Investigating the Relationship Between Estimated Dollar Loss of a Fire and the Extent of the Fire and the Number of People Displaced from the Fire through Frequentist Linear Regression

Assignment 2

Sophie Berkowitz - 1004768003

Introduction

Fires pose a serious threat to humans, the environment, infrastructure, and more. When people are exposed to fires, they risk breathing in toxic gases that can enter the respiratory system and cause damage to the body (Noble et al., 1980). Urban fires that occur in Toronto are especially dangerous as cities are more dense, consist of at-risk populations, and can spread more easily as everything is closer together (Ferreira et al., 2016). It is essential to reinforce urban fire risk mitigation and prevention in order to reduce the dangerous effects that fires have on the City of Toronto.

Although wildfires are associated with rural areas and natural forests, research is showing that wildfires are encroaching more on urban areas (Gee & Anguiano, 2020). Climate change has played a crucial role in increasing the risk that wildfires pose, thanks to the warmer and drier conditions in many vulnerable areas (Gee & Anguiano, 2020). As the climate crisis strengthens, more areas will be considered vulnerable to fires (Gee & Anguiano, 2020). This leaves cities in a delicate position, as urban fires already pose a threat.

The objective of this report is to evaluate the characteristics within a dataset containing information regarding fire incidents in Toronto that explain the estimated dollar loss in various fires that occurred. In market based economies such as Canada, the financial impacts of natural disasters influence how governments prioritize management strategies and how society's outlook on natural disasters evolve (Urry, 2015). The average natural disaster decreases economic growth by 1% and GDP by 2% (IBC, 2014). Beyond the grave environmental consequences, this impact extends beyond macro institutions such as governments and economies, to individual households and small business (IBC, 2014). This report will contribute to the academia regarding the cost of natural disasters that will further inform the importance of natural disaster mitigation and climate change policy. **We will test whether there is a relationship between the estimated dollar loss of a fire and the extent of the fire and the number of people displaced from the fire using multiple linear regression analysis.** Regression analysis allows us to predict future observations using the relationships between variables (Seber & Lee, 2012). We are able to understand the importance of key predictor variables and how they influence each other, and the response variable (Seber & Lee, 2012). Investigating key factors involved in fire incidents is essential to inform the trajectory of future fire incidents and how they relate to costs and affect humans. This model should assist in providing valuable information about the intensity of a fire and how it affects Torontonians.

Data

Data Collection Process

The data being used is a subset of the Office of the Fire Marshal and Emergency Management historical database, which acquires fire incident information from the fire departments on each call, as the Toronto fire services must report to the office for every call attended (TFS, 2021). Fire incidents are defined by the

Ontario Fire Marshal and only included in this report if they fit the stated criteria (Ontario Ministry of the Solicitor General, 2019).

Once the entire dataset was acquired, it was subsetted into a smaller dataframe dedicated to the analysis in this report. The relevant variables used for the regression analysis were included in the new data frame. These variables are listed as follows: estimated dollar loss, extent of fire, status of fire on arrival, estimated number of persons displaced, latitude, and longitude (TFS, 2021). Subsetting the data is an important aspect of the data cleaning process as it increases efficiency in the data analysis. The categorical variables, status of fire on arrival and extent of fire, were visualized as a numeric score based on the criterion imposed by the Marshal followed by a character description of the score as described in Table 4 of the Appendix. In order to improve efficiency in the data analysis portion and avoid redundancy, both variables were reduced to their number value, as the description for each fire incident is not required for the statistical analysis. Additionally, these categorical variables were converted from vector forms to numeric levels as categories. This allows R to clearly categorize the variables into bins.

Throughout the data, NA values were present for various reasons, such as the fire department lacking the proper knowledge to report the known information. As part of the data cleaning process, all rows that consisted of the NA values were omitted from the data frame for reasons of completeness and precision. Additionally, all fire incidents that exhibited empty cells in their rows were omitted. It is unclear why these cells were kept blank, therefore the most satisfactory decision was to omit this data.

The data consisting of NA and empty values are considered a limitation. It is unclear why this data is specified as empty and NA. Since we must resort to omitting this data, our sample size is more constricted and therefore, we might fail to grasp valuable information gained from this data. Given further investigation, there is a possibility that the data with these values demonstrates a relationship among fire incident data. For example, the data with empty cells for the extent of of fire variable, might correspond to fire incidents that exhibit an estimated lower or higher cost. Omitting this data has the possibility of introducing biases if it exhibits certain relationships.

Data Summary

The important variables in this report are estimated dollar loss, extent of fire, status of fire on arrival, estimated number of persons displaced, latitude, and longitude. Estimated dollar loss is a continuous numerical variable that demonstrates the financial impact that resulted from the fire incident. In our analysis, estimated dollar loss will be used as the response variable. Hence, the optimal model will be able to predict future observations of estimated dollar loss of fire incidents. This variable is of interest because the financial impact of a fire incident has far-reaching consequences in Canada. In terms of climate change, demonstrating the financial loss of fires should promote the mitigation of climate associated fires that ultimately addresses the climate crisis.

Extent of fire is a categorical variable that represents the intensity of the fire incident, in ascending order, which follows criterion set by the Ontario Fire Marshal. For instance, '1 - Confined to object of origin' is the most minor extent defined by the Marshal and '11 - Spread beyond building of origin, resulted in exposure fire(s)' is the most severe (Ontario Ministry of the Solicitor General, 2019). Similarly, status of fire on arrival is also a categorical variable that describes the severity of the fire at the time of the Toronto Fire Services arrival, where 1 is scored when the fire was extinguished prior to arrival and 8 is scored when there is exposure involved (TFS, 2021). Both categorical variables included categories of unknown data where the fire instance could not be awarded a category. The values were 9 and 999 for status of fire on arrival and extent of fire, respectively. This data was not removed to prevent bias. Estimated number of persons displaced is a powerful numerical variable that represents the amount of citizens that were forced to flee from their residence due to the fire incident. Longitude and Latitude correspond to the east-west and north-south coordinates on the Earth of the fire incident, respectively.

