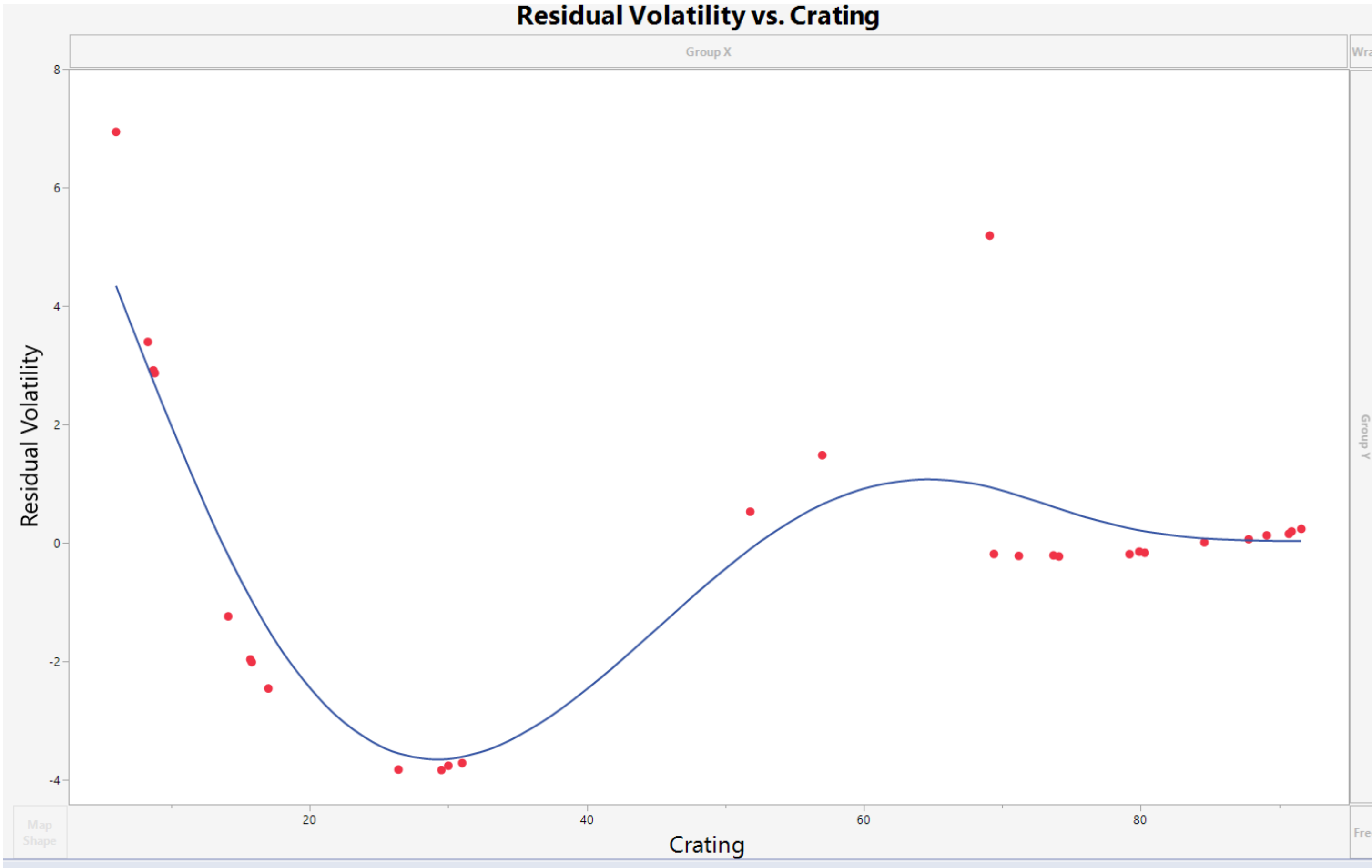


Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	47.96487	1.91994	24.98	<.0001*
Crating	-0.374843	0.039744	-9.43	<.0001*
Market_D	-0.470333	2.434936	-0.19	0.8485
(Market_D-0.5)*(Crating-51.8464)	0.3421999	0.079488	4.31	0.0002*

Probability of the Interaction term is less than 0.01 alpha level, which means that this predictor is statistically significant and its coefficient is not 0. It also means that there is statistical evidence to conclude that the slope of the linear relationship between volatility y and credit rating x1 depends on market type.

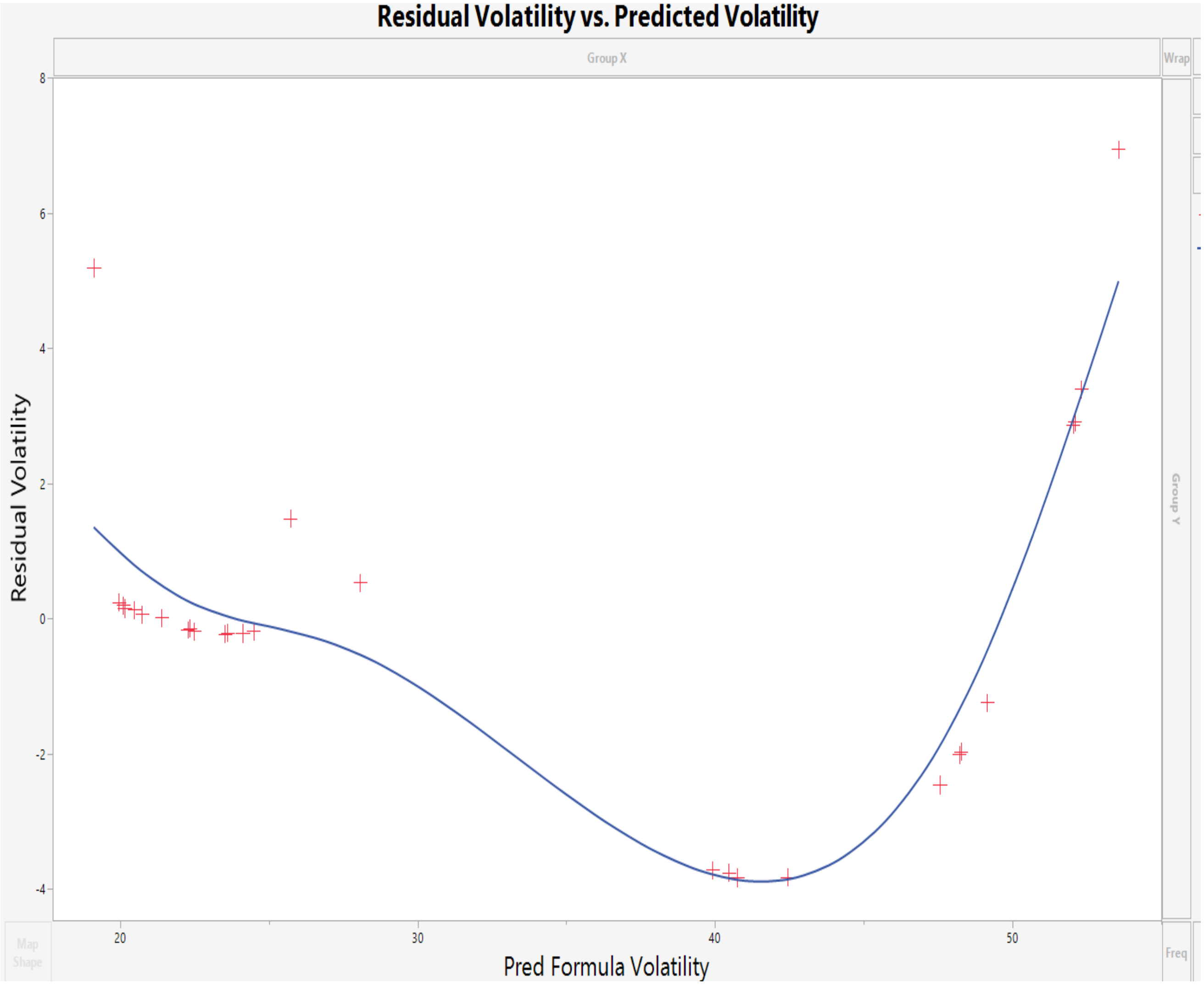
f. Conduct a residual analysis for the model to check the assumptions on the error term.

