THEFTS IN CHICAGO RESIDENTIAL AREA – AN ANALYSIS

IE6200 – Engineering Probability and Statistics

Summer 2 Semester – 2020

PROJECT REPORT

Instructor: *Prof. Nizar Zaarour*

By: **Team 5**

Yuan Ran (NUID 001569510)

Akshaya Balan (NUID 00151091597)

Sai Vaibhav Kandagattla (NUID 001581078)

Northeastern University

CONTENTS

EXECUTIVE SUMMARY

Robberies, burglaries and thefts continue to plague the length and breadth of the world. Indeed, home safety is a serious concern for the residents as well as the police and the authorities with each year witnessing increased number of cases. The crime rate in that particular area helps the State to allocate necessary resources. Several businesses such as insurance companies and safety equipment manufacturers rely on the statistical data in their business decisions. Moreover, safety is a dominant factor when people plan to move into a neighborhood.

This project focuses on analyzing if cases of fire breaking out have any influence on the cases of theft in the different areas of different zip codes in Chicago, USA. By collecting data about the rate of theft in each area and the frequency of fire breaking out, and using statistical methods to analyze this data, we can present a more intuitive data graph to depict the relationship between the fires and thefts to show how safe/dangerous an area is. This will eventually give people more inference and help them make a better housing choice.

Regression analysis by Microsoft Excel and Minitab to analyze the correlation between the two variables;

The independent variable/predictor	The dependent variable
X = Fires per 1000 housing units	*Y = Thefts per 1000 population*

The sample contains **42 elements** (sample size) which are the different zip codes in the Chicago metro area. The reference for this data is from *"Life in America's Small Cities, By G.S. Thomas"* compiled by Cengage Learning.

A closer look at the sample reveals that the area with the highest *fires per 1000 housing units (X)* also has the highest *thefts per 1000 population (Y)*. Moreover, the areas with fewer fires have fewer thefts. The analysis tries to establish a relationship between both the variables and check for its significance. This trend can be depicted as a straight line using a **scatter plot.** We observe that these two variables are positively correlated.

A **simple linear regression** model approximates or estimates the relationship between an independent and dependent variable using a straight line. The resultant linear equation we get from this dataset is **Y = 17.00 + 1.313 X.** This also indicates that a mean of 17 thefts take place when there are no incidents of fire breaking out. Furthermore, the **correlation coefficient** of **0.551** provides evidence that these two variables are positively correlated.

To prove that a relationship between the variables exists, **hypothesis testing** for the slope using the **t-test** and **F-test** were carried out for a **95% level of significance**. As with any estimation and analysis, the variability of the variables and errors associated needs to be considered. In this scenario, **30.37%** of the variability of *Y* can be explained by the variability of *X*. The variability around the reference line is given to be **19.46.**

ANALYSIS AND RESULTS

The United States Commission on Civil Rights (UNCCR) contains data on the thefts committed and the fires that broke out in the Chicago metropolitan area. We calculate the various parameters for the individual variables and conduct several analysis and tests to prove there is a significant relationship between the occurrence of fires and thefts committed.

Descriptive Statistics

From the descriptive statistics of X (*Table 1*), the average number of fires breaking out per 1000 housing units in the 42 zip codes is **12.69** with the standard deviation of **9.67**. The coefficient of variance is **76.2** and this shows extremely high variability around. There are two observations at the 24th and 25th zip code that are at an abnormal distance from the other observations in the sample. These are called outliers (*Figure 3*) and affect the average of the sample. Without the outliers, the average number of fires is **11.92**. Within the 42 areas, the minimum number of fires is just **2** while the maximum number of fires is **39.7**. Only 25% percent of areas have a low number of **5.5** fires, 50% percent of areas have a number of **17.63** fires and 75% percent of areas have a number of **17.65** fires happening per 1000 housing units.

Conducting a similar analysis on the sample of thefts (*Table 2)* we obtain the average of thefts per 1000 population in the area of 42 zip codes to be **33.67** with a standard deviation of **23.04**. The coefficient of variance is **68.4** which is also high and indicates that different zip codes have very different environments concerning the safety index. Again, we can see that there are two outliers – **75** and **147** (*Figure 4*) that affect the average and by removing the outliers the average becomes **31.03**, which can be more representative of the whole data set. The minimum number of thefts is just **4**, while the maximum number of them is **147**. Only 25% percent of areas have a lower number of **21.25** thefts, 50% percent of areas have a number of **31** thefts and 75% percent of areas have a number of **40.24** thefts happening per 1000 population. The co-variance of variables X and Y is **122.76**.

From the given data (*Figure 5*), we can conclude that the region under the 24th zip code is the most dangerous because the fire and theft seems to be the maximum there whereas the 18th zip code is the safest since the fire and theft is relatively low compared to the others which is also evident in the scatter plot. The co-variance of variables X (Fires per 1000 households) and Y (thefts per 1000 population) is **122.76** which is positive. Hence, we can conclude that the variables are linearly related to each other.

Scatter Plot and Residual Plot Analysis

Scatter Plots' (*Figure 1*) primary uses are to observe and show relationships between two numeric variables. It cannot be a coincidence that the area with a lot of fires also has the highest thefts and the area with less fires has less thefts. The graph clearly shows an observable trend which is known as the positive correlation with a correlation coefficient (R_{xy}) of **0.551**.

