Is Fire Alarm System Really Helpful?

Lingjun Meng

10/17/2021

Introduction

Is the Fire Alarm System in your home working properly? What was the last time you check it? There are fire happening nearly everyday in Toronto, especially as winter comes, The whether gets dryer and it could be easier for fire to take place.[1] No matter you are living in a house in the suburb or an apartment in downtown, you have to be more careful to prevent fire from happening.

Some people may have different ideas on the Fire Alarm System, some people might think it does not matter that much as long as he/she can run fast once fire was spotted. In order to make people to pay more attention on fire prevention, I will demonstrate the consequences that a fire can bring to us with or without a properly functioning Fire Alarm System. In other words, I will find out if the loss caused by fire is related to a working Fire Alarm System.

The data I will be using in this report comes from The Toronto Open Data Portal. There is a data set called Fire Incidents, the data set includes detailed information like Fire incidents type, ignition source, time TFS(Toronto Fire Service) was notified, etc.

There are 17536 cases included in this data set and I will be using information in the data set to fit a linear regression model in order to determine if there are relationships between the loss caused by the fire and the status of the Fire Alarm System. Other variables of interests will also be included, for example the time of the fire and location of the fire.

To make it more clear, a linear regression model is to find the possible relationship between variables fitted using given data set. In this case, the model will give me a coefficient regarding each variable of interests(i.e. the status of the fire alarm system) . If the coefficients are very small with a p-value calculated to be greater than 0.05, then we can conclude that there is no significant relation between the variables. (i.e.It does not matter if we have a properly functioning alarm or not regarding the loss)

Hypothesis

There is statistically significant relationship between fire alarm system, location, time and the loss brought by fire.

Data

Data Collection Process

I loaded the whole "Fire Incidents" data set directly to R from Open Toronto data portal.[2] to make further modifications, doing that is easy. There are many approaches you can do this. The easiest approach is to follow the instructions provided by the portal Fire Incidents.[3]

Note that it is okay if you are not using R as your programming language, the portal also provided the instructions for other programming languages including Python and Node.js. The data is also downloadable as CSV, JSON, and XML file if you want to get access to the raw data.

Data Summary

The raw data sets consists of 17536 observations(fire incident cases) and 43 variable including area of fire, address, ward code and other detailed description of the fire incidents including causalities, cause of fire, losses etc.

Cleaning process

Although there are 17536 observations in total, some of them has to be expelled because it has missing values(i.e. NA) in the variables that we are interested in.

So I started cleaning the data by removing all the missing values in our variables of interests: Estimated Dollar Loss, Latitude, Time the TFS receives the alarm and the status of fire alarm system.

Then I worked on modifying the variables to make it easier for me to fit the linear regression model. The Estimated Dollar Loss and Latitude variables can be used directly, However, the time TFS receives the alarm can be hard to exploit without any modification as well as the status of the fire alarm system variable.

I decided to make the fire time to be a categorical variable, that can be achieved by extracting the numerical information from the TFS_Alarm_Time variable, and create a new variable named "Fire_Time", the value of the fire time will be "day" if the hour information extracted from the "TFS_Alarm_Time" variable is within 6 to 18; and the value will be "night" otherwise.

I modified status of fire alarm system also by creating a new variable named "Alarm_Status" by checking the "Fire_Alarm_System_presence" variable in the raw data. if the value of the variable is equal to: "1 - Fire alarm system present", the value of the "Alarm_Status" variable will be "Yes", and will be "NO" otherwise

In addition, I noticed that there are some extremely large numbers in Estimated Loss variable in the raw data set. I choose to remove all the values that is greater than 5 million to make it easier for us to find the relationship between variables.

After every variable of interests are modified, I only chose to keep the variables of my interests and store it to a new data set named "mydata".

