# Final Project Part 3 - Data Analysis Report

Mohsin Reza

2021-12-17

## Introduction

The goal of this study is to answer the following research question: Does the nature and efficiency of the response to a fire incident impact the severity of the damage it causes?. Fire incidents are faced around the world on a daily basis, and can cause serious damage to people's lives, as well as property. Therefore, this study is crucial as it has the potential to identify ways of limiting the damage caused by fires. Although widespread research has been conducted on fire incidents, most research papers that were found focused mainly on factors that caused a fire. For example, a study published by National Center for Biotechnology Information (Song, Kwan, & Zhu, 2017) found that road density and population distribution had the most positive influence on fire risk. Almost no studies were found that investigated factors that affect the severity of the damage caused by fire incidents, which further highlights the importance of our study.

## Methods

Linear regression was the primary tool used to answer the proposed question. The selected data has 17536 observations and 43 variables related to fire incidents in the City of Toronto between 2014 and 2019. In our analysis, we chose the "Estimated Dollar Loss" variable as our response. To measure the "nature of the response", the number of responding personnel and apparatus were chosen as potential predictors. Also, the number of minutes taken to respond to the fire was chosen as a predictor to measure the "efficiency of the response" in our research question.

### Variable Selection

First, the dataset was divided into two datasets, called training and test, of equal size. The model was then built using the training dataset. A variation of the all possible subsets method was used to select variables. It involved fitting all seven possible combinations of predictors, and comparing them using the measures adjusted R squared, AIC, corrected AIC, and BIC. Additionally, the models were compared in how well they satisfied the assumptions, how many influential observations they had, and how much multicollinearity was present. How these three things were analyzed will be explained in the model diagnostics section. The model which satisfied assumptions best, had the least amount of influential observations/multicollinearity, had a significant linear relationship (using the p-values from the partial F test), had the highest adjusted R squared, and had the lowest AIC/BIC was selected as the final model. If there was no model that satisfied all of these properties, then the model which satisfied most of them was selected.

### Model Validation

To validate the model, we first fit it using the test dataset. Then, we compared the coefficients, p values for the partial F and t-tests, adjusted R squared, and model assumptions across the two models. If the coefficients were similar, it would be appropriate to conclude that a similar relationship was estimated in the two datasets. Additionally, a similar adjusted R squared would indicate that both models are similarly good at explaining variation, and similar p-values i.e. the same significant predictors would indicate that the test

dataset was not overfit. If all of these measures were similar, we would conclude that the model was correctly validated.

### Model Violations and Diagnostics

To check if model assumptions were violated, we first checked if additional conditions 1 and 2 were met. We concluded that condition 1 was satisfied if a plot between the actual and predicted response showed an association between the two. Condition 2 was checked by visually examining pariwise relationships between predictors, and ensuring they were linear. Then, to check normality was violated, a qq plot was used, with a straight diagonal string of points indicating normality was met. To check the other assumptions, the residuals were plotted against fitted values as well as predictors. Any systematic pattern indicated a violation of the other three assumptions. If assumptions were not satisfied, we attempted to transform the predictors/response to better satisfy them.

Additionally, the presence of influential observations was checked by computing the Cook's distance, DFFITS, and DFBETAS values. If a measure was above its respective cutoff, we had an influential observation. Finally, multicollinearity was detected by computing the VIF for each variable. If the VIF was above 5, multicollinearity was deemed to be unacceptably high. Influential observations and multicollinearity were only handled by selecting predictors in a way that minimzed their impact. # Results

## Discussion

### Final Model Interpretation and Importance

Recall that our final model is:

$$\text{Estimated Dollar Loss}^{1/4} = 3.702381 + 0.111750 * (\text{Number of responding personnel}) + 0.298717 * (\text{Response time}) + \epsilon$$

In terms of interpretation, we know that the mean fourth root of the estimated dollar loss is 3.702381 when there are no responding personnel, and the fire is responded to instantly (0 response time). Additionally, 0.111750 is the average change in the fourth root of the estimated dollar loss when the number of responding personnel are increased by 1 and the response time is held fixed. Similarly, 0.2987171 is the mean change in the fourth root of the estimated dollar loss when the response time is increased by one minute and the number of responding personnel are held fixed. This model is relevant to the goal of the study as it tells us that the slower a fire is responded to i.e. the lower the efficiency of the response, the higher the monetary damage it can cause. Additionally, the model also shows that a larger number of responding personnel is associated with a fire that causes more monetary damage. This tells us that the best nature of response to a fire may not necessarily be to have as many firefighters as possible.

### Limitations of Analysis

Despite efforts to build the best model possible, there are still some lingering issues in the model that could not be corrected. Notably, the final model has a rather large number of influential points. In the table in figure x, we can see that the final model has 335 influential points according to the DFFITS measure and at least 303 influential points if we take the lowest DFBETAS value. This impacts the usefulness of the final model as these influential points likely drag the regression line towards them and further away from the true response values. Unfortunately, this limitation cannot be corrected. The only possible way to correct the impact of these points is to remove them. However, this is unethical as there is no valid reason to remove them, other than to make the model look "better".

Another limitation of the analysis is that the transformation of the response by taking the fourth root makes it harder to interpret the model. The difficulty in interpretation results in our model being less useful to answer the question the study aims to answer. Unfortunately, this could not be corrected either as without this transformation, the assumption of normality was being violated. Violating the normality assumption would make any results meaningless. Therefore, a small amount of interpretability was sacrificed in order to correctly meet model assumptions.