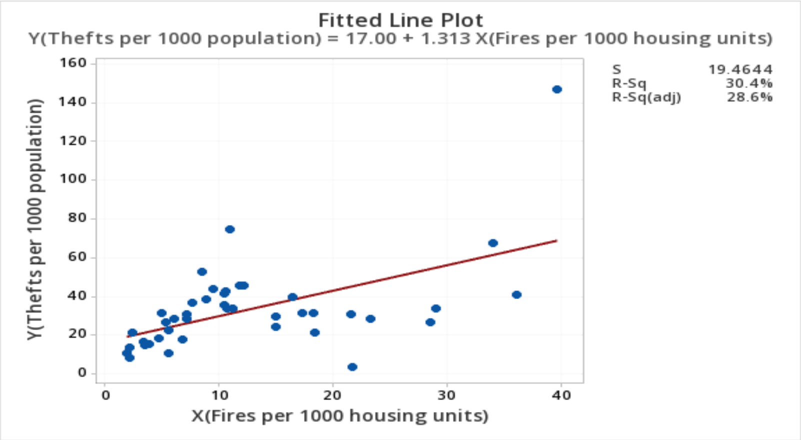


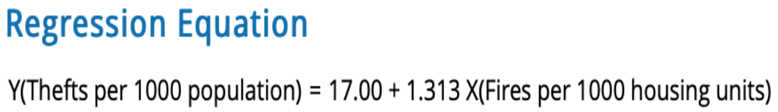
Figure A: Fitted Line Plot (Scatter Plot)

A simple linear regression model (*Table A, Table 5*) approximates or estimates the relationship between an independent (Fires = X) and dependent variables (thefts = Y) using a straight line. An easier way to understand this is to view the scatter plot and visualize a straight line that passes through most the points (*Figure A*). The resultant linear equation we get is, Y = 17.00 + 1.313 X (*Equation A*). The Residual Plot (*Figure 2*) gives us a measure of how much our regression line misses a data point. Since we can see that the points are randomly spread around the line at 0, we can think of the regression line as a good fit for the analysis and the model. To progress with the regression analysis, we assume that the histogram of residuals is normally distributed (*Figure 6*).

ANOVA and Output

The correlation coefficient (R_{xy}) measures the strength and direction of a linear relationship between two variables. A R_{xy} value of magnitude 1 depicts strongest linear correlation while 0 depicts no correlation. As mentioned before, we obtained a value of **0.551** (*Table 5)*. Thus, the strength between the fires and thefts is 0.551 and the direction is positive as seen from the scatter plot (*Figure A*). In simple terms, this means that fires breaking out has an influence on the thefts committed and might also be a reason for increased thefts. However, this value is usually an overestimate of the relationship between two variables since it is dependent on the sample size and the number of independent variables (predictors). With a small sample size of 42 and only one predictor in this case, we are bound to get a decent value for the strength and therefore further analysis is required to arrive at a conclusion.

The next step is to analyze the regression equation which forms the basis of our hypothesis testing. This equation (*Equation A)* obtained from the scatter plot (*Figure A*) links the two variables. What this means is that, for a unit change in the fires/1000 housing units there is a positive change in thefts/1000 population by 1.313 units. This follows the equation of a straight line and hence we can derive two important numerical data – the slope and the y-intercept given by b1 and b0 respectively (*Table B*, *Table 5*). The positive slope of **1.313** indicates that the variables have a positive correlation and the y-intercept of **17** shows that without any incidents of fires breaking out 17 thefts are committed per 1000 population.



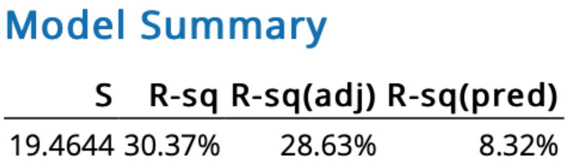
Regression Equation

Y(Thefts per 1000 population) = 17.00 + 1.313 X(Fires per 1000 housing units)

Equation A: Regression Equation

The further action in the analysis would be to calculate the Sum of Squares Total (SST), Sum of Squares Error (SSE) and Sum of Squares Regression (SSR). The values of SST, SSE and SSR are **21765.333, 15154.447** and **6610.887** respectively (*Table 4*). The sum of the squared error due to residuals is high compared to the regression. This means that the deviation between the actual number of thefts and predicted number of thefts is greater than the difference between predicted thefts and mean thefts. The SST is the sum of squared differences between the observed dependent variable and its mean. It is a measure of the total variability of the dataset. The SSE is the difference between the observed value and the predicted value. We usually want to minimize the error since the smaller the error, the better the estimation power of the regression. The rationale is the following: the total variability of the data set is equal to the variability explained by the regression line plus the unexplained variability, known as error. Given a constant total variability, a lower error will cause a better regression. Conversely, a higher error will cause a less powerful regression. The SSR is a measure that describes how well the line fits the data. If this value of SSR is equal to the sum of squares total, it means the regression model captures all the observed variability and is perfect.

It’s time to introduce the R-squared value. It is equal to variability explained by the regression, divided by total variability. It is a relative measure and takes values ranging from 0 to 1. An R-squared of zero means our regression line explains none of the variability of the data and an R-squared of 1 would mean our model explains the entire variability of the data. Unfortunately, regressions explaining the entire variability are rare. The analysis gives an R-squared value of **0.3037 (***Table A*). This indicates that **30.37%** of the variability of the thefts can be explained by the variability of fires breaking out. Observed values usually range from 0.2 to 0.9. Thus, the R-squared is the measure of the goodness of fit. Unfortunately, there is no definitive answer as to how good the value of the R-squared is. It depends on the variables and the various factors. The value obtained is good enough to go ahead with the analysis however the inclusion of more predictors to factor for the variability might lead to a better R-squared value. The adjusted R-square value comes into play when we use multiple regression and hence can be deemed redundant in this case.



Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
19.4644	30.37%	28.63%	8.32%

Table A: Regression Model Summary using MiniTab

The Mean Squared Error (MSE) indicates how close a regression line is to a set of points. It does this by taking the distances (errors) from the points to the regression line and squaring them. It also gives more weight to larger differences. The value of MSE obtained is **378.86** (*Table 4,5*). The square root of the mean square error gives us the standard error of estimate (S_e) which is **19.46** (*Table 4*,5) in this case. This is the standard variability around the reference line. The estimated standard deviation (Sb_1) is **0.314** (*Table 4,5*) for the slope b1.

Hypothesis Testing

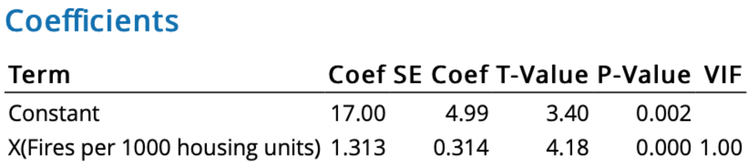
Previously, we had made an assumption that a relationship between the two variables (fires and theft) exists. Upon initial analysis, we did find that there is a positive correlation between the variables. The aim of hypothesis testing is to find out if there is actually a significant relationship between the two variables. We test for significance in the regression line by using the slope since conceptually the slope of the line is the basis of the entire model. Hence by testing the significance of the slope, we can test the significance of the entire model.

