Can external factors reduce fire losses in Toronto?

Jie Huang(1004925156)

2020/12/22

Code and data supporting this analysis is available at:https://github.com/JaneHHH/STA304-Final-Project (https://github.com/JaneHHH/STA304-Final-Project)

Abstract

Fires are adverse events that cause real damage to property and human life. In addition to these material costs, the extent to which fires are harmful to the natural environment is less direct and obvious. In this report, I would like to build a statistical model and aim to find the relationship between the extent of the fire and the method of fire control and estimate the values of civilian casualties, count of persons rescued, estimated dollar loss based on the number of responding personnel from Toronto Open Data Portal 2020 About Fire Incidents in Toronto. In other words, this model justifies the relationship between causes and control method of fire; meanwhile, it explains that the people and apparatus on-site help reduce the loss of fire incidence.

Keywords

Propensity Score, Observational Study, Fire incidence, Causal Inference, Prevention Method

Introduction

The Canadian government widely uses statistical analysis in daily life. The most common ones are the crime rate and divorce rate. At the same time, statistical analysis can also summarize the recent fire incidences in Toronto and the casualties and property losses in the incidence. It is scientifically that "fire presents a significant risk to businesses. It can kill or seriously injure employees or visitors and can also damage or destroy buildings, equipment or stock."[1] In other words, with these data, the government can analyze these cases, then improve fire protection measures, or cause attention to reducing the fire incidences as possible.

Fire is a kind of harm to people, and it will pollute the environment and cause a certain degree of loss of people or society. To use the Toronto Open Data Portal data for analyzing the external factors that affect the loss in the fire incidence. The external factor is the number of responding personnel. I want to find the relationship between the extent of the fire and the method of fire control at the beginning to find the most common cause of the fire and the fastest way to extinguish the fire. Then I will find the connection between the fire loss and the external factors.

In the step-by-step analysis, the first step, I sort the dataset and clean the data to keep the main variables that I need, removing the NAs to reduce errors. Then I summary each variable and build a multiple linear regression model based on sorted data. Finally, I make a discussion on the results, strengths, and weaknesses of the model.

Methodology

Data

I download the datasets from the Toronto Open Data Potral. Then I use R studio to run code to help us select the valid respondents and clean the data with missing values(NA) and the undetermined data(99 - Undetermined) in the variable Extent_Of_Fire. In this project, the target population is all fire cases in Toronto in 2020; the frame is 17536 cases; the sample is 1000 cases. The 1000 cases are randomly selected from the dataset.

Table1: Extent of Fire

Number	Description of fire extent
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)
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)

Table1 shows that the extent of fire control and the ordering number will determine the different extent.

Table2: Extent of Fire

Number	Description of fire control
1	Extinguished by fire department

2	Extinguished by automatic system
3	Extinguished by occupant
4	Fire self extinguished
5	Action taken unclassified

Table2 shows that the method of fire control and the ordering number will determine the different styles. Still, I try to change these two tables' descriptions as numerical variables, summarized as a table to see the sum of various types and the relationship by plots.

Below is the overall summary table for our dataset: Table3: Summary Table for the Overall Dataset

```
##   Extent_Of_Fire      Method_Of_Fire_Control Civilian_Casualties
##   Length:1000         Length:1000            Min.   :0.000
##   Class :character    Class :character       1st Qu.:0.000
##   Mode  :character    Mode  :character       Median :0.000
##                                              Mean   :0.104
##                                              3rd Qu.:0.000
##                                              Max.   :4.000
##   Count_of_Persons_Rescued Estimated_Dollar_Loss Number_of_responding_personnel
##   Min.   :0.000            Min.   :        0     Min.   :  4.00
##   1st Qu.:0.000            1st Qu.:      200     1st Qu.: 20.00
##   Median :0.000            Median :     2000     Median : 22.00
##   Mean   :0.051            Mean   :    36671     Mean   : 28.76
##   3rd Qu.:0.000            3rd Qu.:    10000     3rd Qu.: 34.00
##   Max.   :7.000            Max.   :13000000      Max.   :246.00
##    extent_fire       fire_control
##   Min.   : 1.000    Min.   :1.000
##   1st Qu.: 1.000    1st Qu.:1.000
##   Median : 2.000    Median :1.000
##   Mean   : 2.024    Mean   :1.837
##   3rd Qu.: 2.000    3rd Qu.:3.000
##   Max.   :11.000    Max.   :5.000
```