Description of Important variables

Now we can have a look of the final data set we will be using, there are 11214 cases and four variables in it. Below is a glimpse of the cleaned data set:

Table 1: First eight rows of data

Estimated_Dollar_Loss	Latitude	Fire_Time	Alarm_Status
0	43.70866	Night	No
2000	43.71481	Night	No
100000	43.74858	Day	Yes
5000	43.65215	Night	Yes
500	43.65681	Night	Yes
0	43.69711	Day	No
15000	43.71171	Night	No
0	43.76780	Night	Yes

the four variables are as follows:

The estimated dollar loss is the estimated loss brought by fire; The latitude is the latitude of the fire where it took place(location); The Fire time indicates whether the fire took place at day or night; And lastly, the Alarm status variable indicates whether the place where the fire took place have a properly functioning fire alarm system.

It is shown below the summary of the estimated loss brought by fire in dollar:

Variable	Mean	Standard Deviation
Estimated loss	3.5022801×10^4	1.6323028×10^5

Below is the distribution of the locations(represented by latitude) of the fire incidents.

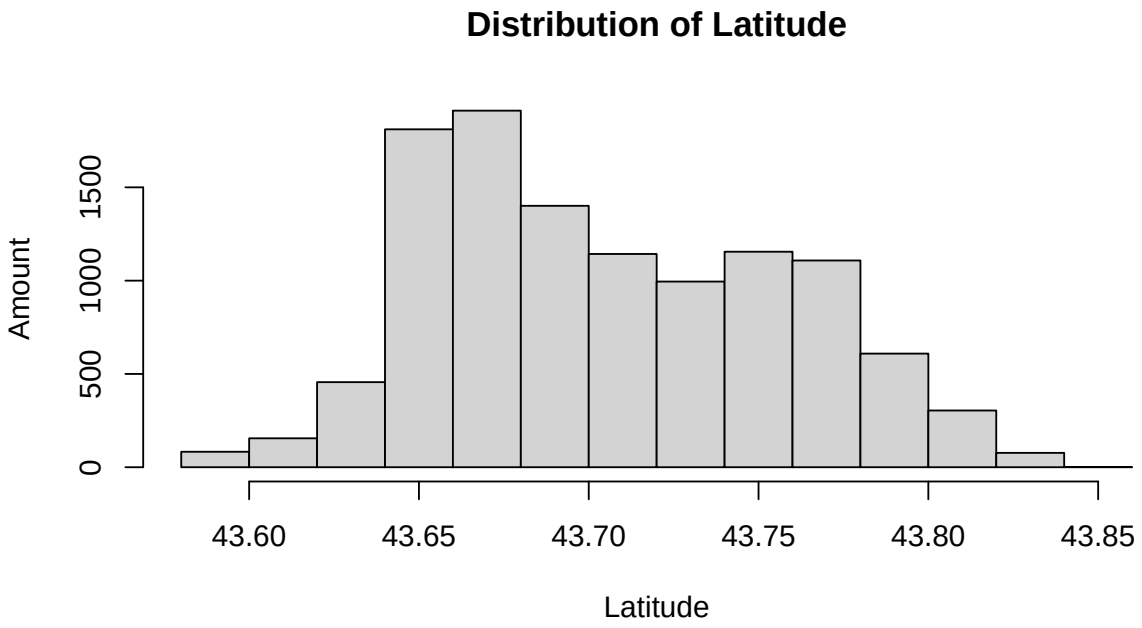

Figure 1: Distuibution of Latitude of the Fire Incidents

The rest two variables are categorical variables indicating the status of the fire alarm system(Yes or No) as well as the happening time(Day or Night) of the fire incidents.

