Predicting Dollar Loss Brought by Fire Incidents Using Multiple Linear Regression Model

Lingjun Meng - 1005712579

10/17/2021

Intoduction

Research Question

I believe all of you who live in a highrise have been disturbed by the fire alarm either early in the morning or in the midnight. It seems that at 99% of the time, the alarms were false alarm caused by BBQ. And the rest 1% is where there was a fire that is 10 Levels away from your home.

I have been waken up in the midnight by the alarm 3 times last month, every time I got dressed and went out to find nothing happening. I started to wonder the significance of the fire alarm, sometime I hope I would just die in fire rather than hearing the alarm sounds.

That arouses my interest of investigating how bad the fire can be, how vast the losses caused by fire can be and what does the loss relate to. Here I define the Loss bought by fire the Dollar loss.(economical loss) In this report I will try to find out what is related to the Dollar Loss brought by fire and then find a way to predict the Dollar Loss brought by fire incidents.

This is a important research question as this can relates to every UofT students, and even every individual that lives in an apartment. Every time when the fire alarm sounds, There are only a few people going out to check what is happening. People can hopefully raise extra attention to the prevention of fire incidents by knowing how consequent the loss caused by fire can be after reading my report.

Note that my goal is not to create a model that can accurately predict the economical loss brought by fire incidents. Instead, a model that is easier to interpret will be more suitable as we want to include predictors that people are familiar with. Thus, people can easily use the predictors included in the model to make a prediction of the dollar losses and hence get a sense of how consequent the fire incidents can be. This claim is important to make as the results can be significantly different if we want to build a model that can make the most accurate predictions.

Goal: Build a linear regression model to predict dollar loss brought by fire that is easy to interpret.

Background Information

I searched the UofT library about similar topic and I did find a journal that is related to the question I am interested in. The journal reports the fire incidents in the U.S. and relating analysis including the Dollar loss by fire trend, Firefighter Casualties trend in 10 years.(Federal Emergency Management Agency et al., 1978)[4] In the journal I found a PowerPoint Slides summarizing the fire incidents from 2008 to 2017.

The Slides(Federal Emergency Management Agency, U.S. Fire Administration, 2019)[5] indicates that although the total fire incident case in U.S. has a decreasing trend for year 2008 - 2017, The deaths and fire dollar loss is still increasing. This can give us a impression that we do need more attention for the fire incidents.

This is related to my research question on predicting the Dollar loss brought by fire incidents. The report shows that it might relates to the deaths caused by the fire incidents since they are having the same trend.

The research question I will study is related to these report because the report provides many possible relationship between Dollar Loss and other variables including fire cause, and Time it takes to extinguish the fire(how long the fire last) etc. That really provided me some valuable thoughts on how to start my research question. i.e. Predicting the Dollar loss by the time it took to extinguish the fire.

In addition, There is also a journal(Tobin, W. A. et al., 2000)[6] studying a fire case at Dogwood Elementary School, Reston, Virginia. The fire caused a estimated dollar loss of 12 million while there was no casualties. The main reason of that was concluded to be a broken Fire Alarm System, which provided a chance for the fire to propagate and thus the time it took to extinguish the fire was extremely long.

From the data provided by National Fire Protection Association in the U.S., there is a fire incidents taking place in every 24 seconds across the States.[7] This once again reminds us how frequent the fire can be and showed the significance of this report.

Method

Variable Selection

As discussed above, The purpose of this report is to raise people's attention on fire prevention by letting people know how consequent a fire incidents be. And the goal is to build a linear regression model to predict the dollar loss brought by fire incidents that is more descriptive and easier to interpret. Thus the model will have to be simple and yet can still make good predictions on the outcome variable.

There are 43 variables available to be included in the model, the method of comparing all the possible subsets of the available variables would be extremely time-consuming and thus impractical. Therefore here I will first manually create the model using the predictors that were shown to be related to the outcome variable through background research. For example, the time that fire incidents last will be included in the model. And then the predictors that is shown to be significant in the full model will be added in the model, also remove predictors that has no or few impact on adjusted R squared. And then The result generated will be compared with the model created using automated selection method.