The testing is carried out by formulating the null and alternate hypothesis. Here, the null hypothesis states that there is no relationship between the variables i.e. the slope is zero and the alternate hypothesis states that there is a significant relationship i.e. slope is not equal to zero. This is given by;

$H_0 : \beta 1 = 0$

$H_a : \beta 1 \neq 0$

To calculate the significance, we compute the test statistic and plug it into the distribution. If the test statistic falls in the rejection region, it means that we have sufficient evidence to reject the null hypothesis. The level of significance level (alpha) is taken to be **0.05**.



Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	17.00	4.99	3.40	0.002	
X(Fires per 1000 housing units)	1.313	0.314	4.18	0.000	1.00

Table B: Coefficients of line equation and t-value

When α = 0.05, the t-score value = **2.021** (*Table 7)*. From the graph (*Figure 7*), we can see that the test statistic t = **4.18** (corresponding p value = **0.000**) is very much into the rejection region since **0.000 < 0.05** (Figure 6). Thus, we have sufficient evidence to reject the null hypothesis and can conclude that there is a strong relationship between the two variables i.e. fires and thefts.

We can also perform another test to determine this relationship. The F-test uses a chi-squared function and is a one-sided test as opposed to the t-test which uses a normal distribution and is a two-tailed test. However, the procedures including the hypotheses remain the same in both the cases. Since our model has only one independent variable (predictor) the result from both F-test and t-test will be the same. The p value under the F statistic of 'x' will also be similar to the p value of the test statistic of the t-test. Ideally the F-test is better than the t-test since the interest of ANOVA also includes the locations of the distributions represented by means. The F-test gives the overall fit of our linear regression model.

The test gives us the values Fα = **4.08** and the calculated $F_{statistic}$ = **17.45** (*Table 6*). Here, $F_{statistic}$ > Fα. Hence, we have sufficient evidence to reject the null hypothesis and we can conclude that there is a strong relationship between both the fires and thefts.

Confidence and Prediction Intervals

A confidence interval gives a range of plausible values for the parameter of interest. It gives a percentage confidence level where the mean value of the theft variable will lie within a range. In this dataset, a 95% confidence interval is built wherein the mean number of thefts will fall.

A prediction interval is a type of confidence interval that predicts the value for a new observation from the given regression model. The prediction interval is usually wider compared to the confidence interval since it has a bigger margin of error.

We build a confidence and prediction interval for a given mean of fires which is **12.69**. The confidence interval is given as **27.6 ≤ x_0 ≤ 39.7** and the prediction interval is **-6.1 ≤ x_0 ≤ 73.5.** The number of fires cannot be a negative value; hence we assume the lower limit to be zero (*Table 8*).

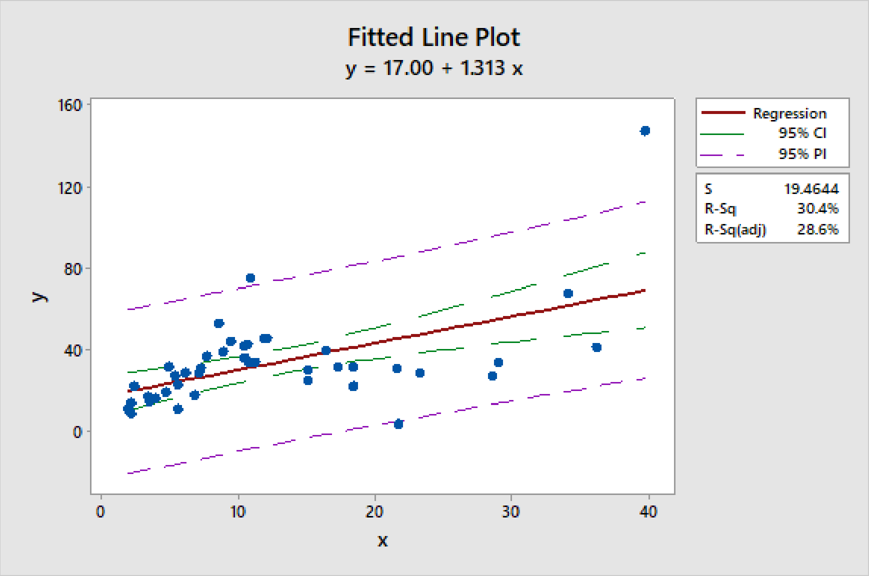


Figure B: Fitted Line Plot (Confidence & Prediction Interval)

Assuming that there are 20 fires per 1000 housing units, we can be 95% sure that the mean number of thefts fall between (**35.6217, 50.9069**) and number of thefts fall between (**3.18983, 83.3387**). In this case, the prediction interval is extremely wide and therefore we need to consider other factors to get a more accurate range (*Table 9*).

CONCLUSION

The linear relationship between Fires and Thefts in Chicago indicates that more fires in one place could lead to more thefts. The reason behind this might be that after the fire breaks out, criminals can take advantage of the distraction the fire has caused and commit more thefts. It might also indicate there is a lack of investment in police or social security leading to poor law enforcement in the area which causes increased frequency of fires and thefts.

Using the data from the analysis we can differentiate between the ‘safe’ and ‘unsafe’ areas. For people who choose to live in Chicago, we recommend areas with zip code 18, 14, 12 and avoid areas with zip code 24, 23, 6 (*Figure 5*).

Most insurance companies cover for the damage caused when burglary instigates the fire accident. However, they do not cover the theft that occurs after the fire accident as they believe that the theft or burglary is not linked to the fire accident. There are several articles that state that there has been an increase in the number of thefts after a fire accident. Especially in the recent times when protests broke out all across the United States, this positive relationship between fires and thefts came to light. This is further demonstrated by the hypothesis testing which clearly shows fires instigate thefts.