<u>Linearity</u>



Residuals were plotted against the Crating explanatory variable. If the model has a linearity, there shouldn't be any pattern in the Residuals vs Crating plot, however there is a clear **Linearity issue** as the plot shows non-linear pattern.

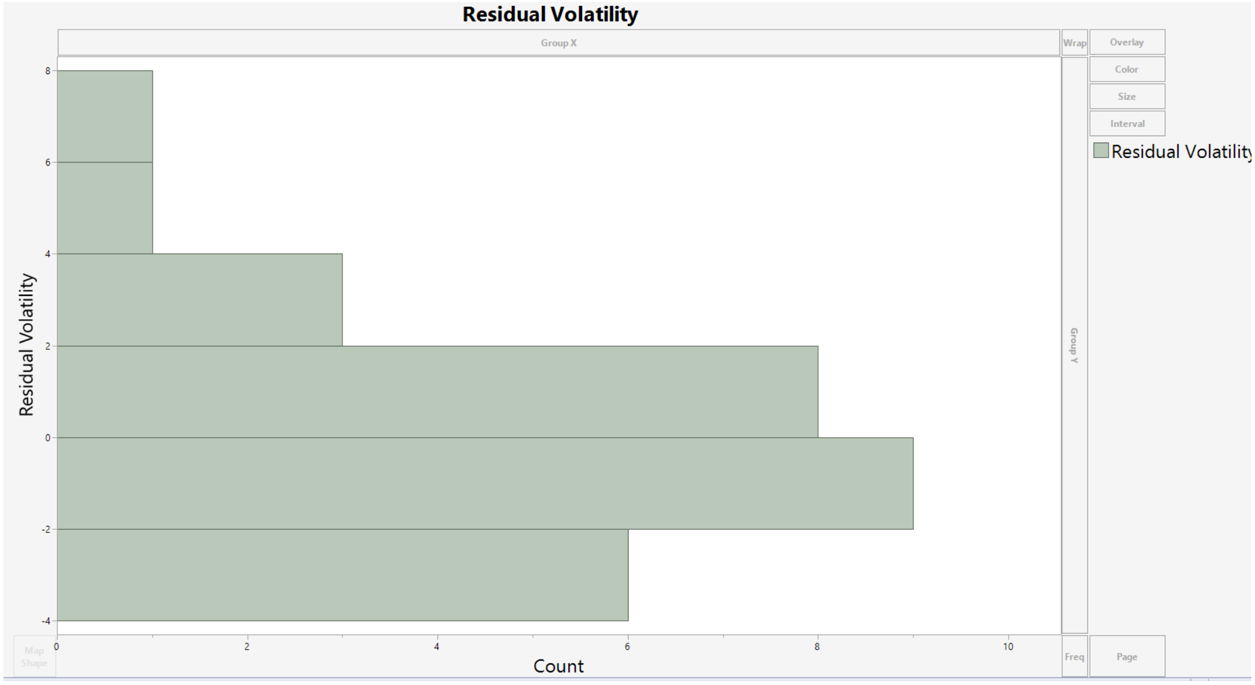
Constant Variance



By plotting the residuals against predicted values, **the Heteroscedasticity** problem was discovered. Values have residuals variance changing across the range or predicted values which violates the **constant variance** rule.

Normality

If data follows the normality assumption, histogram plot of residuals should be relatively bell shaped.