Stepwise Selection using BIC

BIC stands for Bayesian Information Criterion, which is a measure that tells us how well a model performs. This measure is chosen as it punishes harder on the complexity of the model and that fits our goal to make the model simple. Generally we the smaller the BIC the better.

Depending whether we will start with a full model(including all predictors) or a null model(with 0 predictors in the model), the Stepwise Selection method will work differently. If we choose to start with a full model with P predictors, then the first step will be a elimination step where the program will calculate the BIC values for all the possible models with *P-1* predictors. If the smallest BIC for the smaller model is smaller that the previous model, then the smaller model with smallest BIC value will be chosen to go to the next addition step. Adding a predictor is similar with elimination, BIC for all the possible models with one more predictor will be calculate and the smallest one will be selected. And then the program will iterate between addition and elimination until we can not add or remove anymore.

The reason why Stepwise Selection method is used instead of Forward or Backward is obvious, as Stepwise method takes the conditional nature of regression in to consideration. Thus, the algorithm will check and make sure to include the predictors that explains a significant portion of the variation regardless the changing conditional relationship between the response variable and the p predictors.

However, there are drawbacks of the automated selection method, please refer to limitations section in the Discussion part.

Model Validation

There are many reasons why we should validate the model we generated. For instance, we have to make sure the model created does not only work well with the original data that was used to create the model, the model also should perform well on unseen data. Especially when we are trying to get a model that can make the most accurate prediction, we can easily overfit the data. That means even the model can make good prediction on the data used to generate it, there could be a possibility that the model will make bad predictions on a new dataset. Thus, in order to prevent that from happening, we apply the method of model validation.

The most commonly used method of Model Validation is to randomly dividing up the original dataset into training dataset and test dataset. There are many possible training/test split method available, we just have to make sure there is enough observation in each of the dataset after split. In our case, we have 17536 observation, therefore I will split the data by 50/50, meaning that half of the data from the original dataset will be randomly selected to be included in the training dataset and the rest goes to the test dataset.

After training/test split, the model will first be generated using the training data. Which is followed by creating a new model using the test dataset. Note that we have to use the same method when generating a new model, all that changes is the dataset used. Next, the two models generated will be compared and this is the process of validation. If the two models created have the same significant predictors included in the model and the estimates of the coefficients, adjusted R squared are similar, then we will conclude that the model is validated.

In addition, we also have to make sure we do not have new or worse model violations in the model generated using test dataset. Please see next part for more detailed information about model violations.

Model Violations and Diagnostics

Assumptions of Linear Regression Model

Linearity of the relationship

The relationship between the response variable and the predictors should be linear, this ensures we are fitting the right model. Since we are fitting a Linear Model, it will be inappropriate if the true relationship between the outcome and the predictor is non linear. Thus we will have to check the relationship by observing a linear pattern in the scatter plot.

If the pattern we observe from the scatter plot is not linear, then we we have to make modifications to correct the violation. Normally we just have to make a log transformation on the response variable to fix the non-linear violation.

Independence, the constant variance and the normality of the error

We have to make sure the responses(or equivalently, error) are independent with each other, so that we do not work with fewer amount of data than we have expected. The constant variance of error is also important to guarantee the model perform the same for different values of predictors. The normality of the error ensures the application of the properties of normal distribution. Therefore the P-values we get are meaningful.

Checking these three assumptions can be done by observing the residual plots. If there are clusters, then the independence might be violated. If there is a fanning pattern where the spread keeps increasing, then the constant variance might be violated. If the histogram of the errors are not a bell shaped curve or there are many points landing away from the diagonal in the QQ-plot, then the normality might be violated.

Correcting these violations will require the method of boxCox Transformation. This is a program that will calculate the power that should be applied to each variable to correct the violations.