Table 1: The following Table 1 demonstrates the numerical summaries of the variables in the regression analysis. The summaries of the categorical variables, extent of fire and status of fire on arrival, reflect the count of of the six most prevalent categories within each variable. The remaining categories are grouped into their own element, (Other), as their respective counts are signficiantly low. The numerical variables – estimated dollar loss, estimated number of persons displaced, latitude, and longitude, are displayed along with their Minimum, 1st Quartile, Median, Mean, 3rd Quartile, and Maximum values. The Minimum and Maximum values represent each variable's lowest and highest numeric values, respectively. The 1st Quartile corresponds to where 25% of the data lies. Similarly, the 3rd Quartile corresponds to where 75% of the data lies. The median separates the top half of the data from the bottom half. The mean is equivalent to the sum of the components of the variable divided by the number of components. The numerical summaries of each element allow us to investigate the spread or distribution, which will inform our linear regression analysis. The NA values that appear in Table 1 under the numerical data are present due to the categorical variables being presented as 7 rows, whereas the numerical data information is presented in 6 rows. (continued below)

Estimated_Dollar_Loss	Extent_Of_Fire	Status_of_Fire_On_Arrival
Min. : 0	1 :5123	1 :3586
1st Qu.: 250	2 :4084	2 :2888
Median : 2500	3 : 481	3 :2628
Mean : 42944	4 : 469	4 :1415
3rd Qu.: 15000	9 : 342	5 : 250
Max. :50000000	7 : 274	7 : 238
NA	(Other): 441	(Other): 209

Estimated_Number_Of_Persons_Displaced	Latitude	Longitude
Min. : 0.00	Min. :43.59	Min. :-79.64
1st Qu.: 0.00	1st Qu.:43.66	1st Qu.:-79.48
Median : 0.00	Median :43.70	Median :-79.40
Mean : 17.27	Mean :43.71	Mean :-79.40
3rd Qu.: 1.00	3rd Qu.:43.75	3rd Qu.:-79.34
Max. :999.00	Max. :43.85	Max. :-79.12
NA	NA	NA

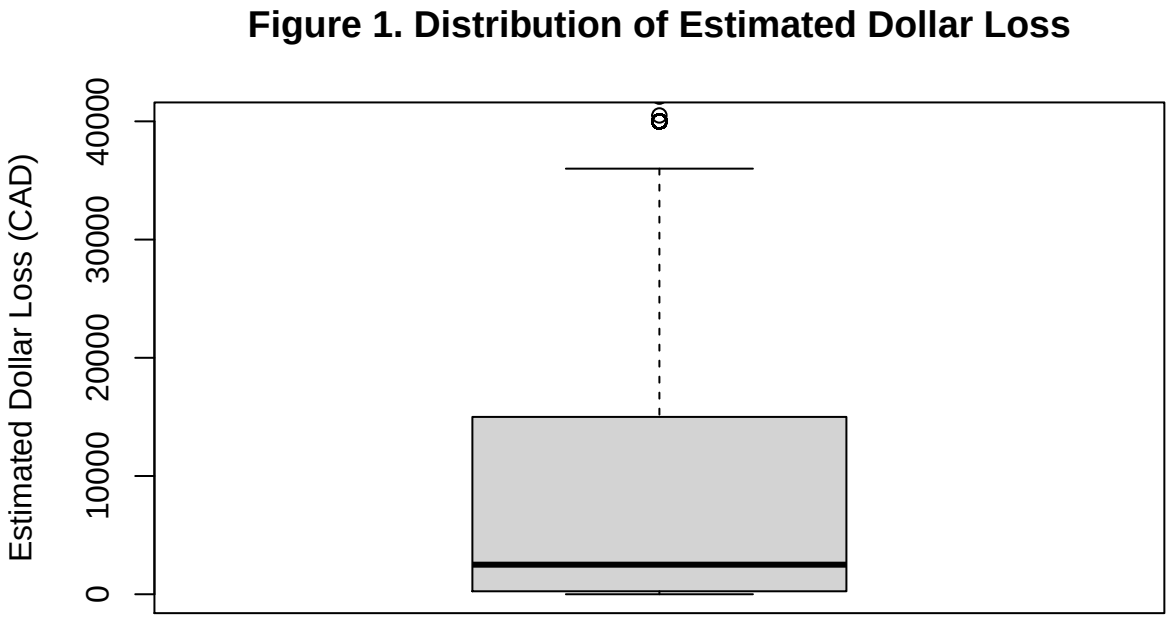

Figure 1. The box plot in Figure 1 demonstrates the distribution of the response variable, estimated dollar loss. The y-axis was restricted to $40,000 CAD, which only neglected the outliers, in order to highlight the 5 number summary and interquartile range (IQR). The IQR depicts the middle 50% of the data. The box plot appears positively skewed, as the median lies towards the lower values of the interquartile range and the distance between the 3rd quartile and the maximum is significantly greater than the distance between the 1st quartile and minimum. The positive skew provides evidence that the probability distribution of estimated dollar loss is NOT normally distributed, which violates the linear regression assumption of normality.