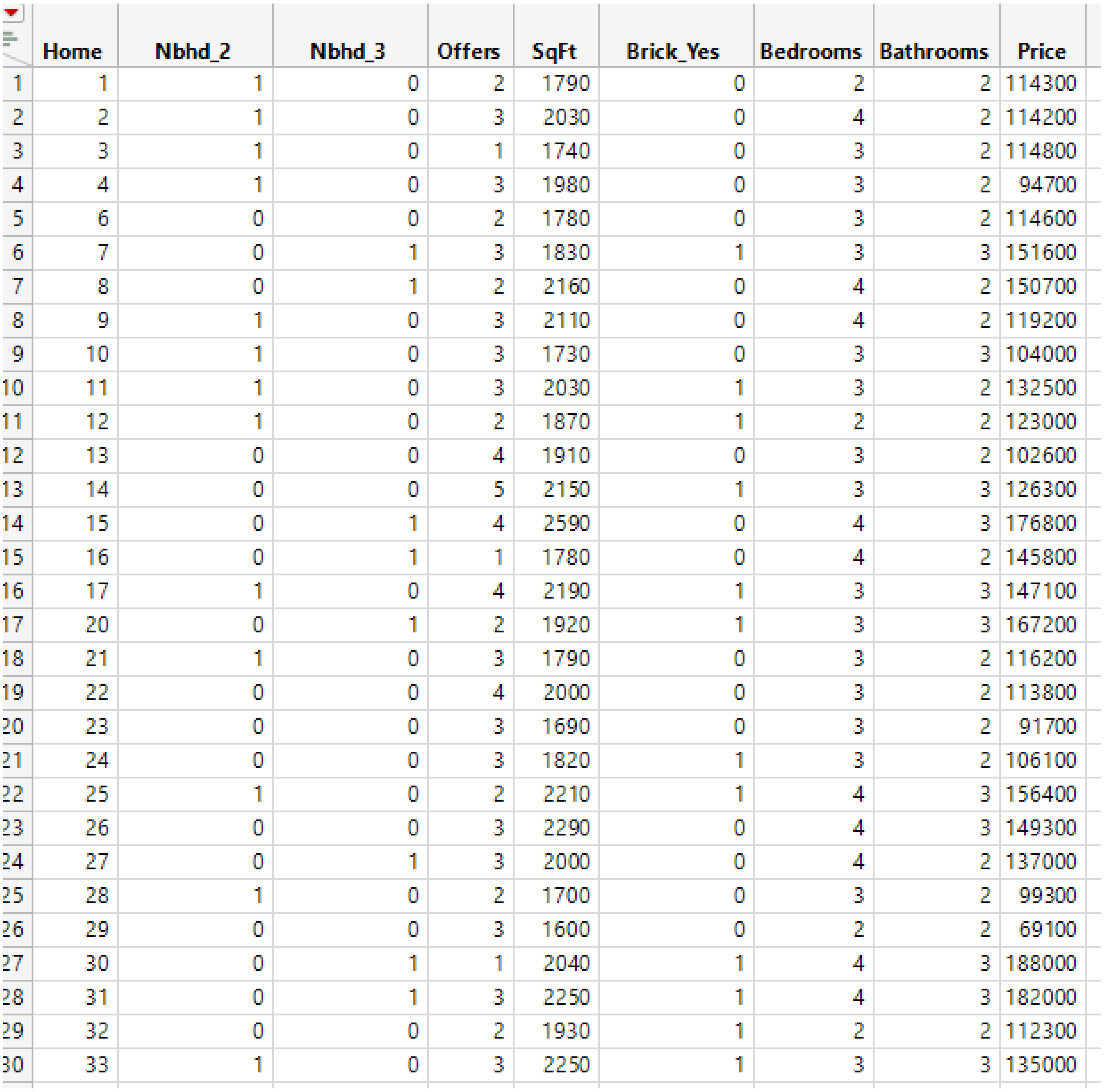


However, current residuals show that data has a left-skewed distribution and it is not normally distributed, which violates the **normality** assumption.

All 3 assumptions are violated by given data.

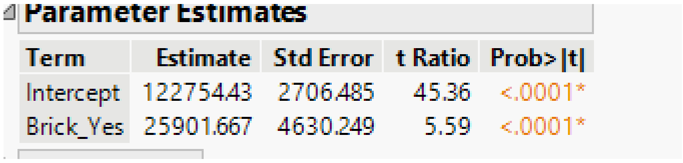
Question 2

a) **Do buyers pay a premium for a brick house, all else being equal?**
First, Neighborhood and Brick variables were transformed into dummy variables, first indicators were dropped.



	Home	Nbhd_2	Nbhd_3	Offers	SqFt	Brick_Yes	Bedrooms	Bathrooms	Price
1	1	1	0	2	1790	0	2	2	114300
2	2	1	0	3	2030	0	4	2	114200
3	3	1	0	1	1740	0	3	2	114800
4	4	1	0	3	1980	0	3	2	94700
5	6	0	0	2	1780	0	3	2	114600
6	7	0	1	3	1830	1	3	3	151600
7	8	0	1	2	2160	0	4	2	150700
8	9	1	0	3	2110	0	4	2	119200
9	10	1	0	3	1730	0	3	3	104000
10	11	1	0	3	2030	1	3	2	132500
11	12	1	0	2	1870	1	2	2	123000
12	13	0	0	4	1910	0	3	2	102600
13	14	0	0	5	2150	1	3	3	126300
14	15	0	1	4	2590	0	4	3	176800
15	16	0	1	1	1780	0	4	2	145800
16	17	1	0	4	2190	1	3	3	147100
17	20	0	1	2	1920	1	3	3	167200
18	21	1	0	3	1790	0	3	2	116200
19	22	0	0	4	2000	0	3	2	113800
20	23	0	0	3	1690	0	3	2	91700
21	24	0	0	3	1820	1	3	2	106100
22	25	1	0	2	2210	1	4	3	156400
23	26	0	0	3	2290	0	4	3	149300
24	27	0	1	3	2000	0	4	2	137000
25	28	1	0	2	1700	0	3	2	99300
26	29	0	0	3	1600	0	2	2	69100
27	30	0	1	1	2040	1	4	3	188000
28	31	0	1	3	2250	1	4	3	182000
29	32	0	0	2	1930	1	2	2	112300
30	33	1	0	3	2250	1	3	3	135000

Brick yes dummy variable was fit to answer the question.



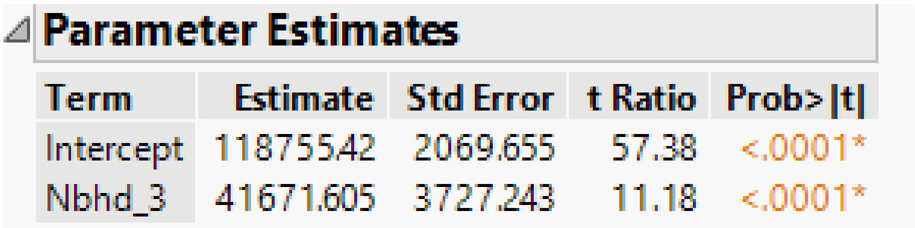
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	122754.43	2706.485	45.36	<.0001*
Brick_Yes	25901.667	4630.249	5.59	<.0001*

According to the t-test, the p-value of Brick_Yes (brick house) is less than 0.0001 at 0.0001 level of significance, which indicates that this predictor is statistically significant and has predictive value.

The estimated value of Brick_Yes is 25901.667. I**t means while anything else is constant, buyers are willing to pay a premium of ~25,902.667 dollars if it is a brick house.**

b) Is there a premium for a house in neighborhood 3, all else being equal?



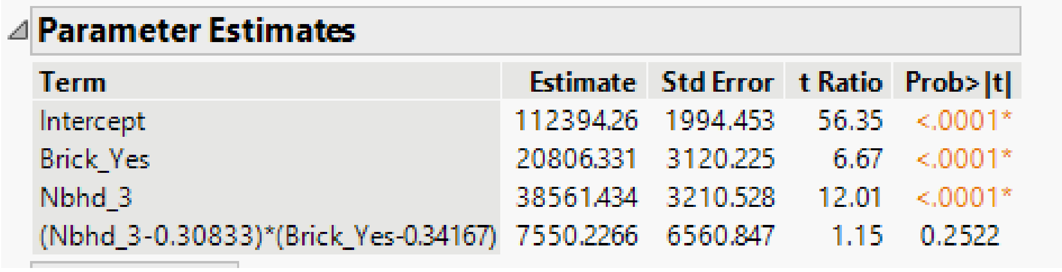
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	118755.42	2069.655	57.38	<.0001*
Nbhd_3	41671.605	3727.243	11.18	<.0001*

Neighborhood 3 dummy variable (Nbhd_3) is also statistically significant as it has a p-value less than <0.0001. Its estimate is 41,671.605. **It explains that there is a ~41,671.605 dollars premium for a house in neighborhood 3**, all else being equal.

c) Is there an extra premium for a brick house in Neighborhood 3, in addition to the usual premium for a brick house?

To test that we need to add a dummy brick_yes variable and interaction term between brick_yes and Nbhd_3 dummy variables.