Multicollinearity

We also have to check if more than two predictors are correlated with each other, This can be done using a VIF function in R where we want a value that is smaller than 5. If we indeed find some predictors' VIF value

is bigger than 5, these predicts will be removed from the model. We then apply VIF function again to make sure we fixed the problem.

Influential Observations

We will first identify the problematic observations using methods including Cook's distance and DFFITS and then assess the impact of the influential points to the model. If the observation are not mistakenly collected, then the influence of these observations will be discussed in the Limitation section.

Result

Data

Data Source

The data I collected comes from Open Toronto Data Portal. The name of the data is "Fire Incidents", which includes 17536 records of fire incidents in Toronto. This data set comes from a larger data set containing all the fire incidents in Toronto and it only keeps the cases that with enough information in it.(see bibliography 8 for more details)[8]

The data set have 43 variables that is related to the recorded fire incidents, for example, the location, The time when TFS(Toronto Fire Service) receives Alarm, The time when the fire is under control, the estimated Dollar loss etc.

Data Cleaning summary

One new variable named Time Took Extinguish is created using the time TFS received Alarm as well as the time when the fire is under control. Here I define the time that fire lasts(time took to extinguish) to be the difference between the time when the fire is under control and the time when the TFS received the Alarm.

Then all the variables that are locations; names; ids were removed. The rest of the variables were then put together into the cleaned dataset and NAs were removed. After cleaning, the data set have 11214 observations with 10 numerical variables. Note that only numerical variables were included in the cleaned data set as there are at least 10 different categories for all the categorical variables. Our goal is to make the model simple and easy to interpret, thus all the categorical variables were not included in the cleaned dataset. However, there are drawbacks of doing this, please refer to Limitations section for more information.

Data Summaries

Below is a table summarizing all the variables split in to training and test data set:

Table 1: Summary statistics in the training and test dataset, each of size 5607.

Variables	mean (s.d.) in training	mean (s.d.) in test
Estimated_Dollar_Loss	4.9997683×10^4 (7.2596897×10^5)	3.58362×10^4 (2.0756898×10^5)
Time_took_Extinguish	1163.032 (2744.42)	1190.816 (3168.504)
Civilian_Casualties	0.114 (0.49)	0.111 (0.421)
Count_of_Persons_Rescued	0.055 (0.492)	0.07 (1.269)
Persons_Displaced	17.948 (121.541)	16.602 (116.557)
Latitude	43.704 (0.051)	43.706 (0.051)
Longitude	-79.403 (0.1)	-79.403 (0.102)
Number_of_responding_apparatus	9.203 (7.831)	9.239 (9.254)
Number_of_responding_personnel	30.16 (24.063)	30.254 (28.026)
TFS_Firefighter_Casualties	0.023 (0.187)	0.019 (0.167)

Note that all the variables have similar mean and standard deviation except the response variable: Estimated Dollar Loss. This is caused by extremely large values in Estimated Dollar Loss variable. However, the observation with that value can not be removed as all the observations were accurately collected. This can be the reason why the model generated using the Training dataset can not be validated using Test dataset.

Process of Obtaining the Final Model

Manual Method

First of all, a model using all the available predictors will be generated to find significant predictor. Also, as stated above, the predictor Time_Took_Extinguish is shown to be related the response variable from background research and will be included in the model.

The predictors that are significant in the full model are Time_Took_Extinguish, Number_of_responding_apparatus, Number_of_responding_personnel and TFS_Firefighter_Casualties. This is easy to understand as the longer the fire last, the more the loss can be. Also, the more responding apparatus, the bigger the scale of the fire can be and hence the loss can be bigger. The new model with only four predictors has even higher adjusted R squared, thus only these four predictors will be included in the model.