Even though there is a relationship between fires and thefts, the R-squared value of 30.37% is rather on the lower side. This might be due to the fact that only one predictor was used to carry out the analysis. We can consider other factors such as the income level of the residents, gun ownership rates etc. into the built model which might give us a better R-squared value. This opens up the opportunity to venture into other types of regression such as multiple regression. Furthermore, using a polynomial regression analysis gives us an R-squared value of **35.9%** which is a slight improvement to the R-squared value using simple linear regression.

Another area of improvement would be the sampling. The analysis was carried out using 42 elements. Although this sample size is good enough for us to perform an analysis using simple linear regression, it still is small compared to the whole population. Since the error decreases as our sample size increases, we will end up with more accurate results. Moreover, this analysis is limited to the Chicago area. The trends might differ as we expand to other parts of the country and further analysis needs to be performed. Moreover, we need to make sure the data is relevant and does not become void with the passage of time. There is always a possibility that the data collected a few years ago might not apply to the current scenario. In this case, there is always a probability that since the data was collected, the law could have become stricter and well enforced or the infrastructure funding could have increased.

This project is certainly helpful for people who might want to move to Chicago and forms an outset for analysts and companies venturing into safety and crime-based analysis.

APPENDIX

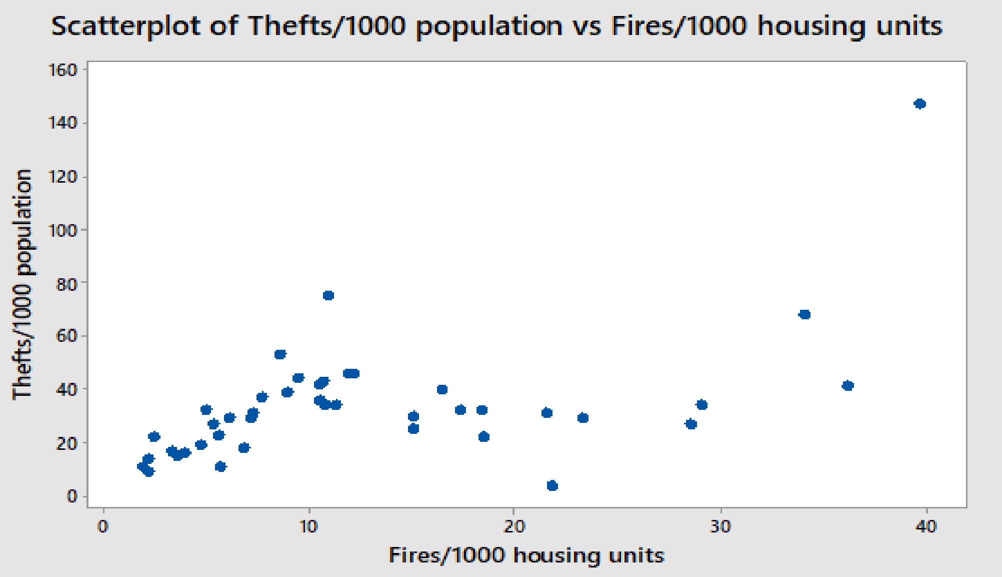


Figure 1: Scatter Plot

The plot shows the basic relationship (correlation) between the independent variable (x-axis) and the dependent variable (y-axis).

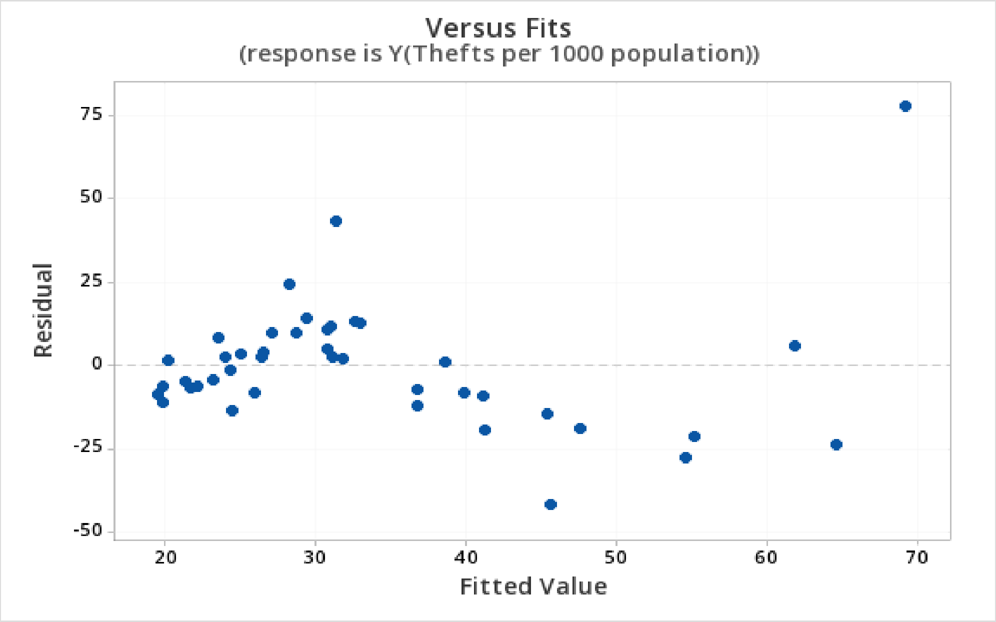
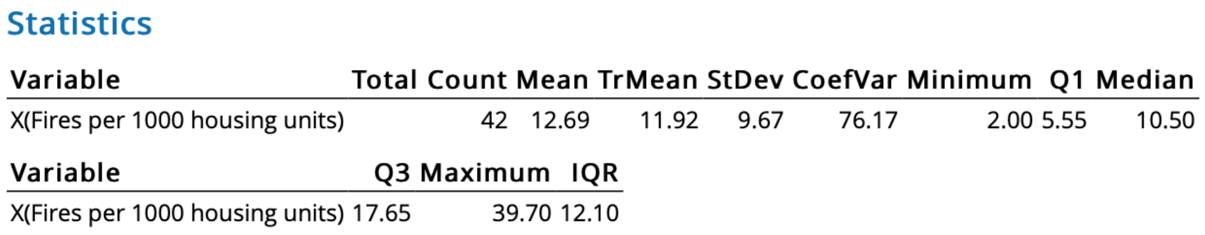


Figure 2: Residual Plot

The residual plot shows the relationship between the regression line and the data points and is used to find the problems in regression. The horizontal axis displays the independent variable and the vertical axis has the residual values.



