Multiple Linear Regression Approach on Primary Factors in Toronto Fire Incidents Financial Loss

Meiwen Ding, Id 1005119863

December 6, 2020

Introduction

Toronto has the most population as a city in Canada. Because of the large population density in the downtown area, fire incidents can be extremely dangerous to the general public. This analysis paper will look into the factors of fire incidents happened in the GTA in 2019, and build relationship between the financial loss and multiple explanatory variables such as responding apparatus, casualties of firefighters, etc. The dataset we will be using is from the Toronto Open Data Portal. This will be a more in depth study of Problem Set 1.

The main approach for this analysis is to establish a multi-linear regression model based on the dataset. The method is considered an extension of a simple linear regression model, by having more than one explanatory variables for the dependent variable. In this case, we want to find out the primary factors that are associated with financial loss in fire accidents. In addition, the fitness of this model will be assessed by giving a look into the corresponding coefficients.

The only dataset used in this report is the fire incidents data.csv from the Toronto Open Data Portal. In the Method section, a specific description of the modelling will be presented. I will also summarize the data cleaning and simulation procedures. Results of the multi-linear regression will be displayed in the Results section, along with graphs that help the audience to comprehend. Finally, in the Discussion section, I will make summary and conclusion of the entire report, and discuss possible shortcomings and improvements to take in the future.

```
## -- Attaching packages ----------------------------------------------------
--------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts -------------------------------------------------------------
--------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   `_id` = col_double(),
##   Civilian_Casualties = col_double(),
##   Count_of_Persons_Rescued = col_double(),
##   Estimated_Dollar_Loss = col_double(),
##   Estimated_Number_Of_Persons_Displaced = col_double(),
##   Exposures = col_double(),
##   Ext_agent_app_or_defer_time = col_datetime(format = ""),
##   Fire_Under_Control_Time = col_datetime(format = ""),
##   Incident_Station_Area = col_double(),
##   Incident_Ward = col_double(),
##   Last_TFS_Unit_Clear_Time = col_datetime(format = ""),
##   Latitude = col_double(),
##   Longitude = col_double(),
##   Number_of_responding_apparatus = col_double(),
##   Number_of_responding_personnel = col_double(),
##   TFS_Alarm_Time = col_datetime(format = ""),
##   TFS_Arrival_Time = col_datetime(format = ""),
##   TFS_Firefighter_Casualties = col_double()
## )

## See spec(...) for full column specifications.
```

Abstract

This project aims to find predictors for Toronto fire incidents. We will explore how 1. number of responding apparatus, 2. number of responding personnel, and 3. TFS firefighter casualties, these predictors can predict estimated dollar loss.

Keywords

Multi linear regression, fire incidents, financial loss, number of responding apparatus, number of responding personnel, TFS firefighter casualties.

Methodology

The original dataset is obtained from the Toronto Open Data Portal. According to their website, “this dataset provides information similar to what is sent to the Ontario Fire Marshal relating to only fire Incidents to which Toronto Fire responds in more detail than the dataset including all incident types”. This analysis will pick the following variables for the MLR model: Estimated_Dollar_Loss, Number_of_responding_apparatus, Number_of_responding_personnel, and TFS_Firefighter_Casualties. Here is a table that shows every variable we will make analysis on, as mentioned before. The NA values have been cleared from the original dataset, and all of the variables are numerical.

```
## # A tibble: 251 x 4
##    Estimated_Dollar_~ Number_of_respondi~ Number_of_respondi~ TFS_Firefigh
```

```
ter_C~
##                  <dbl>               <dbl>               <dbl>
 <dbl>
##  1             500000                  28                  83
     0
##  2            1000000                  37                 113
     0
##  3             100000                  11                  37
     0
##  4               7500                  10                  33
     0
##  5            1000000                  33                  90
     0
##  6             250000                  17                  53
     0
##  7             500000                  18                  57
     0
##  8              50000                  23                  64
     0
##  9              20000                   6                  21
     0
## 10             100000                  12                  34
     0
## # ... with 241 more rows
```

We will use a MLR model with estimated dollar loss as the response variable and estimated number of persons displaced, number of responding apparatus, number of responding personnel, and TFS firefighter casualties as the explanatory variables. For firefighter casualties, there are only four categories: 1, 2, 3, 4. Thus, we will use dummy variable coding here for them as one explanatory variable with four categories. The beta coefficients measure the association between the predictor variable and the outcome. The MLR model we use:

$$EstimatedDollarLoss$$
$$= \widehat{\beta 0} + \widehat{\beta 1} apparatus + \widehat{\beta 2} personnel + \widehat{\beta 3} firefighter1 + \widehat{\beta 4} firefighter2$$
$$+ \widehat{\beta 5} firefighter3 + \widehat{\beta 6} firefighter4$$

Each coefficient represent on average, how much of an increase in the response variable is caused by one unit increase of the exploratory.