Checking Assumptions

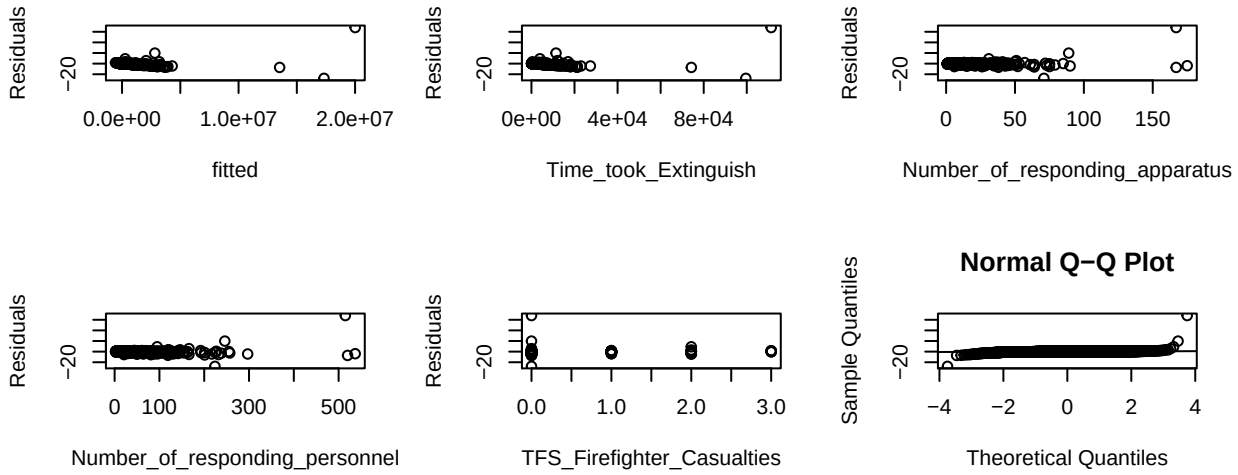

Figure 1: Residual and Q-Q Plots of the initial model

Form the residual plots and Q-Q Plot, power transformation is required to correct non-normality. However, the transformation required is too complicated so that it will no longer be easy to interpret the model. The violation here is not that severe, and note that our goal is to make the model as simple as possible. Therefore no powerTransformation is done. However the impact of the influential points in not negligible.

Mullticolinearity

Form the calculation we know only Number_of_responding_apparatus and Number_of_responding_personnel have very high VIF values. This means that these two variables are highly correlated. Therefore one of them have to be removed. The VIF of the new model is all below 5, and we don't have multicollinearity problem.

Automated Selection Method.

Here the model is generated using Stepwise Selection method using BIC penalty measure. the model that the function thinks to be the best consists of Time_took_Extinguish, Number_of_responding_apparatus,

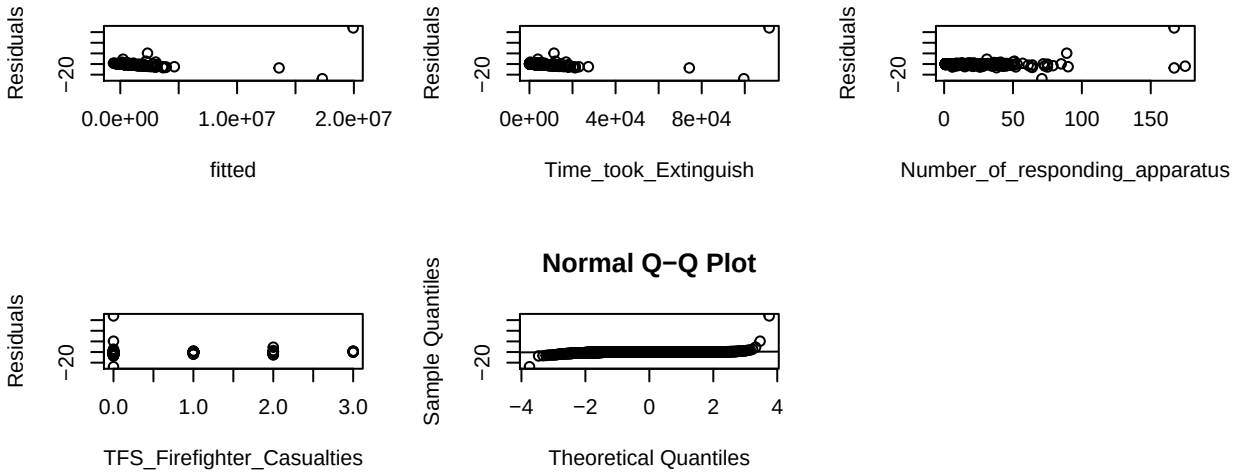

Figure 2: Residual and Q-Q Plots of the Fianl Training model

Number_of_responding_personnel as well as TFS_Firefighter_Casualties.