Statistics

Variable	Total Count	Mean	TrMean	StDev	CoefVar	Minimum	Q1	Median
X(Fires per 1000 housing units)	42	12.69	11.92	9.67	76.17	2.00	5.55	10.50

Variable	Q3	Maximum	IQR
X(Fires per 1000 housing units)	17.65	39.70	12.10

Table 1: Descriptive Statistics of X

The table shows all the values of the independent variable 'X' which is the occurences of fires per 1000 households. The average value of thefts is 12.69 which is calculated by dividing the sum of all the thefts by the sample size. The standard deviation shows how much of a deviation exists around the mean value. The Q1 and Q3 tell us where 25% and 75% of the data fall respectively, with the median being the middle value.

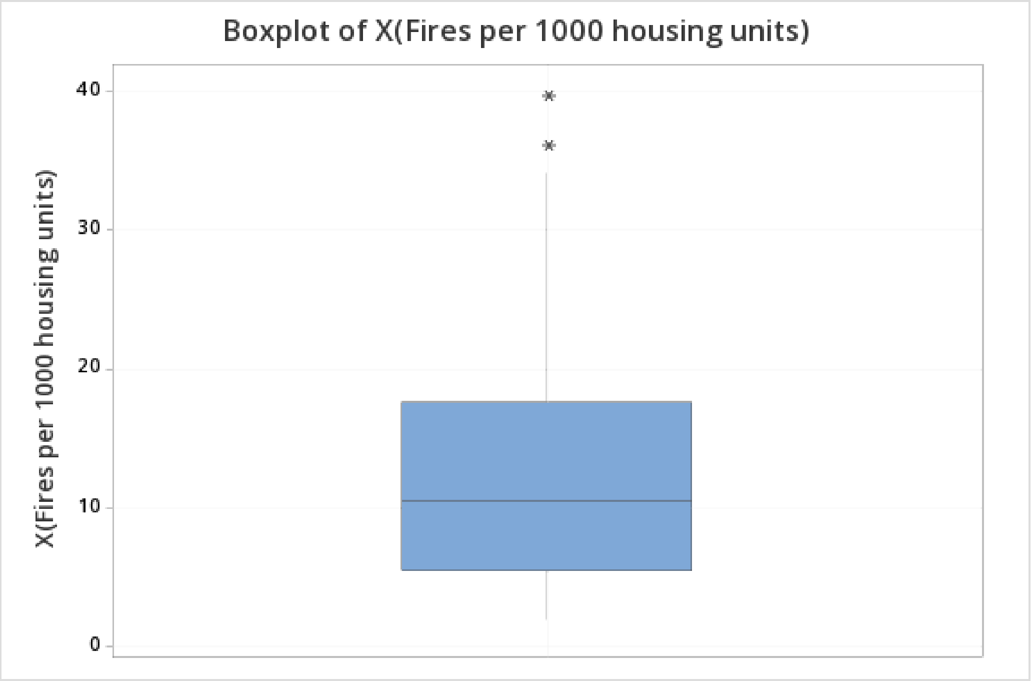
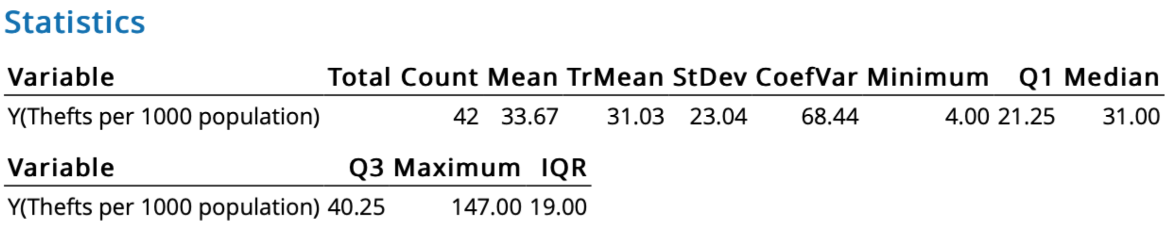


Figure 3: Boxplot of X

The boxplot is a graphical representation of the quartiles (Q1, Q3 & the median) and helps us find out if there are any outliers present. Outliers are values/points beyond the specified range and may negatively affect the analysis. As shown, there are two outliers present.



Statistics

Variable	Total Count	Mean	TrMean	StDev	CoefVar	Minimum	Q1	Median
Y(Thefts per 1000 population)	42	33.67	31.03	23.04	68.44	4.00	21.25	31.00

Variable	Q3	Maximum	IQR
Y(Thefts per 1000 population)	40.25	147.00	19.00

Table 2: Descriptive Statistics of Y

The table shows all the descriptive values of the dependent variable Y i.e. the number of thefts per 1000 population. The average number of thefts are 33.67 with the variability around this value being 23.04 which is the standard deviation. The mid value is 31 which is close to the average which is a good sign.