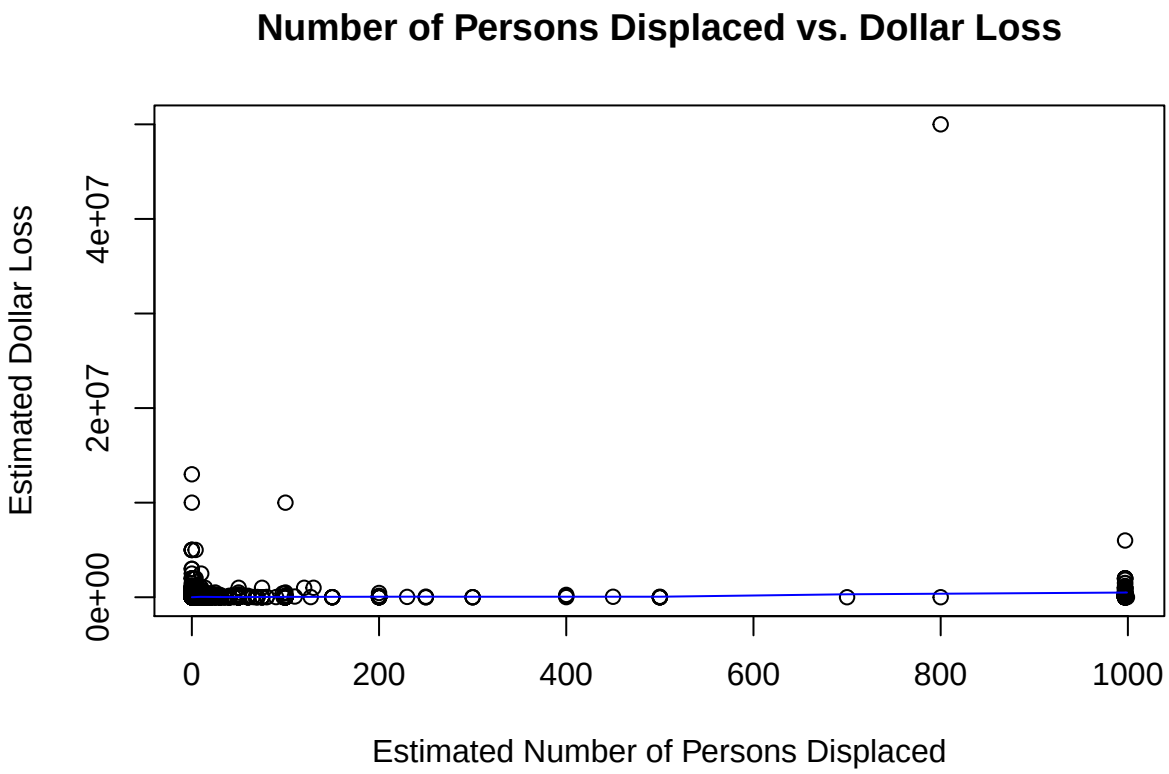

Figure 2. The scatterplot demonstrates the relationship between estimated number of persons displaced and estimated dollar loss. Evidently, there lacks a clear linear relationship based on the plot. Both variables appear concentrated at their lower values. This is expected based on the dataset. It should be less common for fire incidents to result in dramatically high costs and displace a high number of people. The lowess line, which produces a line through the data to demonstrate the general relationship appears to exhibit a slight upward trend as estimated number of persons displaced increases. Based on the plot, there seems to be an extreme point that exhibits high estimated number of persons displaced and estimated dollar loss values. This extreme point might be an influential point or outlier that is influencing the relationship between the two variables, and therefore the linear regression equation. This data point might explain the lack of clarity in the plot, as the y axis upper limit is set very high in order to accommodate for this point.

All analysis for this report was programmed using `R version 4.1.1`.

Methods

Linear regression analysis is a statistical tool that is used when trying to evaluate if a statistical linear relationship exists between the response variable and the various predictor variables (Seber & Lee, 2012). The explanatory or predictor variables correspond to the values that might predict or explain the trends visible in the response or dependent variable. Linear regression analysis rests upon the notion that our model will be able to predict the value of the response variable we expect to see given a value of the explanatory variable, which is known as the conditional mean response. However, real data does not lie perfectly according to the conditional mean. The error term accounts for the difference between the independent realizations of the random variable Y and this conditional mean response (Seber & Lee, 2012).

The multiple linear regression equation is: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_1 x_n + \epsilon$. β_0 represents the unknown parameter that is the intercept of the regression line, $\beta_1 ... \beta_n$ represent the unknown parameters that are the average change in estimated dollar loss when all predictor variables are held constant for each explanatory

variable corresponding to that coefficient, and ϵ represents the random error in the response, y. For the categorical variables, a reference category is assigned and all other categories within the variable will be compared to the reference category (Seber & Lee, 2012). The coefficients for the categorical variables represent the difference in intercepts between each variable and the reference. We create a random variable for all categories that are not the reference category (Seber & Lee, 2012). In the multiple linear regression model, the categorical random variable that is present at the time, can be factored into the model by setting it equal to 1, and the random variables representing the rest of the categories equal to 0.

Multiple Linear Regression makes key assumptions: 1) linearity 2) pairwise independence of errors 3) common error variance 4) normality of errors (Seber & Lee, 2012). Linearity assumes that the relationship between each explanatory variable and the predictor is linear. Ideally, we are looking for a linear pattern in the scatterplot and the residual plot to possess no systematic pattern. Pairwise independence of errors occurs if the random variables are independent of every other random variable and the observations, such as each fire incident, are sampled independently. Common error variance holds when the standardized residual plot exhibits uniformity with no verifiable effect, such as a funnel or filter. Normality of errors can be checked using Q-Q plots to test if two quantiles are both normally distributed. An optimal linear regression model will meet all 4 assumptions, otherwise it may distribute biased predictions. Multicollinearity tests whether the explanatory variables are highly correlated with each other, which can be measured using the variance inflation factor (VIF). The VIF indicates the ratio of the model's variance to the variance of a model that only includes the explanatory variable of interest. Generally, a VIF value that is greater than 10 is a cause for concern. If multicollinearity does exist, then the independent variables with a high VIF value should be removed from the model.