The visualization of the data is as follows:

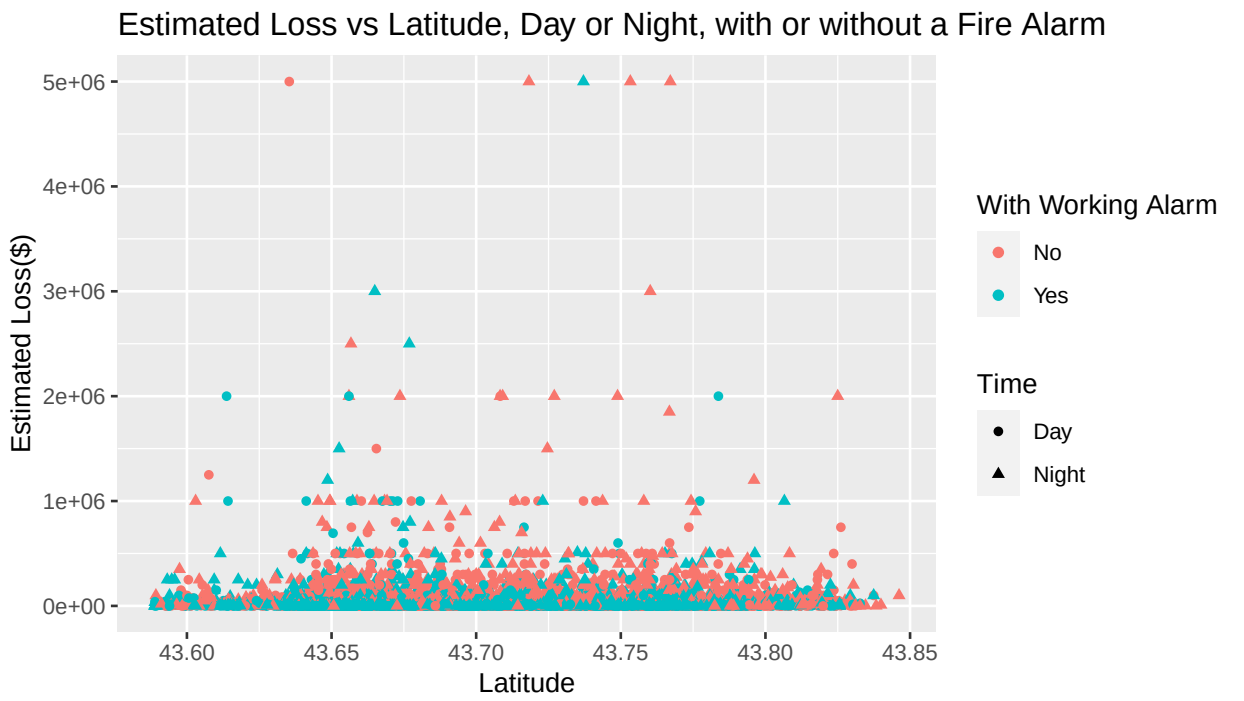

Figure 2: Estimated Loss vs Latitude, Time, Fire Alarm System

From the Visualization generated above, those with out a working fire alarm(dots in red) seem to have higher loss than those who have a properly functioning one(dots in blue).

In addition, fire incidents took place at night(triangle dots) seem to have a higher loss than that took place at Day(round dots).

The relationship between Latitude(location) and the loss seems to be casual i.e. no relationship could be identified from the graph.

In order to determine if the observations above are correct, I will carry out a linear regression model.

All analysis for this report was programmed using `R version 4.0.5`.

Methods

The model I will be using is frequentest linear regression model. I will fit the model with the cleaned data with only the variables of our interests. The liner regression model will provide us with the possible relationships between variables fitted in the model. if there are only one dependent variable and one independent variable, the model will be simple linear regression model. Here we have 1 dependent variable: estimated dollar loss and 3 independent variable: Latitude, Time of fire incidents and the status of fire alarm system. Thus the model we will be fitting is a multiple linear regression model.

The reason why I am using linear regression model is first off all, this method is easy to carry out and we have a data set of big enough size(more than 10,000). the model will generate a function, predicting the dependent variable using independent variables. A single coefficient will be provided for each independent variable(predictor). In our example the model will provide us a function in the form of:

estimated dollar loss = a x Latitude + b x fire time + c x Alarm status + d + error

where a, b, c are coefficients and d is the intercept

a, for example, means if we keep all other variables constant(stay same), if Latitude goes up by 1, the expected change in estimated dollar loss is a. And it is similar for that of b and c, however there are some significant differences of interpreting the outcomes between b,c and a since b,c are categorical variable and a is numerical. Please see Result part for more detailed information.

if we set all a,b,c to 0, we will get estimated dollar loss = d.