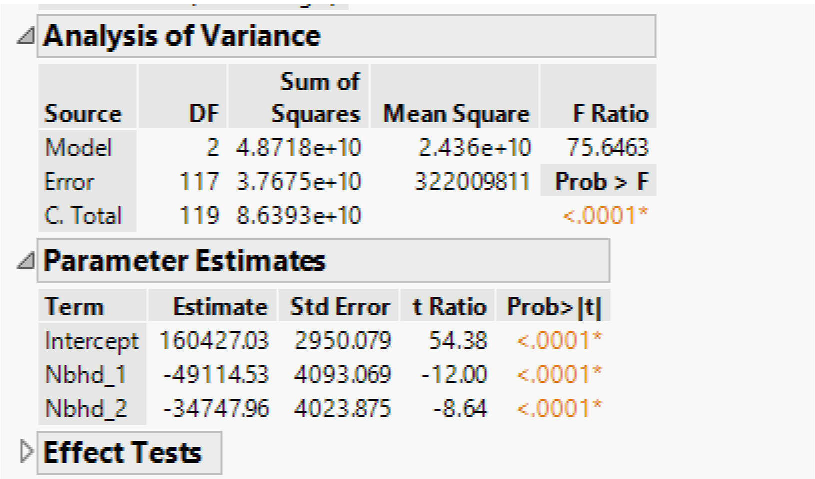


Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	112394.26	1994.453	56.35	<.0001*
Brick_Yes	20806.331	3120.225	6.67	<.0001*
Nbhd_3	38561.434	3210.528	12.01	<.0001*
(Nbhd_3-0.30833)*(Brick_Yes-0.34167)	7550.2266	6560.847	1.15	0.2522

According to t-test, the p-value of the **interaction term between Neighborhood and Brick House dummy variables** is 0.2522 and greater than 0.0001 alpha level, which indicates that this interaction is statistically insignificant and we don't have enough evidence to claim that there is an extra premium for a brick house in Neighborhood3, in addition to the usual premium for a brick house.

d) For purposes of estimation and prediction, could neighborhoods 1 and 2 be collapsed into a single "older" neighborhood?



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	4.8718e+10	2.436e+10	75.6463
Error	117	3.7675e+10	322009811	Prob > F
C. Total	119	8.6393e+10		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	160427.03	2950.079	54.38	<.0001*
Nbhd_1	-49114.53	4093.069	-12.00	<.0001*
Nbhd_2	-34747.96	4023.875	-8.64	<.0001*

Effect Tests

First, ANOVA TEST:

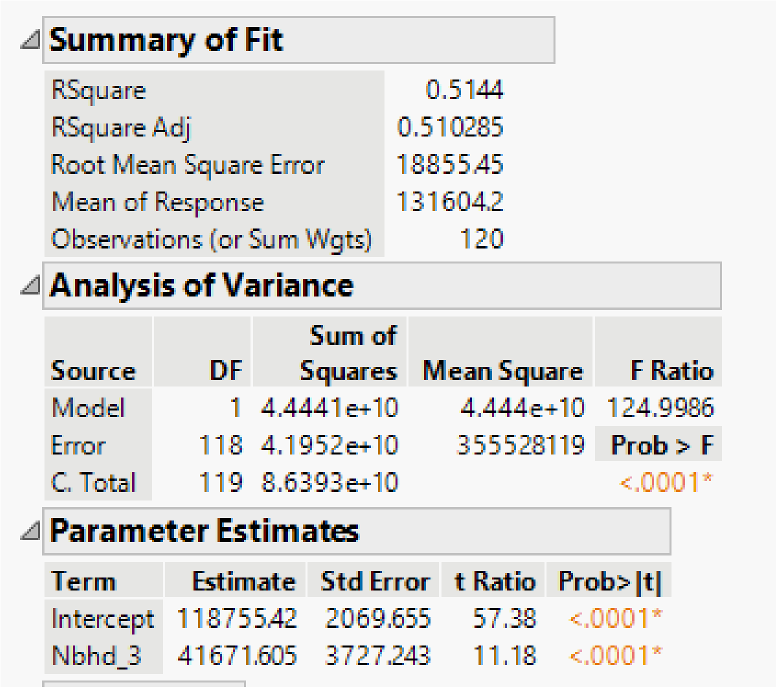
H0 (Null hypothesis): There is no difference in house price means between Neighborhood 1 and 2
HA (Alternative hypothesis): There is a significant difference in house price means between Neighborhood 1 and 2

P-value is less than 0.0001 significance level, which means we have enough evidence to reject the null hypothesis and state that there is a statistically significant difference in price means between 2 neighborhoods.

T-test:
H0 (Null hypothesis): Predictor is not statistically significant
HA (Alternative hypothesis): Predictor is statistically significant
P-value of both neighborhood variables are less than 0.0001 alpha level, which means that both neighborhood variables are statistically significant as predictors. Furthermore, estimates of two neighborhoods are different: -49114.53 (Neighborhood 1) and -34747.96 (Neighborhood 2).

Metric Results:

(Results where neighborhood_1 and neighborhood_2 were combined and dropped as first dummy indicator, neighborhood 3 was included to the prediction as a predictor)



Summary of Fit

RSquare	0.5144
RSquare Adj	0.510285
Root Mean Square Error	18855.45
Mean of Response	131604.2
Observations (or Sum Wgts)	120

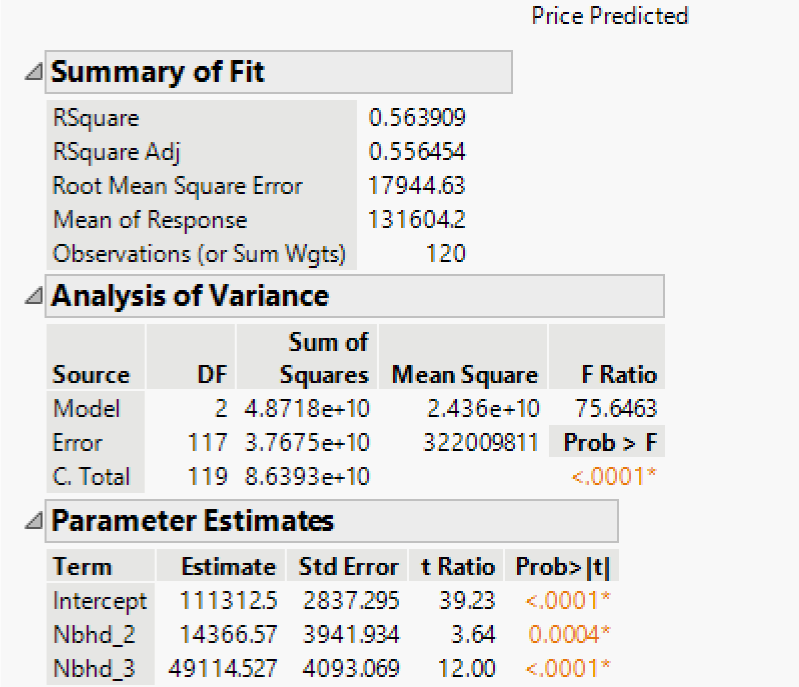
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	4.4441e+10	4.444e+10	124.9986
Error	118	4.1952e+10	355528119	Prob > F
C. Total	119	8.6393e+10		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	118755.42	2069.655	57.38	<.0001*
Nbhd_3	41671.605	3727.243	11.18	<.0001*