The most optimal model needs to consist of significant variables and will usually have a high R^2_{adj} and low information criterion values (AIC and BIC), which account for quality and complexity of the model. AIC attempts to select the model that describes an unknown operating model (Seber & Lee, 2012). BIC attempts to select the model that is closest to the true model (Seber & Lee, 2012). R^2_{adj} will inform us on how well the data fits the regression line. The residual standard error (RSE) evaluates the value of the standard deviation of ϵ and will indicate the amount that the response will differ from the predictions given by the model's regression line. Therefore, we want a small RSE. The p-value of the model as a whole can inform us on the model's significance, where the p-value will report the probability of producing the data results given the model in question. It is also important to inspect each coefficient's p-value in the linear model summary which will test the null hypothesis that the coefficient is equal to 0. In both cases, we are hoping for a low p-value. Typically, a p-value that is less than 0.05 can be considered significant, which indicates that the data compatibility with the null hypothesis is incredibly low. The F-statistic measures the ratio of the model variance to the unexplained noise produced by the model (Seber & Lee, 2012). We are looking for a high F-statistic and a low p-value. Usually, statistically worrisome variables, such as outliers, might produce a less significant model. We will use these measures to compare models consisting of different indicator variables in question.

Results

All analysis for this report was programmed using `R version 4.1.1`. I used the `lm()` function in base `R` to derive the estimates of a frequentist linear regression in this section. The assumption, pairwise independence of errors, was met because we know that the random variables are independent of every other independent variable and each fire incident is sampled independently by the TFS (TFS, 2021). The scatterplots of continuous variables in Figure 5, demonstrate the relationships between the variables and explore the linearity assumption. It is evidence from Figure 5 that the continuous variables lack strong linear relationships. Extent of fire and status of fire on arrival are both considered categorical data. When the full model was run that included all of the variables of the subsetted data set for the regression, extent of fire and estimated number of persons displaced were significant. Once the problematic variables were removed, the F-statistic value increased from 27.24 to 45.21. The F statistic 45.21, explains that at least one explanatory variable in the model produces a significant effect. The p-value is shown for that test, which compares the model that has been outputted to a hypothetical model that only includes the intercept. The null hypothesis states

that there is no linear association between estimated dollar loss, extent of fire, and estimated number of people displaced. We reject the null hypothesis, as p-value < 2.2e-16. The R^2_{adj} value, shows that 5% of the variability in estimated dollar loss is explained by estimated number of persons displaced and extent of fire, which is much lower than we are hoping for.

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_{estimated\ number\ of\ persons\ displaced} + \hat{\beta}_2 x_{Extent\ of\ Fire\ 10} + \hat{\beta}_3 x_{Extent\ of\ Fire\ 11} + \hat{\beta}_4 x_{Extent\ of\ Fire\ 2} + \hat{\beta}_5 x_{Extent\ of\ Fire\ 3} + \hat{\beta}_6 x_{Extent\ of\ Fire\ 4} + \hat{\beta}_7 x_{Extent\ of\ Fire\ 5} + \hat{\beta}_8 x_{Extent\ of\ Fire\ 6} + \hat{\beta}_9 x_{Extent\ of\ Fire\ 7} + \hat{\beta}_1 0 x_{Extent\ of\ Fire\ 8} + \hat{\beta}_1 1 x_{Extent\ of\ Fire\ 9} + \hat{\beta}_1 2 x_{Extent\ of\ Fire\ 99} + e_i.$

$\hat{y} = 5294.6 + 399.8 x_{estimated\ number\ of\ persons\ displaced} + 112708.3 x_{Extent\ of\ Fire\ 10} + 196007.7_{Extent\ of\ Fire\ 11} + 9229.3_{Extent\ of\ Fire\ 2} + 58550.1_{Extent\ of\ Fire\ 3} + 123590.9_{Extent\ of\ Fire\ 4} + 123350.1_{Extent\ of\ Fire\ 5} + 278741.6_{Extent\ of\ Fire\ 6} + 257207.5_{Extent\ of\ Fire\ 7} + 908006.4_{Extent\ of\ Fire\ 8} + 38291.3_{Extent\ of\ Fire\ 9} + 38291.3_{Extent\ of\ Fire\ 99} + e_i$

Model Interpretations:

- The average estimated dollar loss when the predictors are equal to 0 is $5,294.60 CAD.
- For a one-unit increase in estimated number of persons displaced, we see on average an increase in $399.80 CAD in average estimated dollar loss when the predictor extent of fire is fixed at a constant value.
- Fire incidents with extent of fire classified as category 2 result in an average estimated dollar loss of $196,007.7 greater than category 1
- Fire incidents with extent of fire classified as category 3 result in an average estimated dollar loss of $5,8550.1 greater than category 1
- Fire incidents with extent of fire classified as category 4 result in an average estimated dollar loss of $123,590.9 greater than category 1
- Fire incidents with extent of fire classified as category 5 result in an average estimated dollar loss of $123,350.1 greater than category 1
- Fire incidents with extent of fire classified as category 6 result in an average estimated dollar loss of $278,741.6 greater than category 1
- Fire incidents with extent of fire classified as category 7 result in an average estimated dollar loss of $257,207.5 greater than category 1
- Fire incidents with extent of fire classified as category 8 result in an average estimated dollar loss of $908,006.4 greater than category 1
- Fire incidents with extent of fire classified as category 9 result in an average estimated dollar loss of $38,291.3 greater than category 1
- Fire incidents with extent of fire classified as category 10 result in an average estimated dollar loss of $112,708.3 greater than category 1
- Fire incidents with extent of fire classified as category 11 result in an average estimated dollar loss of $196,007.7 greater than category 1
- Fire incidents with extent of fire classified as category 99 result in an average estimated dollar loss of $34,033.2 greater than category 1

The vif values for the explanatory variables in both the full model and the final model were all around 1, which indicates that the independent variables do not exhibit multicollinearity.

```
##                                       GVIF Df GVIF^(1/(2*Df))
## Estimated_Number_Of_Persons_Displaced 1.054759  1        1.027014
## Extent_Of_Fire                        1.054759 11        1.002426
```

```
##                                       GVIF Df GVIF^(1/(2*Df))
## Extent_Of_Fire                        1.633915 11        1.022568
## Status_of_Fire_On_Arrival             1.568070  7        1.032654
## Estimated_Number_Of_Persons_Displaced 1.057734  1        1.028462
## Latitude                              1.121694  1        1.059101
## Longitude                             1.114979  1        1.055926
```