The model generated using automated selection method is exactly the same with the model that I manually created before the adjustment made to correct multicollinearity. Since same model was approached, we do not have to check assumptions and multicollinearity again. Thus, from training dataset, the "best" model can be generated includes predictors: Time_took_Extinguish, Number_of_responding_apparatus and TFS_Firefighter_Casualties.

Model Validation

Using the same approach as above to generate a new model using test dataset.

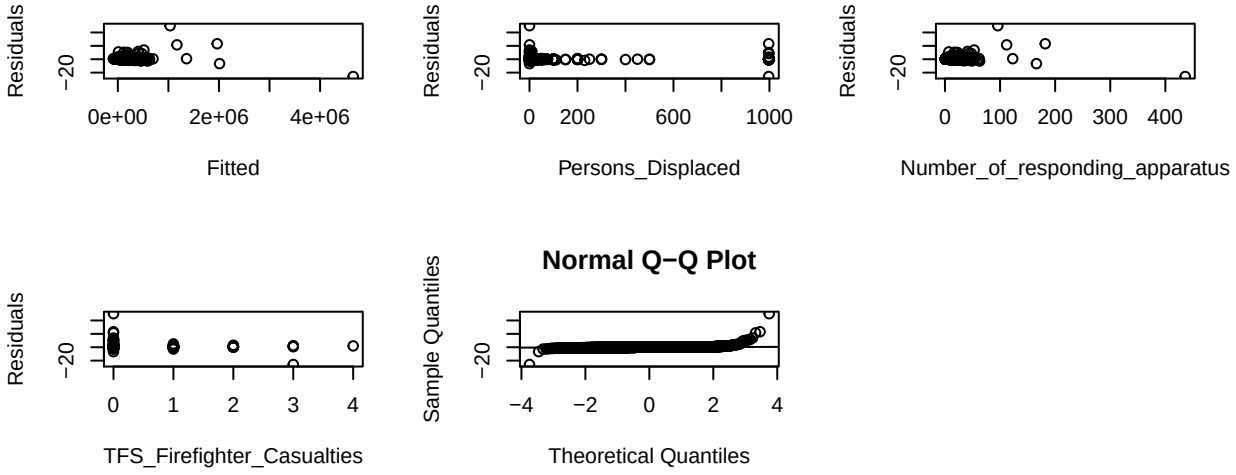

Figure 3: Residual and Q-Q Plots of the test model

Here the model generated using Stepwise Selection method have different predictor with the one created using Training dataset. Time_took_Extinguish is substituted by the new predictor Persons_Displaced. The new adjusted R squared is much smaller while there are not new or worse violations of assumption

If the same predictors are used to create a model using test dataset, Time_took_Extinguish predictor's P-value and its coefficient will be significantly different from the numbers in the Training model. However, if we use a confidence level of 0.05, this predictor is still significant.

Therefore the model is no validated, this is because the influential observations in the training dataset. Among 5607 observations in the dataset, there is a Dollar Loss of 50 million where the mean is only 1163. This is not saying the model is wrong, it just shows that the model generated is not that accurate when we are trying to make predictions.

Final Model

$$Y = 171X_1 + 6549X_2 - 219400X_3 - 204100 + \epsilon$$

Here Y refers to the response variable, the Estimated Dollar Loss;

X_1 refers to the Time Fire Last predictor;

X_2 refers to the Number of responding apparatus predictor;

X_3 refers to the Number of Firefighter Casualties predictor;

ϵ refers to the error(residual) factor.