In the table3, I can find that the mean of the civilian casualties is 0.104, which means that the value of civilian casualties in the fire incidence is not large. Most of people are safe in the fire incidence, meaning that the fire awareness is very strong. The IQR in the count of persons rescued is 0. And the maximum value of the estimated dollar loss is 13000000. When the fire incidence happened, there are minimum 4 responding people around and try to help the firemen to extinguish the fire and save people.

Model

I use R studio to build three simple linear regression models to see the association between the three independent variables and the response variable. In addition, it is appropriate to use a simple regression

model to analyze the dataset, predicting the relationship between the dependent and independent variable. 'Linear regression is the most basic and commonly used predictive analysis. Regression estimates are used to describe data and to explain the relationship between one dependent variable and one or more independent variables. At the center of the regression analysis is the task of fitting a single line through a scatter plot.'[2] Besides, I set up a null hypothesis that the coefficients of predictor variables are equal to zero, indicating no association between the dependent and the independet variable. Meanwhile, I choose the significance level of 0.05, to compare it with the p-value.

Plot1: Scatterplot of personnel and civilian casualties

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, : at
## -0.02
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## radius 0.0004
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, : all
## data on boundary of neighborhood. make span bigger
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## pseudoinverse used at -0.02
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## neighborhood radius 0.02
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## reciprocal condition number 1
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## zero-width neighborhood. make span bigger
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, : at
## -0.02
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## radius 0.0004
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, : all
## data on boundary of neighborhood. make span bigger
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## pseudoinverse used at -0.02
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## neighborhood radius 0.02
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## reciprocal condition number 1
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## zero-width neighborhood. make span bigger
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, : at
## -0.02
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## radius 0.0004
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, : all
## data on boundary of neighborhood. make span bigger
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## pseudoinverse used at -0.02
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## neighborhood radius 0.02
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## reciprocal condition number 1
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## zero-width neighborhood. make span bigger
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, : at
## -0.02
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## radius 0.0004
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, : all
## data on boundary of neighborhood. make span bigger
```
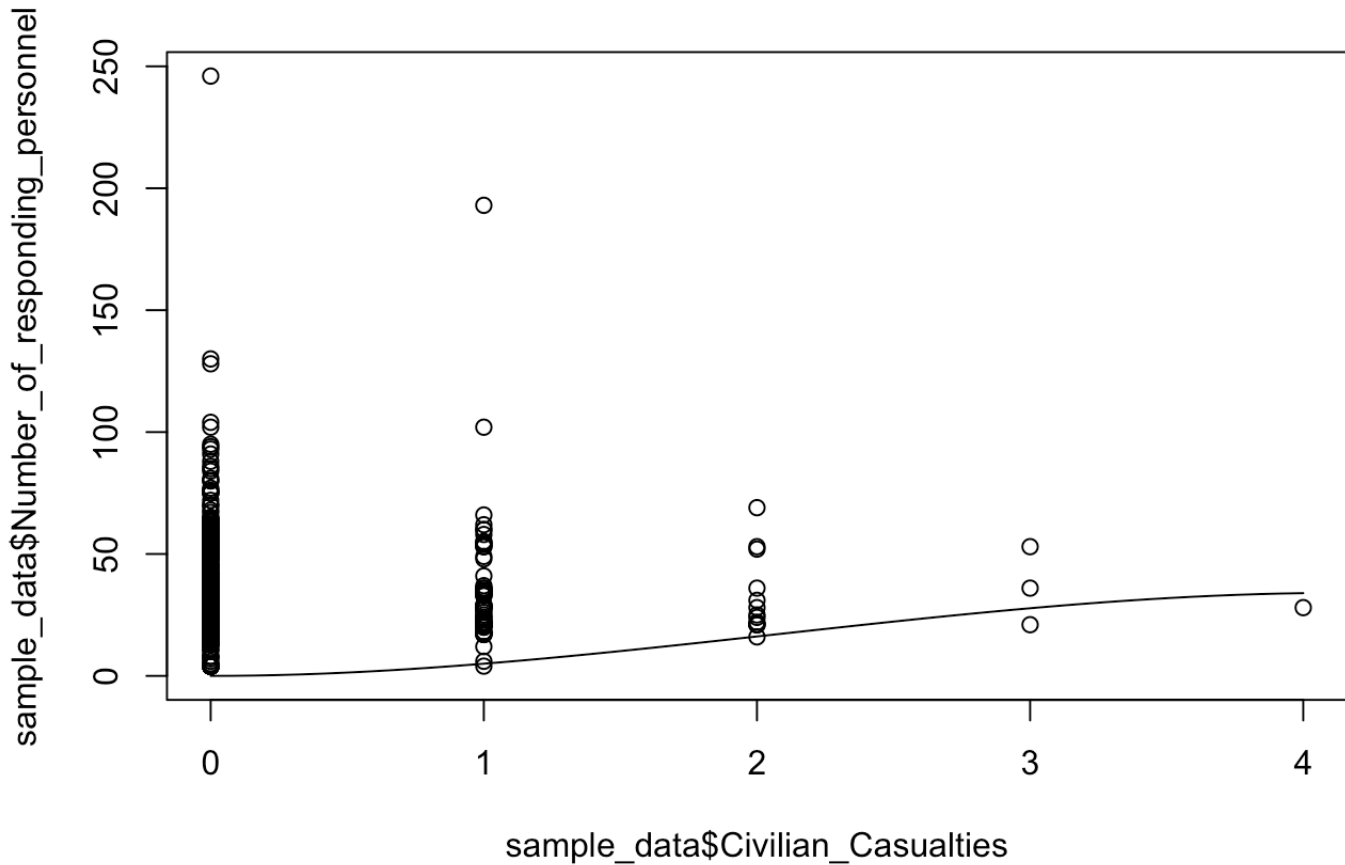
```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## pseudoinverse used at -0.02