Table 2. Statistics for both model outputs for the full model that includes all of the perspective predictor variables and mod 2, which only includes extent of fire and estimated number of persons displaced as predictor variables. The * indicates the chosen model for this report. R^2_{adj} indicates the sample variability in the estimated dollar loss of fire incidents that can be explained by each model and accounts for each additional predictor (Seber & Lee, 2012). We are looking for a higher R^2_{adj} value. The p-value and F-statistic indicate if the explanatory variables are linearly related to estimated dollar loss. Ideally, the best model has a very low p-value and a large F-statistic. The residual standard error also indicates the goodness of fit, as it is equivalent to the square root of the residual sum of squares divided by the degrees of freedom of the residual. AIC and BIC measure goodness of fit and account for model complexity. The best model has a small AIC and BIC.

	R^2_{adj}	p-value	F-statistic	Residual Standard Error	AIC	BIC
Full Model	0.04685	< 2.2e-16	27.24 on 21 and 11192 DF	521300 on 11192 degrees of freedom	327091.1	327259.5
Mod2 (model excluding the insignifcant predictors)*	0.04518	< 2.2e-16	45.21 on 12 and 11201 DF	521700 on 11201 degrees of freedom	327101.7	327204.3

Table 3. The intercept coefficient listed under 'Estimates' indicate the expected response, which occurs when all of the predictors are 0. The remaining estimates correspond to the coefficients that give the expected change in the estimated dollar loss due to a one unit change in the predictor when all other predictors are held constant. The standard error for each estimate corresponds to the variation of an observation compared to the regression line. The t-value indicates the difference between the estimate and the hypothesized value, weighted by the standard deviation estimate (Seber & Lee, 2012). The higher the t-value, the greater the difference. The Pr(>|t|) measures the p-value for the previous t-test. Ideally this value is below the critical value, 0.05.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5294.5691	7295.88079	0.7256929	0.4680422
Estimated_Number_Of_Persons_Displaced	399.8076	42.49701	9.4078992	0.0000000
Extent_Of_Fire10	112708.2940	70343.41045	1.6022580	0.1091268
Extent_Of_Fire11	196007.7206	49005.07449	3.9997433	0.0000638
Extent_Of_Fire2	9229.3159	10946.95946	0.8430940	0.3991939
Extent_Of_Fire3	58550.0800	24893.31749	2.3520401	0.0186880
Extent_Of_Fire4	123590.9279	25236.88715	4.8972334	0.0000010
Extent_Of_Fire5	123350.0825	81829.13568	1.5074103	0.1317337
Extent_Of_Fire6	278741.5853	89853.05868	3.1021936	0.0019257
Extent_Of_Fire7	257207.4776	32417.88314	7.9341232	0.0000000
Extent_Of_Fire8	908006.4164	53086.08876	17.1044136	0.0000000

	Estimate	Std. Error	t value	Pr(>|t|)
Extent_Of_Fire9	38291.2951	29139.17810	1.3140829	0.1888452
Extent_Of_Fire99	34033.2068	55176.73185	0.6168036	0.5373768