Beside the above information, the linear regression model can also provide us the P-value of the coefficients. P-value is a useful tool for us to determine if we are getting statistically significant results or just random outcomes. The P-value is calculated by examining the probability of a result as extreme as our result assuming the null hypothesis. This is also a reason why I am choosing to do linear regression, the P-value provided can be used directly as a model selection method.(determine if the model is good)

For example, here we are summing there is no relationship between these variables(null hypothesis), that is assuming the coefficients to be 0. and the P-value will tell you how extreme your calculated coefficients are under the above assumption. Normally we will use a significance level of 0.05. If the P-value is smaller than 0.05, that is, the probability of getting a result as extreme as ours is smaller than 0.05. Then we can conclude that there is actually relationship between variables.(reject the null)

Multiple Linear Regression model

The multiple linear regression model we are using is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

Note that here: y represents the dependent numerical variable estimated dollar loss; β_0 represents the intercept of the regression line; β_1 represents the coefficient for the first independent variable: Latitude; β_2 represents the coefficient for the second independent variable: Alarm status; β_3 represents the coefficient for the third independent variable: Fire incidents Time; ϵ represents the residual. x1,x2,x3 represent Latitude, Alarm status and fire time respectively.

When we are fitting the linear regression model with these four variables, we are assuming the relationship between these variables are linear. Also we are assuming that no other factor can affect our dependent variable: estimated dollar loss except the 3 independent variables.

Results

Below is a summary for the linear regression model fitted.

Results	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
Estimate	-1508583	35718	-27999	-3013
P-value	0.251	0.235	<2*10^(-16)	0.348

As I mentioned above, β_0,β_1,β_2,β_3 are related to intercepts, Latitude, Alarm Status and Fire Time respectively. The Estimated row in the table is the coefficients, for those variables while the next row is their corresponding P-values.

For β_0 = -1508583, it means if all the variables are 0, then the estimated loss will be -1508583.

For β_1 = 35715, it means if all other variables stay the same and Latitude goes up by 1, then the estimated loss will be expected to go up by 35718.

For β_2 = -27999, it means if all other variables stay the same, those who have a properly functioning Fire Alarm System will be expected to have a estimated loss that is 27999 smaller than that who does not have one.(fire alarm system)

For β_3 = -3013, it means if all other variables stay the same, those fire incidents took place at night are expected to have a estimated loss that is 3013 smaller than that took place during day.

For example, if one day there is a fire took place at Latitude = 43.7 at night where there was a properly functioning Fire Alarm System, Then the Estimated Dollar Loss is predicted to be $\beta_0 + \beta_1$* 43.7 + β_2 +β_3. and the predicted value is 21281.6

However, not every coefficient we got are statistically significant. on a significance level of 0.05, we are saying the coefficients with a P-value bigger than 0.05 are not statistically significant. And if the P-value generated is smaller than 0.05, we say that the percentage of getting a result as extreme as ours assuming the null hypothesis is lower than 0.05 and thus the coefficient generated will be statistically significant.

Here in our case, only the coefficient for Fire Alarm status is significant and the P-value calculated is extremely small(smaller than 2*10^(-16)), which is nearly 0. That indicates the two variables: estimated loss and alarm status are strongly related. On the other hand, the rest of the two variables: Latitude and Fire time do not have strong relationships between the variable estimated loss.

This result is not surprising, it is intuitive that having a working fire alarm system will decrease our loss brought by fire because it can notify us the moment it detects a fire and thus saving our valuable time to either get the fire extinguished before expanding everywhere or evacuate.[4]

All analysis for this report was programmed using `R version 4.0.5`. I used the `lm()` function in base `R` to derive the estimates of a frquentist linear regression in this section [5].