Results

```
##
## Call:
## lm(formula = Estimated_Dollar_Loss ~ Estimated_Number_Of_Persons_Displaced
 +
##     Number_of_responding_apparatus + Number_of_responding_personnel +
##     as.factor(TFS_Firefighter_Casualties), data = data_o)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max 
## -1115685  -123863   -12368    57190  3600294 
## 
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|
t|)
## (Intercept)                            -182894.60   43972.75  -4.159 4.43e
-05
## Estimated_Number_Of_Persons_Displaced       42.95      73.70   0.583   0.5
606
## Number_of_responding_apparatus           24387.72   20764.26   1.175   0.2
413
## Number_of_responding_personnel            -664.86    6613.97  -0.101   0.9
200
## as.factor(TFS_Firefighter_Casualties)1 -171807.79   78860.85  -2.179   0.0
303
## as.factor(TFS_Firefighter_Casualties)2   17197.07  202888.43   0.085   0.9
325
## as.factor(TFS_Firefighter_Casualties)3   18697.73  229007.23   0.082   0.9
350
## as.factor(TFS_Firefighter_Casualties)4 -182101.88  386203.00  -0.472   0.6
377
##                                           
## (Intercept)                            ***
## Estimated_Number_Of_Persons_Displaced     
## Number_of_responding_apparatus            
## Number_of_responding_personnel            
## as.factor(TFS_Firefighter_Casualties)1 *  
## as.factor(TFS_Firefighter_Casualties)2    
## as.factor(TFS_Firefighter_Casualties)3    
## as.factor(TFS_Firefighter_Casualties)4    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 381800 on 243 degrees of freedom
## Multiple R-squared:    0.4,  Adjusted R-squared:  0.3827 
## F-statistic: 23.14 on 7 and 243 DF,  p-value: < 2.2e-16
```

However, most of the explanatory variables have a significant p-value (<0.05), thus, changes in the estimated number of responding apparatus/personnels will not significantly affect estimated dollar loss in fire incidents. When the number of firefighter casualties is one, the result is significant and suggests a negative correlation between firefighter casualties and estimated dollar loss.

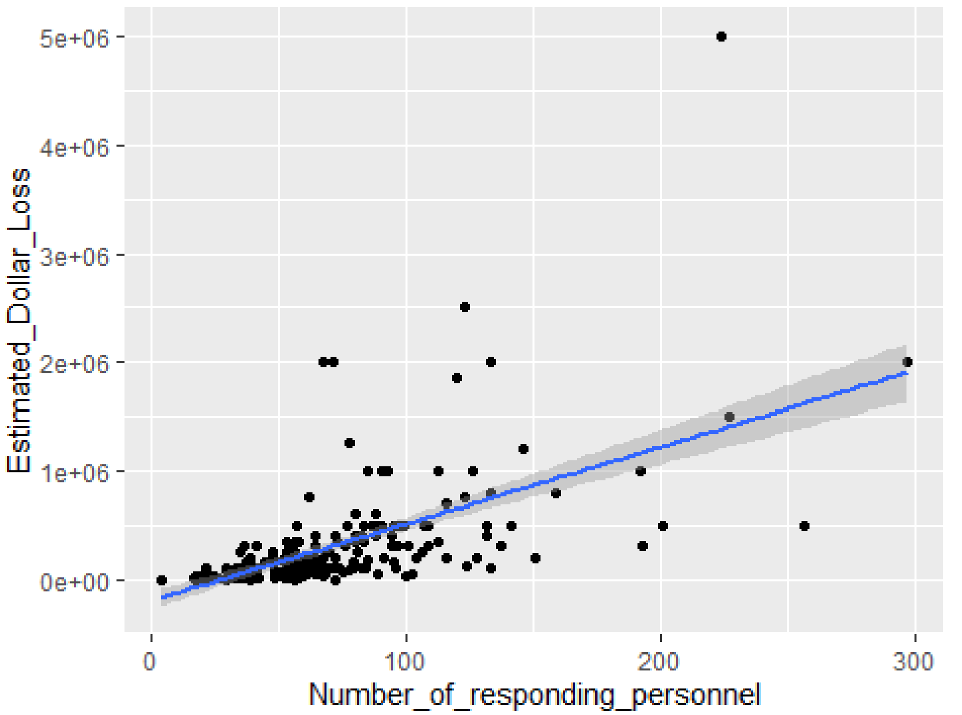
Discussion

In this analysis, we performed a MLR model to predict the explanatory variables for estimated dollar loss in Toronto fire incidents. We found that p-values of number of responding apparatus/personnels, category 2,3,4 of firefighter casualties are not significant enough to predict the estimated dollar loss in the fire incidents. When the casualty of firefighter is one, there is a negative correlation bwteen it and the estimated dollar loss.