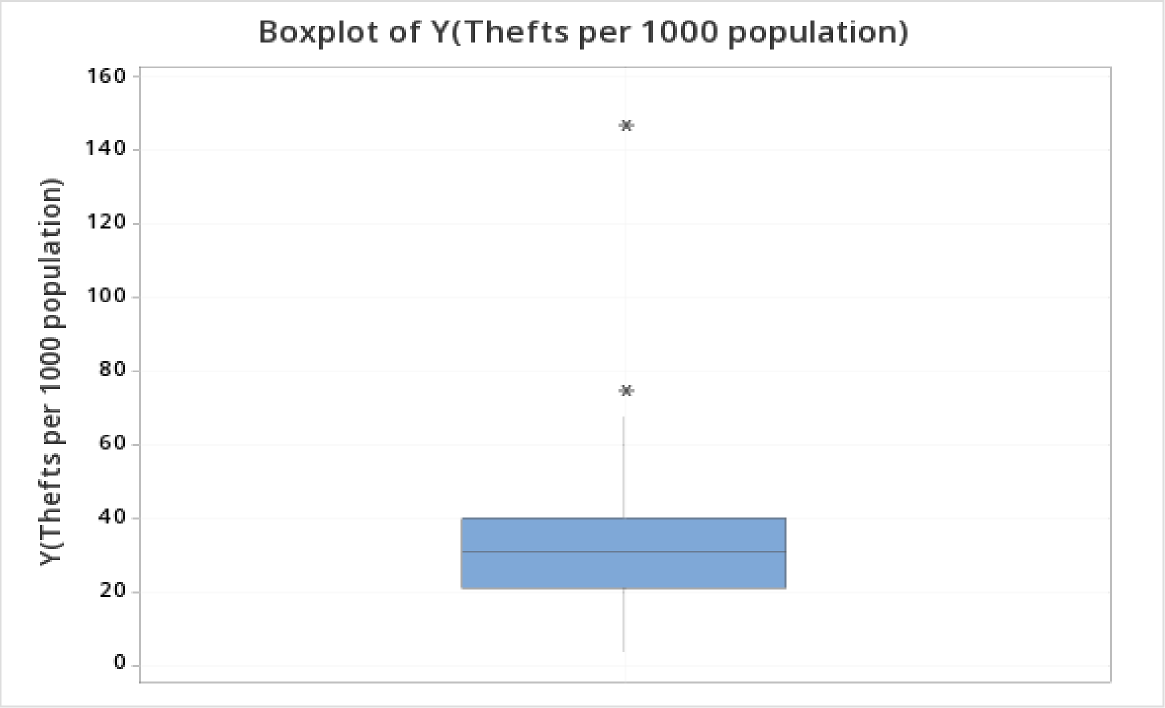


Figure 4: Box Plot of Y

The graphical representation of the quartiles also shows that we have two outliers. These values also correspond to the given independent variable.

Fits and Diagnostics for Unusual Observations

Obs	Y(Thefts per 1000 population)	Fit	Resid	Std Resid	
6	68.00	61.78	6.22	0.35	X
7	75.00	31.44	43.56	2.27	R
23	41.00	64.54	-23.54	-1.33	X
24	147.00	69.14	77.86	4.51	R X
29	4.00	45.63	-41.63	-2.19	R

R Large residual
X Unusual X

Table 3: Places with the most thefts

Analysis of the residual values help us identify the places with the highest occurrences of thefts. The high Y values as well as the bigger residual values are indicative of the safety of the area.

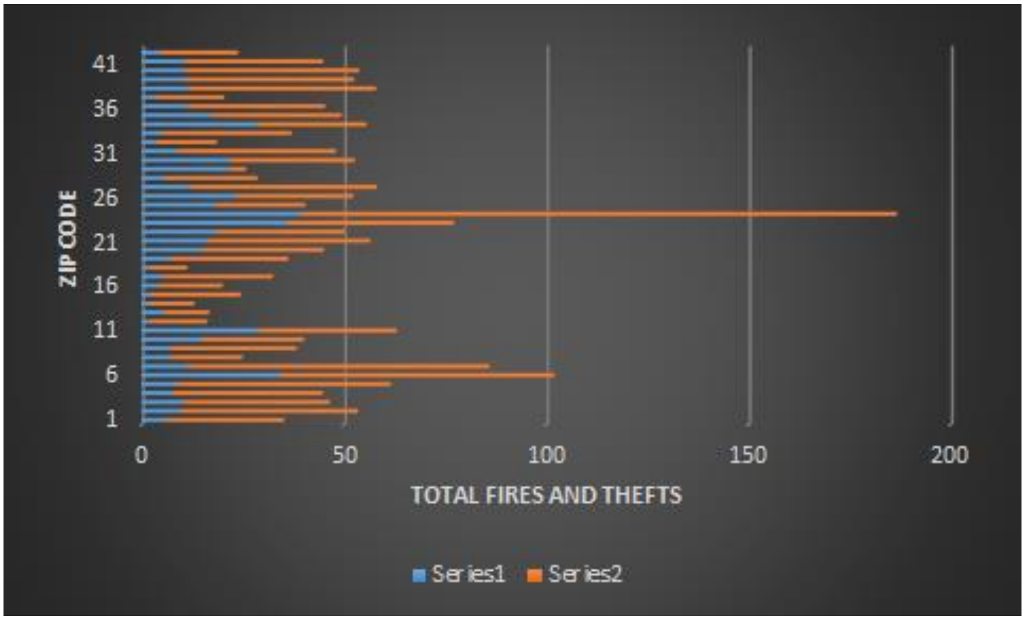
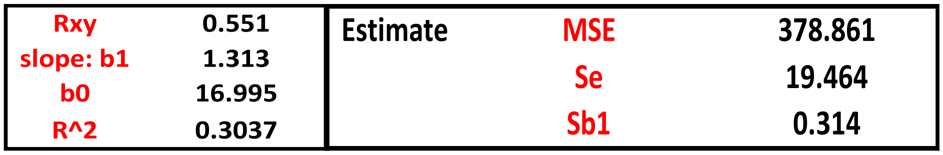


Figure 5: Stacked Bar Graph of Fires and Thefts

This graphical representation of the total number of fires and thefts is to indicate which area is safe and unsafe. The longer the line, the more occurrences of fire/theft and the area is considered unsafe. Simply by looking at the graph, we can see that area 24 is the most unsafe.