Conclusions

We were trying to find if Fire Alarm System is really helpful considering the Estimated Dollar Loss brought by the fire incidents.

And we want to find if there are also relationships between the loss and the location of the incidents(represented by Latitude) as well as the time of the incidents.(Day or Night) and we decided to figure it out using frequentest linear regression model.

From the Linear regression model fitted using the fire incidents data set, we found that there is no significant relationship between estimated loss and Location(Latitude) and fire time.(Day or Night) On the other hand, the loss is strongly related to the status of the Fire Alarm System. To be more clear, we found that with every other factors staying the same, having a properly functioning Fire Alarm System will on average have a Loss that is 27999 smaller than that who does not. Note that the mean value of the Estimated Dollar loss is only3.5022801×10^4

In conclusion, considering the estimated dollar loss brought by the fire incidents, the fire alarm system is really helpful. That is, having a correctly working fire alarm system is expected to significantly reduce the loss brought by fire.

Weaknesses

1.There might be a sampling bias since I removed all the fire incidents information with missing values for Estimated Dollar loss, Latitude, TFS_Alarm_Time and Fire_ Alarm_System_Presence variables from the raw data set. And that might be potentially related to our responses of interests Estimated Dollar loss. For example there were missing values in the estimated dollar loss variables because the loss were too big to estimate or there were civilian casualties due to the fire incidents and thus the loss could not be estimated. This can also be considered as Survival Bias.

2.I removed all the Estimated Dollar Loss that is bigger than 5 million considering them to be Outliers. removing them might also cause a sampling bias.

3.The data set I was using is not the complete data set of all fire incidents in Toronto, instead, it was derived from a Larger data set containing all the incidents. And it chose to keep only the cases with more information was recorded. The result generated using different data set might differ.

4.Be careful when trying to use the model I generated above to make predictions on Estimated Dollar Loss. The coefficient of determination(R^2) was calculated to be 0.07, which means the portion of variation of the estimated dollar loss explained by the model is only 0.07.[6] There should be other predictors added into the model.

5.In the model I used the Latitude variable to represent the location of the fire incidents and found that the loss is not (significantly) related to Latitude. However, there are other variables in the data set that can represent locations. For example ward number, Longitude etc. using different location representations might have a different result.

Next Steps

The model fitted should be improved. I did not fit a regression line on the distribution of the data as there is one categorical variable(Fire Alarm Status) significantly related to the response: Estimated Dollar Loss.

One key predictor that I think could be considered is the time it takes to extinguish the fire. This information was provided indirectly in the raw data set by giving the time alarm sounded and the time the fire was extinguished.

The predictor can intuitively be the difference of these two variables. However, I faced some difficulties when trying to do this. The time information provided were all of type "strings" and we have to convert it to type "numeric" before we can make a subtraction.

One possible way is converting the time all into seconds and then make the difference. However, I failed to find a way to achieve this and thus I made the fire time categorical Day or Night instead.

In addition, since there were only 1 predictor(Alarm Status) that is significantly related to the estimated dollar loss variable in the model, A partial F test can be done to determine if we can remove the rest two predictors.(Latitude and Fire time) This is similar that the P-Value selection process but this method can make more analysis of the co-variance between the variables.[6]

Discussion

To wrap up, do make sure to check your Fire Alarm System on a regular time basis. A properly functioning alarm is proved to significantly saves your economical loss.

Bibliography

1.Nast, C., 2021. House Fires Are Way More Common in Winter—7 Tips to Stay Safe. [online] SELF. Available at: https://www.self.com/story/house-fire-prevention [Accessed 14 October 2021].

2.Open.Toronto.ca. 2021. Open Toronto Portal. [online] Available at: https://open.toronto.ca/ [Accessed 14 October 2021].

3.Ku, K., 2021. Open Data Data set. [online] Open.Toronto.ca. Available at: https://open.toronto.ca/dataset/fire-incidents/ [Accessed 14 October 2021].