Discussion

Final Model Interpretation

Table 2: Summary of the predictors in the final model.

Predictor	Estimated Coefficient	significance
Time Fire Last in seconds(s)	171	***
Number of responding apparatus	6549	***
Number of Firefighter Casualties	-204100	***

From the model we can see all the predictors are extremely significant. Assuming all other variables stays the same, with every second the fire last, the economical loss is expected to grow 171 dollars; with one more apparatus responding to a fire, the economical loss is expected to be 6549 dollars more; with one more firefighter dying from the fire, the loss is expected to be 204 thousand less.

This intuitively and clearly showed how consequent the fire can be considering the economical loss brought by fire, which is consistent with the goal. With the fire going for every second, you are giving away 171 dollars, and one firefighter's sacrifice only worth about 200 thousand dollar. Therefore everyone should raise extra attention to prevent fire from happening.

Limitations

1. The original dataset have 17536 observations, after cleaning, there are only 11214 observations available left. The EDA shows there are differences in the distribution of the variables before and after cleaning. Cleaned dataset might not be representative for the whole population.
2. The original dataset is only a subset of all the fire incidents in Toronto where observations have fewest missing values. Again this make the data not representative enough as there might be sampling bias.

3. The Linear Model is generated using only the fire incidents in Toronto, the model might not alsoperform well at locations outside Toronto.

4. The Final model is supported by the model generated using automated Selection Method using BIC penalty Measure. However not all the possible models have been assessed and thus there could be a possibility that the "best" Model is different with the final model stated above.

5. In order to make the model as simple as possible, all the categorical variables in the dataset were not included in the model. This should be avoided if you are trying to generate a model that can make the most accurate prediction. For next steps, people interested in this topic can clean the data in a manner where categorical variables are transformed to numerical, where the number can be the different classes. This can make further research easier.

6. There are still violations of assumptions of the regression. The normality of the errors should be corrected. However the powertransformation can make the model extremely difficult to interpret, which contradicts my goal. Therefore no transformation was done in the final model. However, people that are interested in generating a accurate model should correct all the violations.

7. The impact of the influential points on the model is huge. without these observations, the significance of the predictor will be different. This is also the reason why the model failed to be validated. Again the goal is to generate a model that is simple and interpretable, the exact influence is not discussed. People focusing on a more precise model should investigate the model with and without the influential observations.

Bibliography

1. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: January 15, 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: January 15, 2021)
4. Fire in the United States (Online). (1978). Federal Emergency Management Agency, U.S. Fire Administration, National Fire Data Center.
5. Fire in the United States 2008-2017 20th Edition (PowerPoint Slides). (2019). Federal Emergency Management Agency, U.S. Fire Administration. Retrieved from: https://permanent.fdlp.gov/LPS16395/20th_edition_fius20th.pdf
6. Tobin, W. A., Stambaugh, H., & Roberson, J. L. (2000). $12 million dollar fire at Dogwood Elementary School, Reston, Virginia. Federal Emergency Management Agency, U.S. Fire Administration.
7. Nfpa.org. 2021. NFPA - Reporter's Guide: The consequences of fire. [online] Available at: https://www.nfpa.org/News-and-Research/Publications-and-media/Press-Room/Reporters-Guide-to-Fire-and-NFPA/Consequences-of-fire [Accessed 13 December 2021].
8. Ku, K., 2021. Open Data Data set. [online] Open.Toronto.ca. Retrieved from: https://open.toronto.ca/dataset/fire-incidents/.
9. Sharla Gelfand. opendatatoronto: Access the City of Toronto Open Data Portal. R package version 0.1.4
10. Yihui Xie.(2021). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.34
11. John Fox et al. (2021). car: Companion to Applied Regression. R package version 3.0-12
12. Brian Ripley et al. (2021). MASS: Support Functions and Datasets for Venables and Ripley's MASS. R package version 7.3-54