x	y	x-xbar	(x-xbar)^2	(y-ybar)	(y-ybar)^2	(x-xbar)(y-ybar)	yhat	y-yhat	y-yhat squared	yhat-ybar	(yhat-ybar)^2
6	29	-6.493	42.157	-4.667	21.778	30.300	25.139	3.861	14.911	-8.528	72.728
10	44	-3.193	10.194	10.333	106.778	-32.993	29.473	14.527	211.034	-4.194	17.587
11	36	-2.193	4.809	2.333	5.444	-5.117	30.786	5.214	27.181	-2.880	8.296
8	37	-4.993	24.929	3.333	11.111	-16.643	27.109	9.891	97.836	-6.558	43.006
9	53	-4.093	16.751	19.333	373.778	-79.129	28.291	24.709	610.541	-5.376	28.899
34	68	21.407	458.266	34.333	1178.778	734.979	61.784	6.216	38.639	28.117	790.585
11	75	-1.693	2.866	41.333	1708.444	-69.971	31.443	43.557	1897.197	-2.223	4.944
7	18	-5.793	33.557	-15.667	245.444	90.755	26.058	-8.058	64.931	-7.609	57.892
7	31	-5.393	29.083	-2.667	7.111	14.381	26.583	4.417	19.506	-7.083	50.173
15	25	2.407	5.794	-8.667	75.111	-20.862	36.828	-11.828	139.910	3.162	9.996
29	34	16.407	269.194	0.333	0.111	5.469	55.217	-21.217	450.150	21.550	464.405
2	14	-10.493	110.100	-19.667	386.778	206.360	19.885	-5.885	34.630	-13.782	189.941
6	11	-6.993	48.900	-22.667	513.778	158.505	24.482	-13.482	181.760	-9.185	84.361
2	11	-10.693	114.337	-22.667	513.778	242.371	19.622	-8.622	74.340	-14.045	197.251
3	22	-10.193	103.894	-11.667	136.111	118.917	20.279	1.721	2.963	-13.388	179.235
4	16	-8.693	75.566	-17.667	312.111	153.574	22.249	-6.249	39.050	-11.418	130.364
5	27	-7.293	53.186	-6.667	44.444	48.619	24.088	2.912	8.481	-9.579	91.754
2	9	-10.493	110.100	-24.667	608.444	258.824	19.885	-10.885	118.478	-13.782	189.941
7	29	-5.493	30.171	-4.667	21.778	25.633	26.452	2.548	6.492	-7.215	52.051
15	30	2.407	5.794	-3.667	13.444	-8.826	36.828	-6.828	46.626	3.162	9.996
17	40	3.807	14.494	6.333	40.111	24.112	38.667	1.333	1.776	5.001	25.005
18	32	5.707	32.571	-1.667	2.778	-9.512	41.163	-9.163	83.956	7.496	56.191
36	41	23.507	552.586	7.333	53.778	172.386	64.542	-23.542	554.238	30.876	953.303
40	147	27.007	729.386	113.333	12844.444	3060.810	69.139	77.861	6062.279	35.473	1258.312
19	22	5.807	33.723	-11.667	136.111	-67.750	41.294	-19.294	372.262	7.627	58.178
23	29	10.607	112.511	-4.667	21.778	-49.500	47.599	-18.599	345.911	13.932	194.101
12	46	-0.493	0.243	12.333	152.111	-6.079	33.019	12.981	168.498	-0.647	0.419
6	23	-7.093	50.309	-10.667	113.778	75.657	24.351	-1.351	1.824	-9.316	86.791
22	4	9.107	82.940	-29.667	880.111	-270.179	45.628	-41.628	1732.932	11.962	143.085
22	31	8.907	79.337	-2.667	7.111	-23.752	45.366	-14.366	206.376	11.699	136.870
9	39	-3.693	13.637	5.333	28.444	-19.695	28.816	10.184	103.709	-4.850	23.526
4	15	-9.093	82.680	-18.667	348.444	169.733	21.724	-6.724	45.207	-11.943	142.637
5	32	-7.693	59.180	-1.667	2.778	12.821	23.562	8.438	71.192	-10.104	102.095
29	27	15.907	253.037	-6.667	44.444	-106.048	54.560	-27.560	759.554	20.893	436.531
17	32	4.707	22.157	-1.667	2.778	-7.845	39.849	-7.849	61.611	6.183	38.225
11	34	-1.393	1.940	0.333	0.111	-0.464	31.837	2.163	4.678	-1.829	3.347
3	17	-9.293	86.357	-16.667	277.778	154.881	21.461	-4.461	19.900	-12.206	148.981
12	46	-0.793	0.629	12.333	152.111	-9.779	32.625	13.375	178.883	-1.041	1.084
11	42	-2.193	4.809	8.333	69.444	-18.274	30.786	11.214	125.744	-2.880	8.296
11	43	-1.993	3.971	9.333	87.111	-18.600	31.049	11.951	142.823	-2.618	6.851
11	34	-1.893	3.583	0.333	0.111	-0.631	31.180	2.820	7.950	-2.486	6.181
5	19	-7.893	62.297	-14.667	215.111	115.762	23.300	-4.300	18.488	-10.367	107.473
12.693	**33.667**	**0.000**	3832.028	0.000	21765.333	5033.200		0.000	**15154.447**	0.000	**6610.887**
xbar	**ybar**		93.464		**SST**	**122.761**			**SSE**		**SSR**
			9.668		530.862	**Sxy**			**378.861**		**6610.887**
			Sx		**23.040**				MSE		MSR
					Sy				**19.464**		
									Se		

Table 4: Error Analysis



Rxy	0.551	Estimate	MSE	378.861
slope: b1	1.313		Se	19.464
b0	16.995		Sb1	0.314
R^2	0.3037			

Table 5: Regression Model Summary using Excel

Analytical software gives us various values and coefficients that help us dig deeper into the regression model and analyze them. The table contains the correlation coefficient (Rxy), the value of the slope (b1), the Y-intercept (b0) and the coefficient of determination (R^2). It also contains the data for the estimated Y values namely the mean square error (MSE), standard error (Se) and the estimated standard deviation (Sb1) of the slope.