(Results where neighborhood_1 variable was dropped as first indicator and neighborhood_2 and neighborhood_3 were both included to the prediction as a predictors)



Price Predicted

Summary of Fit

RSquare	0.563909
RSquare Adj	0.556454
Root Mean Square Error	17944.63
Mean of Response	131604.2
Observations (or Sum Wgts)	120

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	4.8718e+10	2.436e+10	75.6463
Error	117	3.7675e+10	322009811	Prob > F
C. Total	119	8.6393e+10		<.0001*

Parameter Estimates

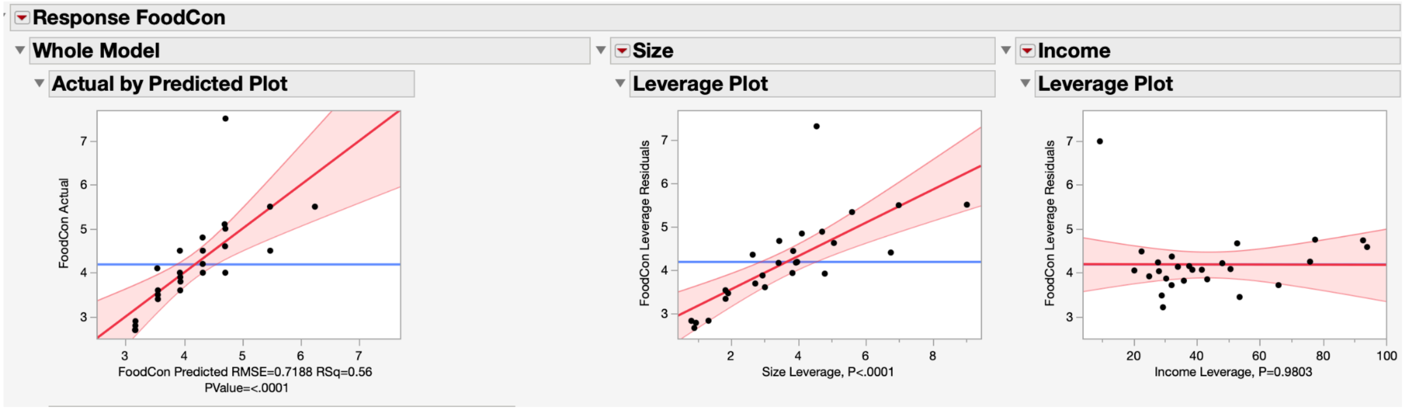
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	111312.5	2837.295	39.23	<.0001*
Nbhd_2	14366.57	3941.934	3.64	0.0004*
Nbhd_3	49114.527	4093.069	12.00	<.0001*

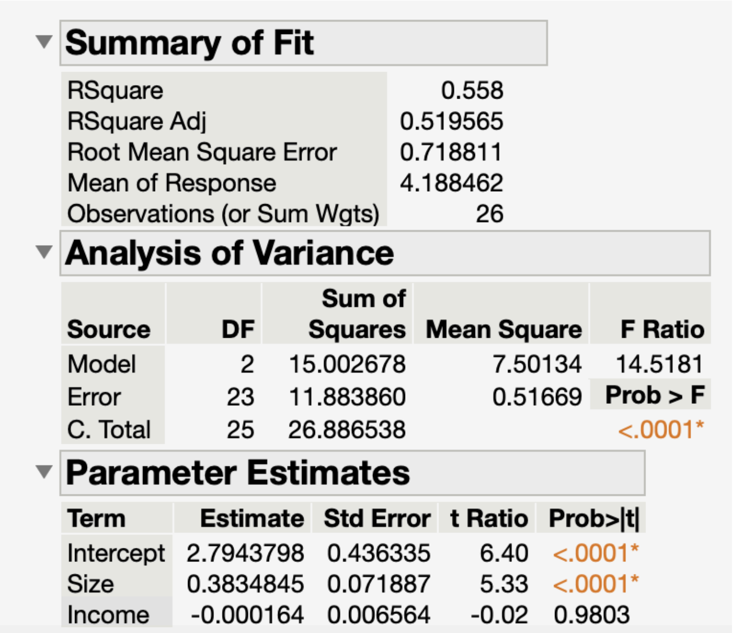
With neighborhoods not combined (2nd picture) there is a clear increase in R2 and R2 Adjusted metrics. It indicates that if we combine neighborhood_1 and neighborhood_2 as old neighborhood, we will loose around ~5% of data variance and our prediction will be less accurate.

Due to the above given results:
For purposes of estimation and better prediction of house prices, we can’t merge neighborhoods 1 and 2 into a single “older” neighborhood.

Question 3

a) **Fit the model to the data. Do you detect any signs of multicollinearity in the data? Explain.**





Summary of Fit

RSquare	0.558
RSquare Adj	0.519565
Root Mean Square Error	0.718811
Mean of Response	4.188462
Observations (or Sum Wgts)	26

Analysis of Variance

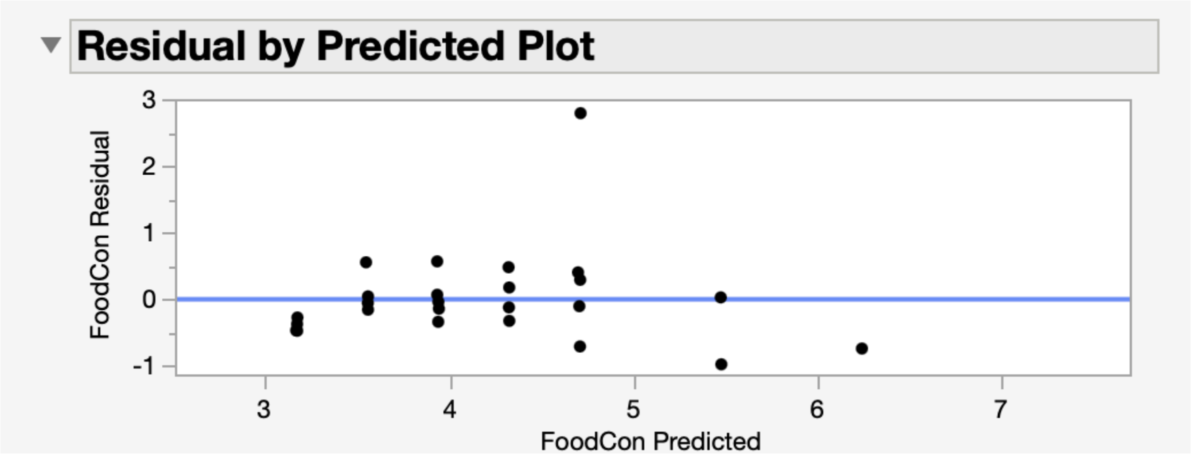
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	15.002678	7.50134	14.5181
Error	23	11.883860	0.51669	Prob > F
C. Total	25	26.886538		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	2.7943798	0.436335	6.40	<.0001*
Size	0.3834845	0.071887	5.33	<.0001*
Income	-0.000164	0.006564	-0.02	0.9803