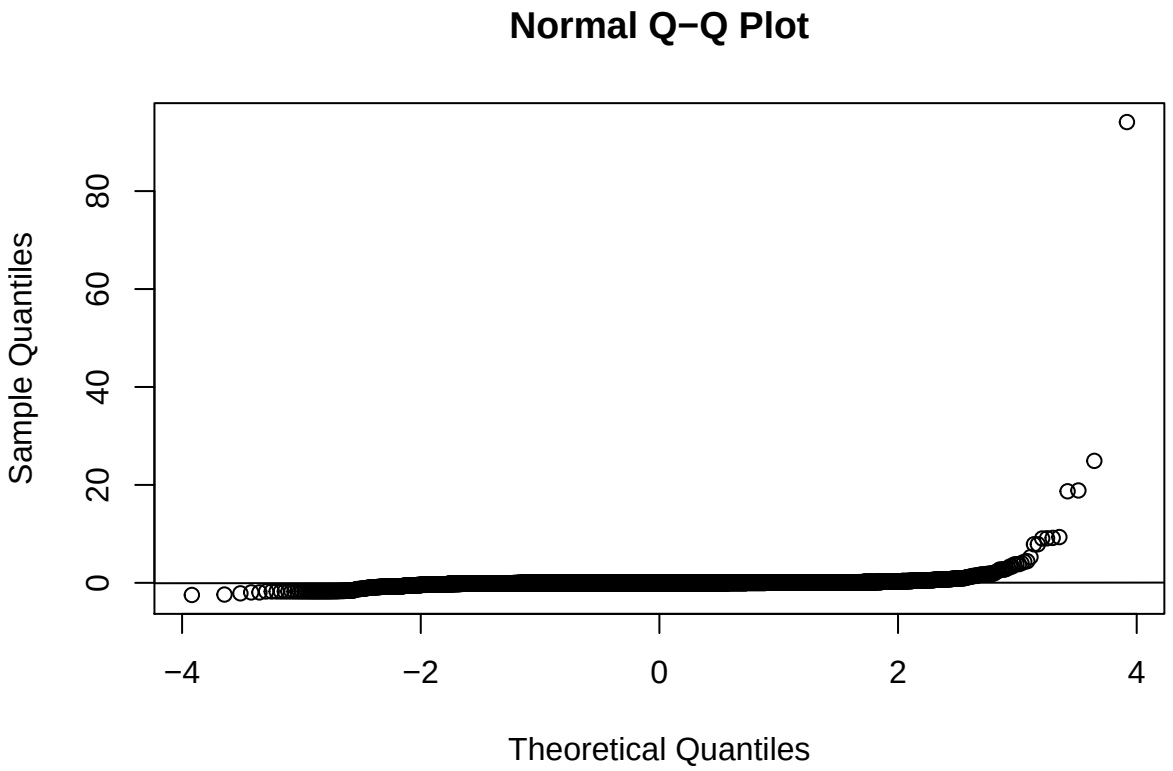

Figure 3. The Normal Q-Q Plot is comparing the probability distributions of the theoretical quantiles and the sample quantiles. It allows us to identify whether both datasets belong to the same distribution, which in our case is normal. If both data sets belong to the normal distribution, then the data points in Figure 3 should lie along the qqline. The data lying along the qq-line indicates that both data sets belong to the same distribution. The data points exhibit an almost flat trend due to the characteristics of the data that show that the data varies slowly over a small range. However, the tails at the higher value axes are heavier, meaning that the data varies more quickly. Due to this, the Normal Q-Q plot seems to violate the normality of errors assumption, as the q-q line should exhibit a greater slope that should lie more diagonal across the plane, and the data points should therefore lie along it. Instead of normal, the data exhibits a more asymmetric right-tailed distribution.

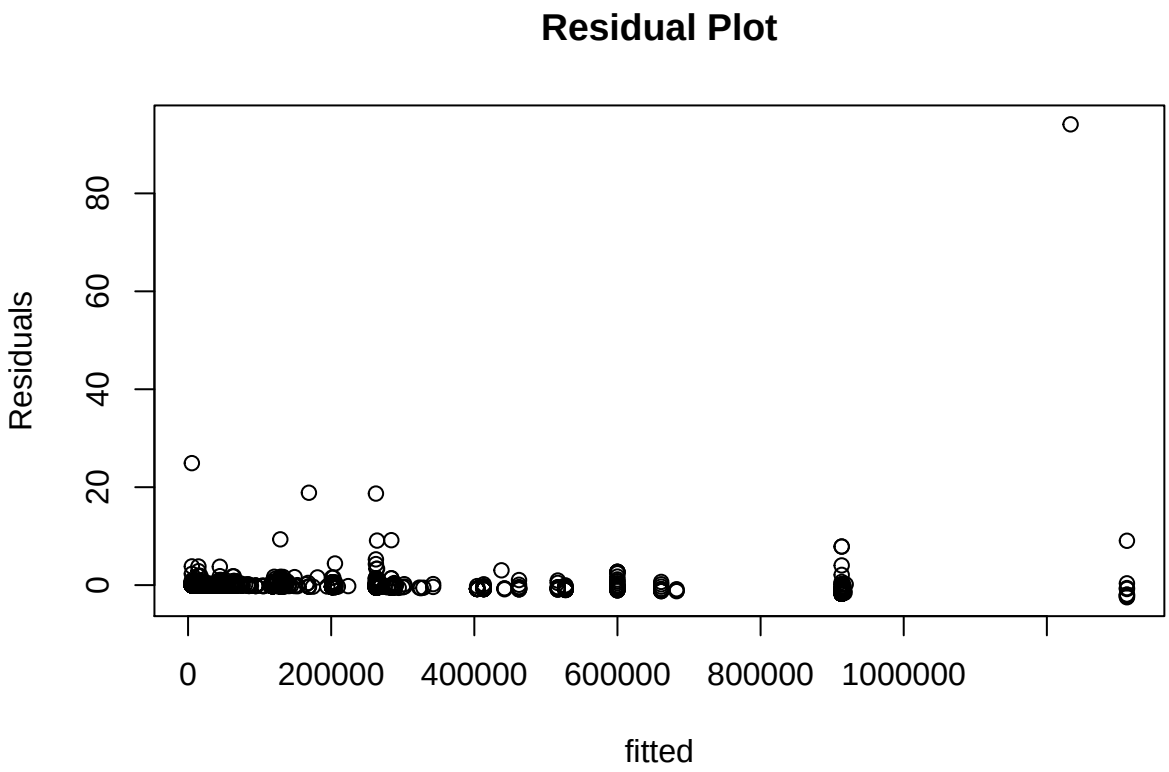

Figure 4. The residual plot reflects both the clumping and skew that was expected prior to the linear regression analysis. However, there does not appear to be any clear systematic pattern or curvature in the residual plot. Ideally, the residuals would randomly lie around the horizontal line at 0. In general, the data is mostly concentrated at around 0 and seems to be present randomly around this value. This indicates that the variances of the error term are approximately equal for most of the data points which meets the common error variance assumption. Therefore, it is fairly reasonable to assume linearity. The extreme influential point or outlier is contributing to the lack of clarity in the residual plot and causing the y-axis upper limit to exist at a high value.

The Response vs. Fitted values plot is located in the appendix, and it visualizes the scatterplot with the regression line overlaid (Figure 6).

Conclusions

As the climate crisis continues to wreak havoc on all walks of life and resources, systematic change will require data driven evidence. In order to do this, it is valuable to investigate the financial impacts of key climate disasters. According to scientists, wildfires will have a greater impact on urban areas, like Toronto. The report seeks to bridge the gap between the science of climate change and the reality of our market based society by investigating the relationship between the estimated dollar loss of fire incidents and the extent of the fire and the number of people displaced from the fire. Multiple linear regression analysis assisted us in understanding the importance of significant explanatory variables and how they influence our response variable, estimated dollar loss. Using the model that identifies the relationships, we are able to predict future observations and determine the certainty of relationships.