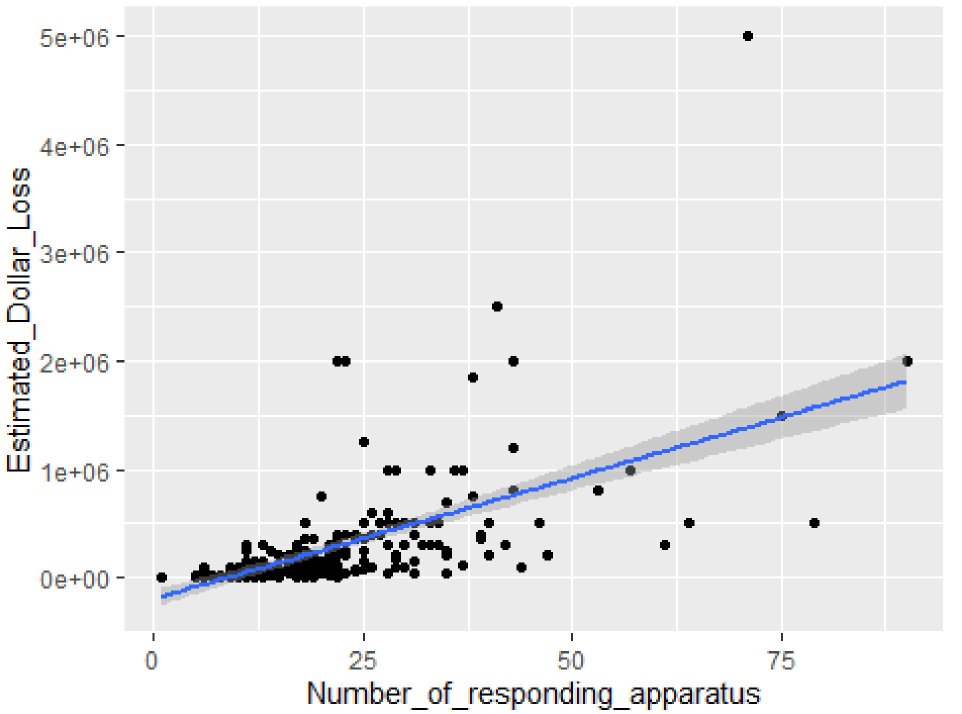
Here are three scatterplots to help interpret such insignificant results. We can tell that the plots for firefighter casualties do not show a linear pattern with estimated dollar loss. The rest of the two plots have weak positive associations as well. This might account for the reason why our MLR model does not have statistically significant results.

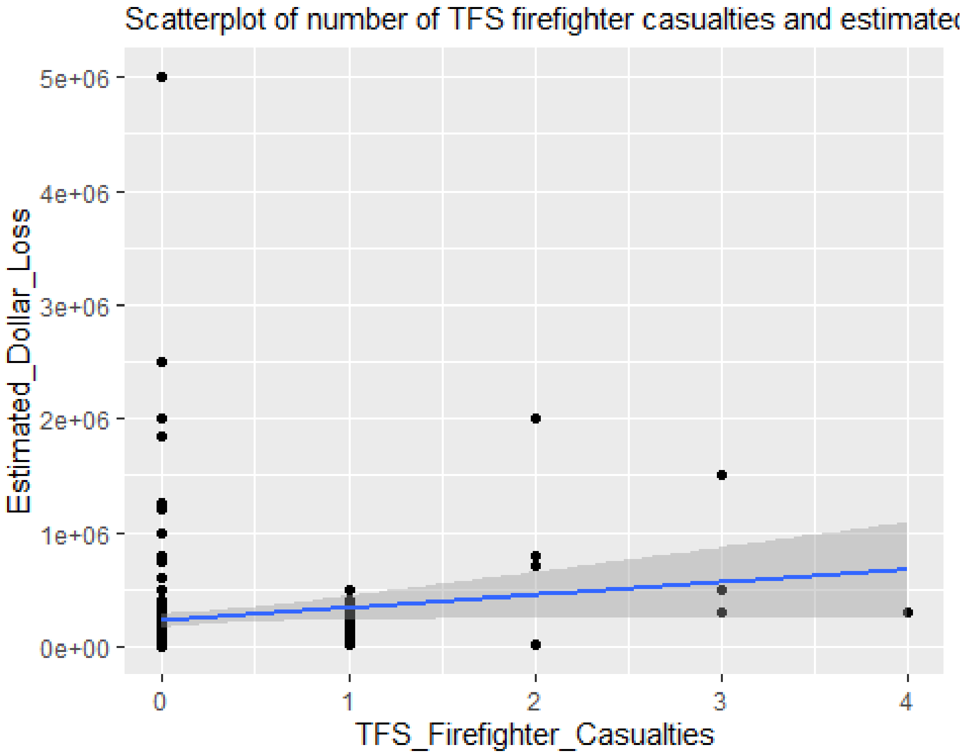
To improve the results, we could try removing some outliers, or using a logistic model for the discrete variables (count of people/casualties). This also gives insight to another possible analysis on this dataset, that is to predict factors that influence the civilian casulties in the fire incidents. This will create a logistic model since the variable of civilian casualties is discrete. Notice that a lot of variables in this dataset of are discrete values, but with text descriptions along side. Thus, if we want to simulate other models which work with the numerical and discrete variables, the next step would be to clean the dataset and remove all the texts.

Scatterplot of number of responding personnels and estimated c



Scatterplot of number of responding apparatus and estimated dc





References

Services, Fire. “Open Data Dataset.” City of Toronto Open Data Portal, open.toronto.ca/dataset/fire-incidents/.