There does appear to be evidence of multicollinearity with this model, specifically as it relates to the income variable. Even though the model overall is shown to be highly statistically significant with a p-value of less than 0.0001, the income predictor has a p-value that is very insignificant at 0.9803, which is much greater than a standard alpha value of either 0.01, 0.05, or 0.1. Additionally, the coefficient for the income variable is negative, which would imply that for every dollar increase in income, a household's expected food consumption would decrease by 0.000164. This does not make much sense theoretically, as it would be more reasonable to predict that a household would increase its food consumption, even if just slightly, if it had more available income to spend on it.

b)Is there visual evidence (from a residual plot) that a second order model may be more appropriate for predicting household food consumption? Explain.

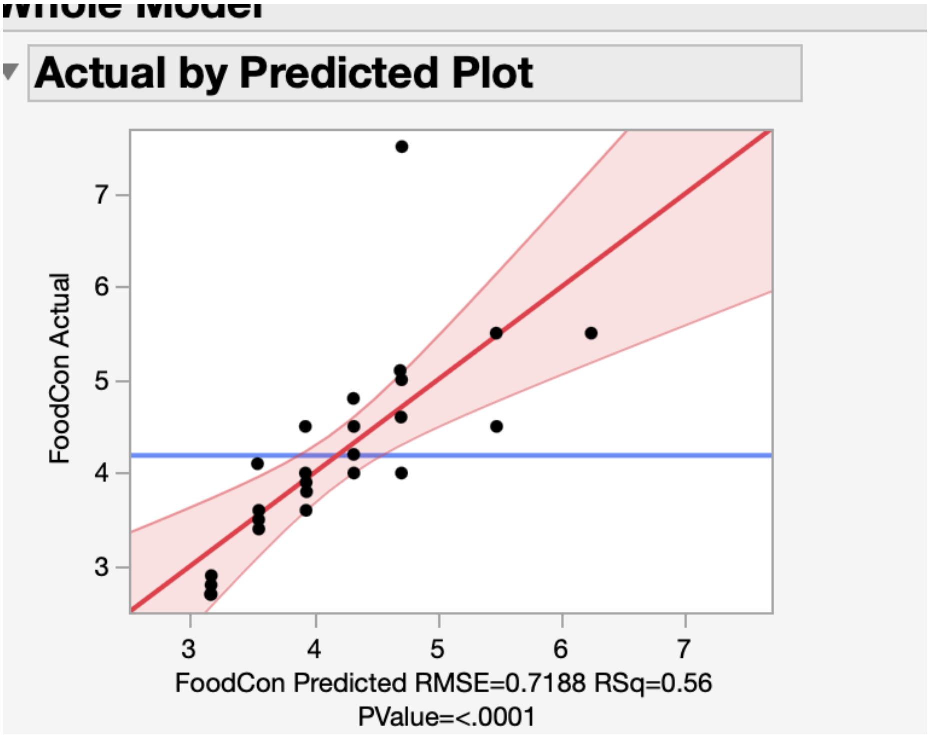


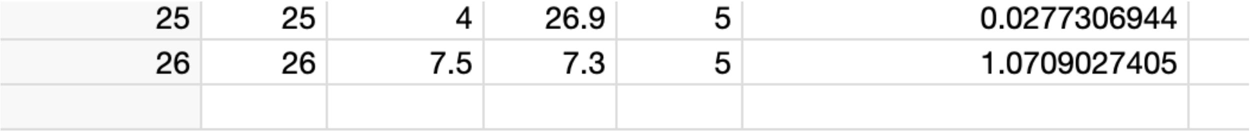
Based on the residual plot above, there is visual evidence that a second order model may be more appropriate in predicting household food consumption. There appears to be a slightly curved shape in the residual plot, which demonstrates that there might not be equal variance among the data points in the model, so a linear model is inappropriate for predicting this data.

c) **Comment on the assumption of constant error variance, using a residual plot. Does it appear to be satisfied?**

Using the same residual plot from the last part, constant error variance does not seem to be satisfied in the residual plot using size to predict food consumption. In this plot, there appears to be a slightly curved shape, in which the data points closer to 3 on the x-axis are below zero and start to rise up to be above zero before they start plummeting once again around the 4-4.5 mark on the x-axis. If the constant error variance condition were visible here, it would mean that there would be a more consistent spread of data points on the residual plot varying around similar distances from zero on the y-axis, both above and below the line for all points across the y-axis.