In order to conduct the regression analysis, we subsetted a large dataset provided by the Toronto Fire Services into the dataset that we were interested in, by removing variables we did not use. Before performing the regression analysis, it was crucial to become familiar with the data and variable properties by investigating

numerical summaries, distribution, and spread of the variables. Most of the variables of interest exhibit skew, which was not alarming based on the context of our data. However, skew might indicate a need for transforming data which was not performed in this analysis. Initially, the full model was run which included all of the predictor variables, extent of fire, status of fire on arrival, estimated number of persons displaced, latitude, and longitude. However, based on the full model output, only extent of fire and estimated number of persons displaced were kept based on the estimate's p-values and the overall model statistics such as R^2_{adj}. The final model only included the two predictors and produced superior statistics compared to the full model. The AIC and BIC values improved, along with R^2_{adj} and the residual standard error. Model assumptions were checked using the Normal Q-Q plot, residual plots, vif tests, and contextual information.

The findings indicate that extent of fire and estimated number of persons displaced have a significant influence on estimated dollar loss. This indicates that as the number of persons displaced increases, as does the estimated dollar loss. Additionally, the category of the extent of fire for each incident influences the estimated dollar loss. In a broader context, this information informs estimated dollar loss of future incidents. In terms of mitigation, this data indicates to government officials, policy makers, business developers, and citizens of the city that more intense fires, based on the extent and the people displaced, will result in greater costs. This analysis should encourage institutions to adopt climate policies to mitigate and prevent dangerous fires that will result in costly damages.

Weaknesses

Although the final model that was produced was considered the most optimal based on certain statistics, there still exists lingering issues. There appeared to be skew among most of the variables, which could be fixed using transformations. It is possible that the outliers affected the regression surface based on the figures and tables that depict the variables and their spreads and the residual plots. However, the outliers weren't removed from the model as there was no reason to remove them based on the context of the data. Both the outliers and the skewed data can alter the equation of the model and influence future predictions made by the model.

Next Steps

If I had more time I would identify influential points and outliers more closely. I would also apply transformations that might improve the model and its performance. Additionally, I would utilize alternative model selection processes, such as stepwise selection.

Discussion

This report investigates whether there is a relationship between fire incident intensity and cost resulting from the fire. In our context, intensity of the fire is defined by the extent of fire and number of persons displaced. Our model indicated that there was a relationship between estimated dollar loss and extent of fire and number of persons displaced. Unfortunately, climate change is associated with increasing the intensity of climate disasters, such as wildfires and wildfires are encroaching on urban areas, such as Toronto (Gee & Anguiano, 2020). This data should indicate that as the climate crisis worsens with the lack of action from policy makers and institutions, then the cost of these disasters will only grow. Therefore, climate policy should be enacted and change should be made.

Bibliography

1. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: October 12, 2021)
2. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.
3. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: October 12, 2021)
4. Peter Dalgaard. (2008) *Introductory Statistics with R, 2nd edition.*
5. Zhu, Hao. (2021). Rstudio. [https://cran.r-project.org/web/packages/kableExtra/vignettes/awesome_table_in_html.html]. (Last Accessed: October 24, 2021)
6. Noble, I. R., Gill, A. M., & Bary, G. A. V. (1980). McArthur's fire-danger meters expressed as equations. Australian Journal of Ecology, 5(2), 201-203.
7. Ferreira, T. M., Vicente, R., da Silva, J. A. R. M., Varum, H., Costa, A., & Maio, R. (2016). Urban fire risk: Evaluation and emergency planning. Journal of Cultural Heritage, 20, 739-745.
8. Gee & Anguiano. (2020). Wildfires are striking closer and closer to cities. We know how this will end. the Guardian. Retrieved from https://www.theguardian.com/commentisfree/2020/sep/12/wildfires-are-striking-closer-and-closer-to-cities-we-know-how-this-will-end
9. Xu, R., Yu, P., Abramson, M. J., Johnston, F. H., Samet, J. M., Bell, M. L., ... & Guo, Y. (2020). Wildfires, global climate change, and human health. New England Journal of Medicine, 383(22), 2173-2181.
10. Urry, J. (2015). Climate change and society. In Why the social sciences matter (pp. 45-59). Palgrave Macmillan, London.
11. Insurance Bureau of Canada (IBC). (March, 2014). Reducing the Fiscal and Economic Impact of Disasters. Retrieved from http://www.ibc.ca/qc/resources/studies/reducing-the-fiscal-and-economic-impact-of-disasters
12. Seber, G. A., & Lee, A. J. (2012). Linear regression analysis (Vol. 329). John Wiley & Sons.
13. Toronto Fire Services (TFS). (2021). Fire Incidents - City of Toronto. Open Data Toronto. Retrieved from https://open.toronto.ca/dataset/fire-incidents/
14. Ontario Ministry of the Solicitor General. (2019). Ontario Fire Incident Summary. Retrieved from https://www.mcscs.jus.gov.on.ca/english/FireMarshal/MediaRelationsandResources/FireStatistics/OntarioFires/AllFireIncidents/stats_all_fires.html

Appendix

Table 4. Extent of Fire Categorical Data. The following table displays the extent of fire category breakdown and each category's corresponding description supplied by the Office of the Fire Marshal and Emergency Management.

category	description
1	Confined to object of origin
2	Confined to part of room/area of origin
3	Spread to entire room of origin
4	Spread beyond room of origin, same floor
5	Multi unit bldg: spread beyond suite of origin but not to separated suite(s)
6	Multi unit bldg: spread to separate suite(s)

category	description
7	Spread to other floors, confined to building
8	Entire Structure
9	Confined to roof/exterior structure
10	Spread beyond building of origin
11	Spread beyond building of origin, resulted in exposure fire(s)
99	Undetermined