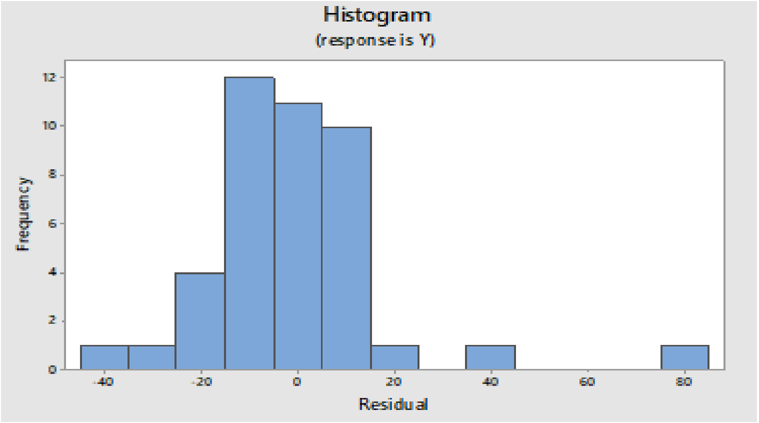
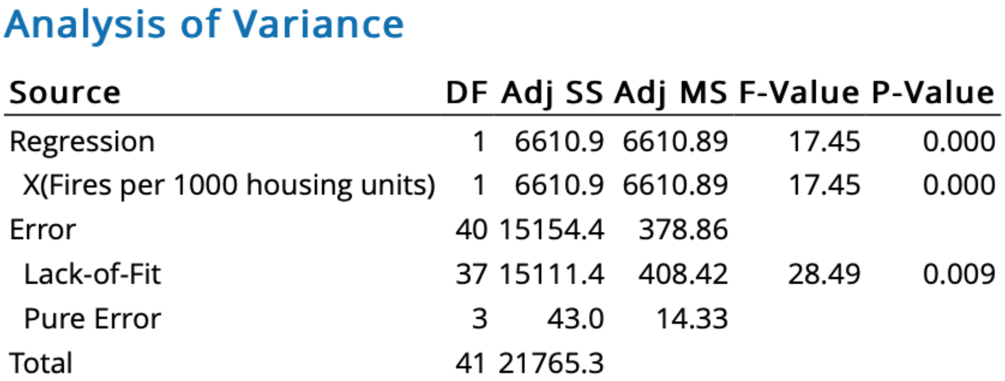


Figure 6: Histogram of Residuals

This histogram reveals that the residuals follow the likes of a normal distribution and hence we can carry on with our regression analysis and hypothesis testing.



Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	6610.9	6610.89	17.45	0.000
X(Fires per 1000 housing units)	1	6610.9	6610.89	17.45	0.000
Error	40	15154.4	378.86		
Lack-of-Fit	37	15111.4	408.42	28.49	0.009
Pure Error	3	43.0	14.33		
Total	41	21765.3			

Table 6: Analysis of Variance

The Analysis of Variance (ANOVA) tests the relationship between two or more variables. We use the F-value and the P-value and compare with our test statistic to prove our hypothesis. The P-value is the probability that the null hypothesis is true and we want the P-value to be as low as possible for us to reject the null hypothesis.

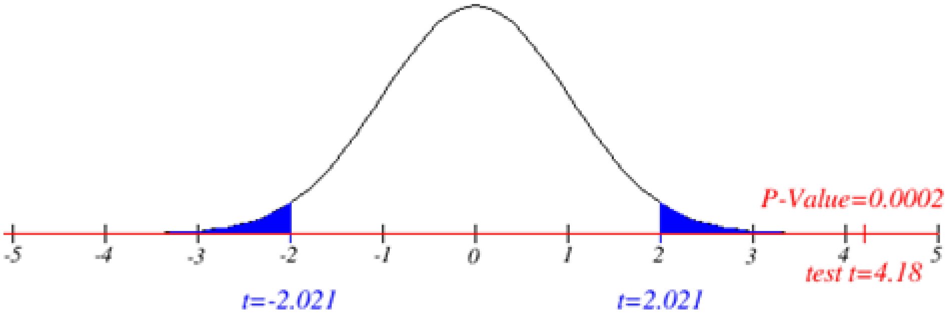
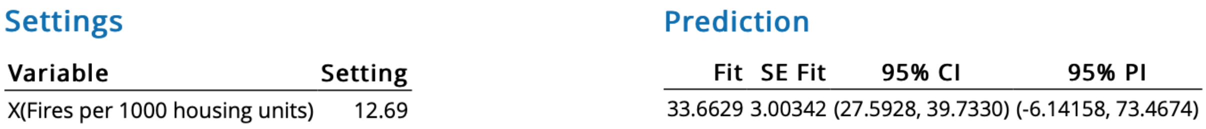


Figure 7: t-test

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	16.99515723	4.994880666	3.402515169	0.001528172	6.900126842	27.09018762	6.90012684	27.0901876
X Variable 1	1.313456005	0.31443126	4.177243716	0.000155322	0.677966723	1.948945286	0.67796672	1.94894529

Table 7: t-test Data

The t-test is used for hypothesis testing and to find a relationship between the means of the data. It follows a normal distribution as depicted by Figure 6. The blue shaded regions indicate the rejection region for a 95% confidence interval. Since we are testing the slope, we consider the t-stat value of 4.18 which is greater than 2.021 and falls in the rejection region.



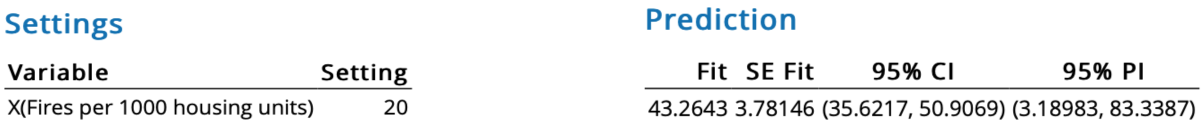
Settings

Variable	Setting
X(Fires per 1000 housing units)	12.69

Prediction

Fit	SE Fit	95% CI	95% PI
33.6629	3.00342	(27.5928, 39.7330)	(-6.14158, 73.4674)

Table 8: Confidence and Prediction Interval for X = 12.69



Settings

Variable	Setting
X(Fires per 1000 housing units)	20

Prediction

Fit	SE Fit	95% CI	95% PI
43.2643	3.78146	(35.6217, 50.9069)	(3.18983, 83.3387)

Table 9: Confidence and Prediction Interval for X = 20

A confidence interval is an estimate of the range of values within which a certain probability of the chosen parameter might exist. The prediction interval indicates that a future value might fall in the given range with a specified probability. For example, for an X value of 12.69, Table 8 shows that there is 95% probability that a new X will fall within the range -6.14158 and 73.4674. It also shows that the probability of finding an X value within the range 27.5928 and 39.7330 is 95%. The same can be inferred from Table 9 which takes a value of 20 for X.