d) **Are there any outliers in the data? If so, identify them.**

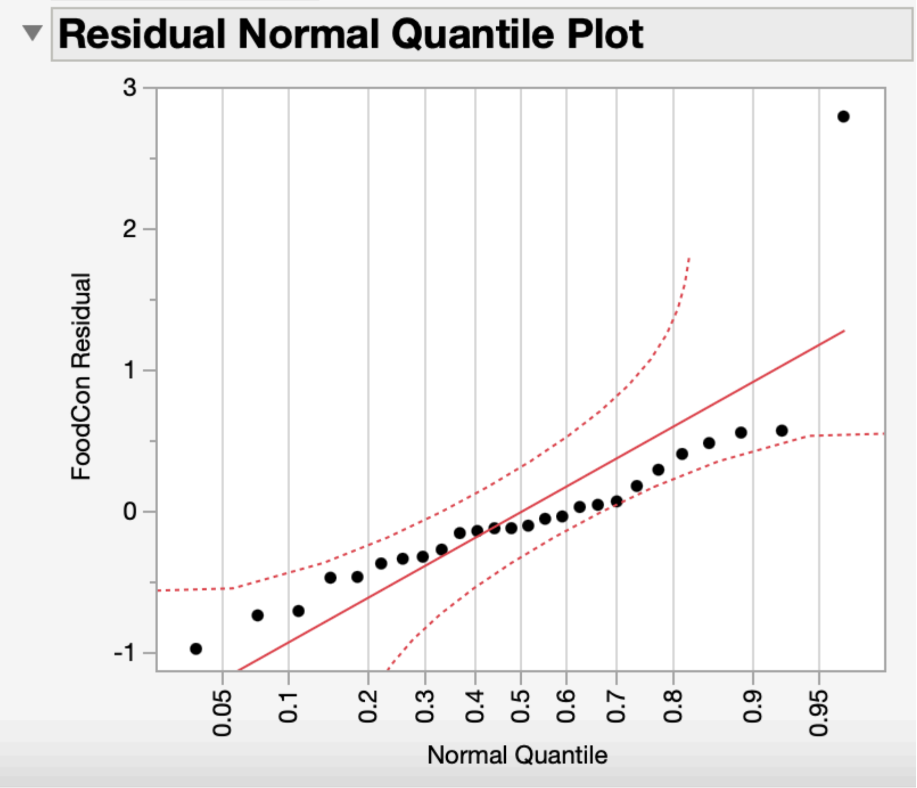




25	25	4	26.9	5	0.0277306944
26	26	7.5	7.3	5	1.0709027405

There appears to be one outlier from row 26 in the data set, in which food consumption is 7.5, income is 7.3, and size is 5. In the predicted plot with all the data points and all the predictor variables taken into account, this point can be seen with a much higher y-value than the rest, though its x-value aligns well within the range of the other points. While there are points that fall outside of the red shaded region on the plot, household 26 has a Cook’s distance of about 1.07, which signifies that it is a highly influential point. It is the only household that has a Cook’s distance of greater than 1. If this point were to be removed from the data set, it is likely that the statistical analysis would become altered.

e) Based on a graph of the residuals, does the assumption of normal errors appear to be reasonably satisfied?

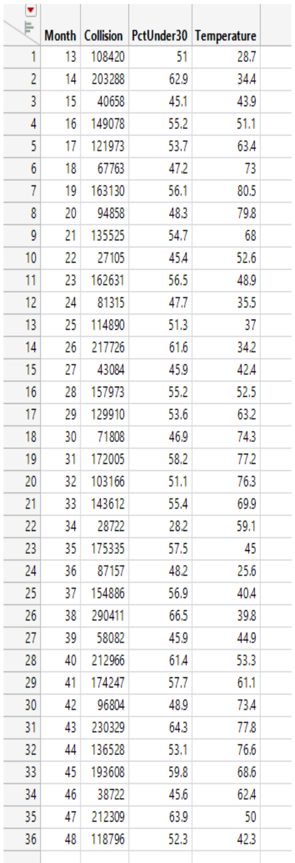


The assumption of normal errors does appear to be reasonably satisfied. Using the residual normal quantile plot for the data, it can be seen that the majority of the points (other than the row 26 outlier identified in the part above) fall within the red lines. Though some of the points are farther from the straight red line in the middle, they stay within the overall range indicating a normal distribution of variances, though the tails of the curve of these residuals might stretch outwards more than they would in an ideal normal distribution, as this tends to be where the points stray more significantly from the middle line.

Question 4

The data are in the attached file. Using only the last 3 years of data (months 13-48), answer the following questions….

First, data was filtered to months 13-48 as asked by the question.

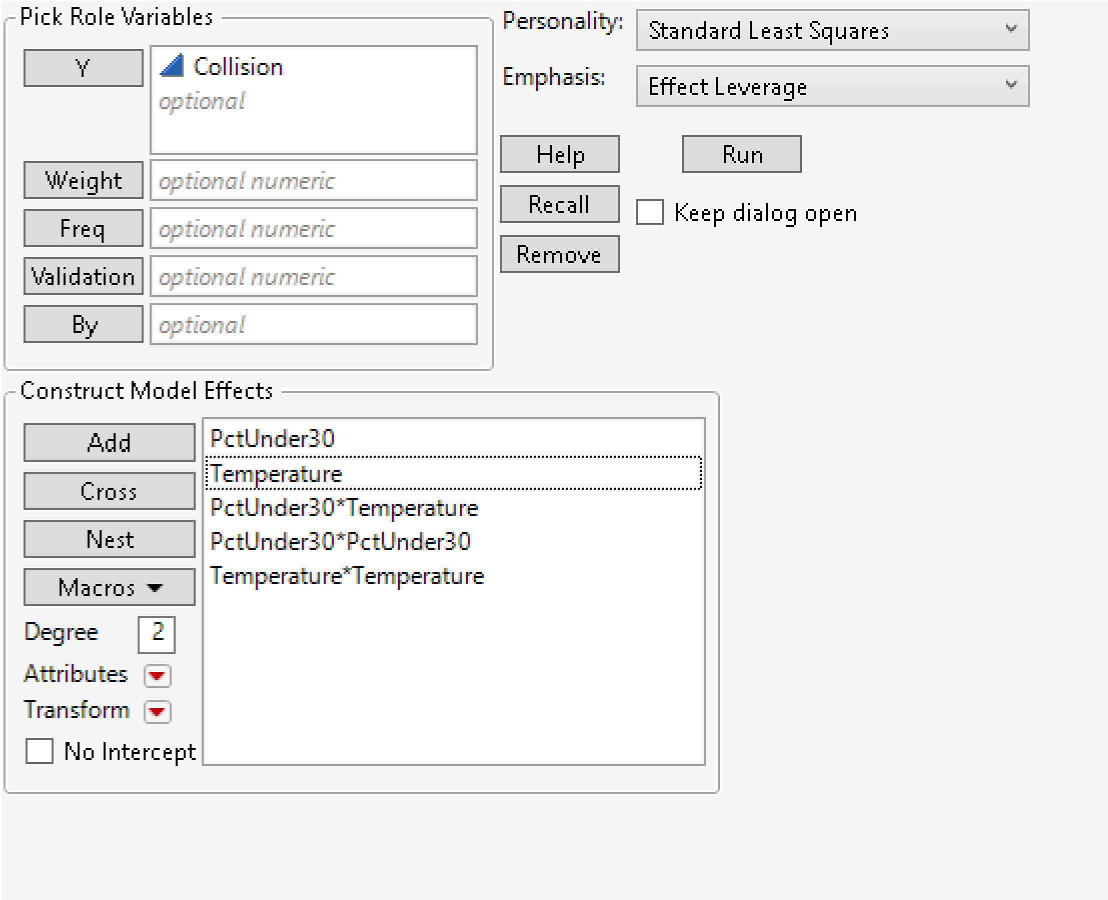


	Month	Collision	PctUnder30	Temperature
1	13	108420	51	28.7
2	14	203288	62.9	34.4
3	15	40658	45.1	43.9
4	16	149078	55.2	51.1
5	17	121973	53.7	63.4
6	18	67763	47.2	73
7	19	163130	56.1	80.5
8	20	94858	48.3	79.8
9	21	135525	54.7	68
10	22	27105	45.4	52.6
11	23	162631	56.5	48.9
12	24	81315	47.7	35.5
13	25	114890	51.3	37
14	26	217726	61.6	34.2
15	27	43084	45.9	42.4
16	28	157973	55.2	52.5
17	29	129910	53.6	63.2
18	30	71808	46.9	74.3
19	31	172005	58.2	77.2
20	32	103166	51.1	76.3
21	33	143612	55.4	69.9
22	34	28722	28.2	59.1
23	35	175335	57.5	45
24	36	87157	48.2	25.6
25	37	154886	56.9	40.4
26	38	290411	66.5	39.8
27	39	58082	45.9	44.9
28	40	212966	61.4	53.3
29	41	174247	57.7	61.1
30	42	96804	48.9	73.4
31	43	230329	64.3	77.8
32	44	136528	53.1	76.6
33	45	193608	59.8	68.6
34	46	38722	45.6	62.4
35	47	212309	63.9	50
36	48	118796	52.3	42.3

a) **Fit the complete second order model:**

$$E(y) = \beta_0+\beta_1x_1+\beta_2x_2+ \beta_3x_1x_2+\beta_4x_1^2+ \beta_5x_2^2$$