Figure 5. Scatterplot of Continuous Variables

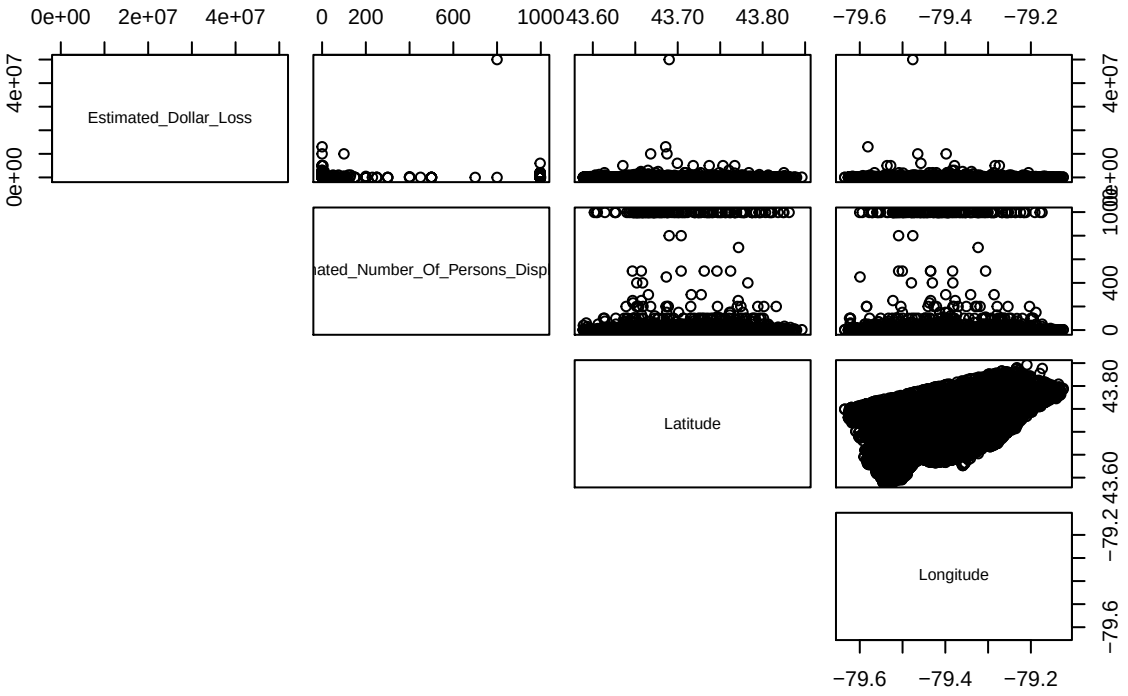

Figure 5. Scatterplots demonstrating the relationship between continuous variables within the fire incident dataset. Latitude and longitude appear to exhibit a somewhat linear relationship. Based on each scatterplot, there is a lack of strong linear relationships among many of the variables. However, there does not appear to be any curvature or systematic patterns either, which might indicate an alternative statistical relationship. Figure 2 demonstrates skew properties in most of the variables, as most of the data is concentrated in certain areas. For instance, excluding outliers, estimated dollar loss data is concentrated in the lower value portion of all of the plots. This is unsurprising as it is less common for fire incidents to result in extremely high dollar losses. Similarly, estimated number of persons displaced is heavily concentrated at lower values, as it significantly less common for fire incidents to cause the displacement of citizens.

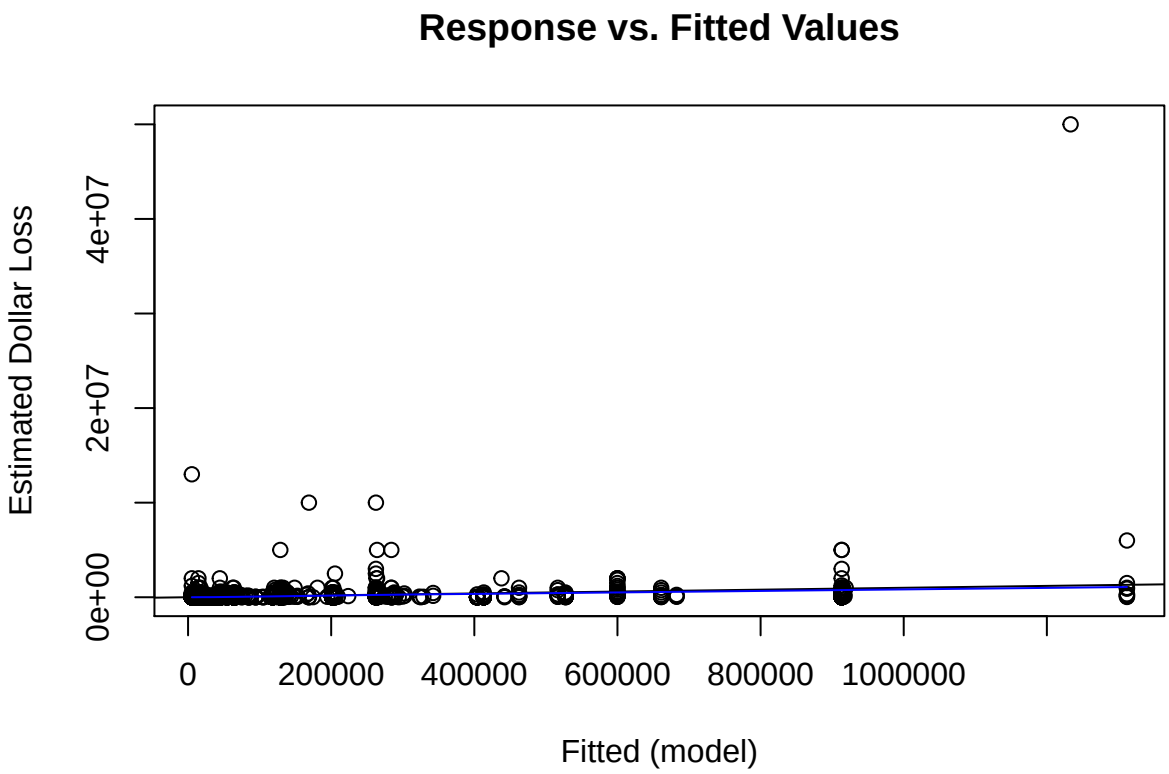

Figure 6. This plot demonstrates the response values against the fitted values. The fitted values indicate the linear regression model's prediction of the mean response value given the values of the predictors. The regression line is overlaid.