4.Dubner, S., 2021. How Many Lives Do Smoke Alarms Really Save? - Freakonomics. [online] Freakonomics. Available at: https://freakonomics.com/2012/02/06/how-many-lives-do-smoke-alarms-really-save/ [Accessed 17 October 2021].

5.Peter Dalgaard. (2008) *Introductory Statistics with R, 2nd edition.*

6.Katherine Daignault. (2021) *Methods of Data Analysis 1* [PowerPoint Slides].(Asynchronous Slides of STA 302)

7.Yihui Xie.(2021). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.34

Appendix

Here is a glimpse of the Raw data set.

```
## Rows: 17,536
## Columns: 43
## $ `_id`                                                          <int> 2122289,~
## $ Area_of_Origin                                                 <chr> "81 - En~
## $ Building_Status                                                <chr> NA, NA, ~
## $ Business_Impact                                                <chr> NA, NA, ~
## $ Civilian_Casualties                                            <int> 0, 0, 0,~
## $ Count_of_Persons_Rescued                                       <int> 0, 0, 0,~
## $ Estimated_Dollar_Loss                                          <int> 15000, 5~
## $ Estimated_Number_Of_Persons_Displaced                          <chr> NA, NA, ~
## $ Exposures                                                      <int> NA, NA, ~
## $ Ext_agent_app_or_defer_time                                    <chr> "2018-02~
## $ Extent_Of_Fire                                                 <chr> NA, NA, ~
## $ Final_Incident_Type                                            <chr> "01 - Fi~
## $ Fire_Alarm_System_Impact_on_Evacuation                         <chr> NA, NA, ~
## $ Fire_Alarm_System_Operation                                    <chr> NA, NA, ~
## $ Fire_Alarm_System_Presence                                     <chr> NA, NA, ~
## $ Fire_Under_Control_Time                                        <chr> "2018-02~
## $ Ignition_Source                                                <chr> "999 - U~
## $ Incident_Number                                                <chr> "F180209~
## $ Incident_Station_Area                                          <chr> "441", "~
## $ Incident_Ward                                                  <int> 1, 18, 2~
## $ Initial_CAD_Event_Type                                         <chr> "Vehicle~
## $ Intersection                                                   <chr> "Dixon R~
## $ Last_TFS_Unit_Clear_Time                                       <chr> "2018-02~
## $ Latitude                                                       <dbl> 43.68656~
## $ Level_Of_Origin                                                <chr> NA, NA, ~
## $ Longitude                                                      <dbl> -79.5994~
## $ Material_First_Ignited                                         <chr> "47 - Ve~
## $ Method_Of_Fire_Control                                         <chr> "1 - Ext~
## $ Number_of_responding_apparatus                                 <int> 1, 1, 6,~
## $ Number_of_responding_personnel                                 <int> 4, 4, 22~
## $ Possible_Cause                                                 <chr> "99 - Un~
## $ Property_Use                                                   <chr> "896 - S~
## $ Smoke_Alarm_at_Fire_Origin                                     <chr> NA, NA, ~
## $ Smoke_Alarm_at_Fire_Origin_Alarm_Failure                       <chr> NA, NA, ~
## $ Smoke_Alarm_at_Fire_Origin_Alarm_Type                          <chr> NA, NA, ~
## $ Smoke_Alarm_Impact_on_Persons_Evacuating_Impact_on_Evacuation <chr> NA, NA, ~
## $ Smoke_Spread                                                   <chr> NA, NA, ~
## $ Sprinkler_System_Operation                                     <chr> NA, NA, ~
## $ Sprinkler_System_Presence                                      <chr> NA, NA, ~
## $ Status_of_Fire_On_Arrival                                      <chr> "7 - Ful~
## $ TFS_Alarm_Time                                                 <chr> "2018-02~
## $ TFS_Arrival_Time                                               <chr> "2018-02~
## $ TFS_Firefighter_Casualties                                     <int> 0, 0, 0,~
```