Second order model was fit using JMP and Standard Least Squares method



Pick Role Variables

Y: Collision

Personality: Standard Least Squares

Emphasis: Effect Leverage

Construct Model Effects

PctUnder30
Temperature
PctUnder30*Temperature
PctUnder30*PctUnder30
Temperature*Temperature

Degree 2

□ Keep dialog open

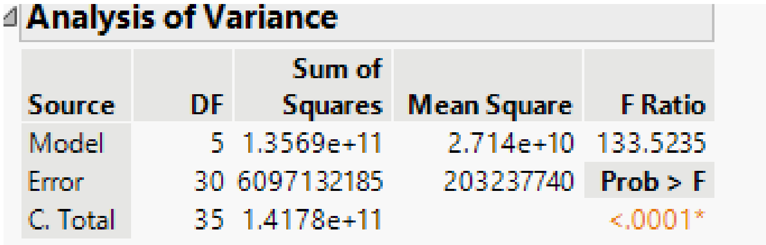
□ No Intercept

b) Test the hypothesis H0: β4=β5=0 using alpha = 0.05. Interpret the results in practical terms.

We will use the F test to check this hypothesis. Then, we will use the t-test to check coefficients individually.

Null hypothesis: B4 and B5 coefficients are 0

Alternative hypothesis: At least one of the coefficients is not 0

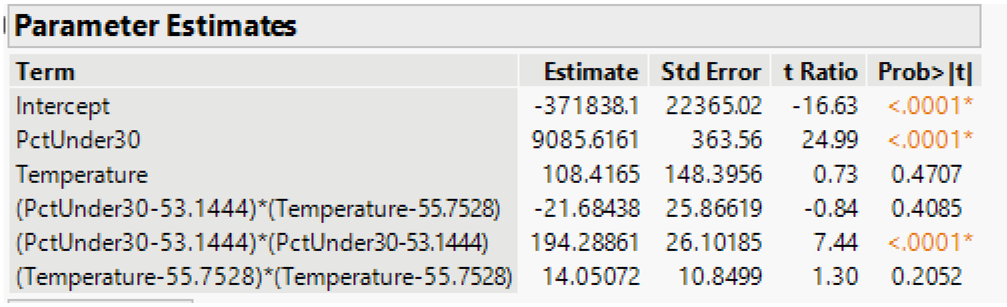


Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	1.3569e+11	2.714e+10	133.5235
Error	30	6097132185	203237740	Prob > F
C. Total	35	1.4178e+11		<.0001*

According to the ANOVA test conducted, the F p-value is less than 0.0001 level and our 0.05 level, which means that at least one of the predictor coefficients is not 0 and significant. We have enough statistical evidence to reject the null hypothesis and **claim that at least one of the predictor coefficients is not 0.**

Now, as we identified that some of the coefficients are not 0, we need to look individually for these polynomial coefficients, using t-test.



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-371838.1	22365.02	-16.63	<.0001*
PctUnder30	9085.6161	363.56	24.99	<.0001*
Temperature	108.4165	148.3956	0.73	0.4707
(PctUnder30-53.1444)*(Temperature-55.7528)	-21.68438	25.86619	-0.84	0.4085
(PctUnder30-53.1444)*(PctUnder30-53.1444)	194.28861	26.10185	7.44	<.0001*
(Temperature-55.7528)*(Temperature-55.7528)	14.05072	10.8499	1.30	0.2052

According to the t-test results, p-value of B4 coefficient (PctUnder30 * PctUnder30) is less than 0.0001 and our 0.05 level of significance, which indicates that the given variable is statistically significant.

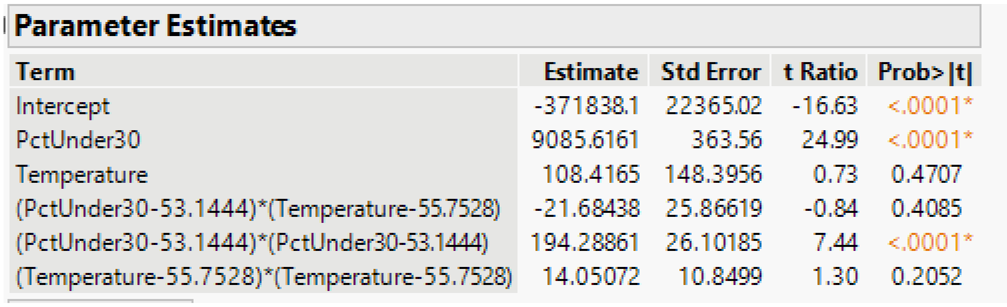
However, the p-value of B5 (Temperature* Temperature) coefficient is greater than 0.05 and 0.0001 levels of significance, which means that this predictor is not statistically significant and doesn’t have any predictive power for the data.

Practical Explanation:
The significance of the B4 coefficient for quadratic variable (PctUnder30 *PctUnder30) explains that there is actually a significant non-linear relationship between **percentage claims by drivers under age 30** and **monthly collision claims**.

The insignificance of B5 coefficient for quadratic variable (Temperature * Temperature) indicates that the relationship between **average daily temperature during month and monthly collision claims** doesn’t have a non-linear relationship. We should be cautious about calling this relationship linear as not all other complex factors possible might have been introduced.

c) Do the results support the analysts’ beliefs? Explain.

The data analysis results support only 1 from analysts' beliefs.



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-371838.1	22365.02	-16.63	<.0001*
PctUnder30	9085.6161	363.56	24.99	<.0001*
Temperature	108.4165	148.3956	0.73	0.4707
(PctUnder30-53.1444)*(Temperature-55.7528)	-21.68438	25.86619	-0.84	0.4085
(PctUnder30-53.1444)*(PctUnder30-53.1444)	194.28861	26.10185	7.44	<.0001*
(Temperature-55.7528)*(Temperature-55.7528)	14.05072	10.8499	1.30	0.2052

PctUnder30 single variable and quadratic variable PctUnder30*PctUnder 30 are both significant, they support the idea of analysts' that as the percentage of claims by driver under 30 age increases, collision claims will rise.

Both quadratic and linear variables of Temperature are not significant and their p-values from t-test are bigger than 0.05 and 0.0001 levels of significance. It means the belief of the analyst that claims will rise as the average daily temperature decreases is False, or at least we don't have enough statistical evidence to support this analyst claim.