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## neighborhood radius 0.02
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## reciprocal condition number 1
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## zero-width neighborhood. make span bigger
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, : at
## -0.02
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## radius 0.0004
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, : all
## data on boundary of neighborhood. make span bigger
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## pseudoinverse used at -0.02
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## neighborhood radius 0.02
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## reciprocal condition number 1
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## zero-width neighborhood. make span bigger
```

Personnel and civilian casualties

The Plot1 along with the smoothing line above suggests a weak linearly increasing relationship between the number of responding personnel and civilian casualties variables. This is a good thing, because, one of the assumptions in linear regression is that the relationship between the response and predictor variables is linear. And there are some outliers in the plot.

Plot2: Scatterplot of personnel and persons rescued

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, : at
## -0.035
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## radius 0.001225
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, : all
## data on boundary of neighborhood. make span bigger
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## pseudoinverse used at -0.035
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## neighborhood radius 0.035
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## reciprocal condition number 1
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## zero-width neighborhood. make span bigger
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, : at
## -0.035
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## radius 0.001225
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, : all
## data on boundary of neighborhood. make span bigger
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## pseudoinverse used at -0.035
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## neighborhood radius 0.035
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## reciprocal condition number 1
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## zero-width neighborhood. make span bigger
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, : at
## -0.035
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## radius 0.001225
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, : all
## data on boundary of neighborhood. make span bigger
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## pseudoinverse used at -0.035
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## neighborhood radius 0.035
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## reciprocal condition number 1
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## zero-width neighborhood. make span bigger
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, : at
## -0.035
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## radius 0.001225
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, : all
## data on boundary of neighborhood. make span bigger
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## pseudoinverse used at -0.035
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## neighborhood radius 0.035
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## reciprocal condition number 1
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## zero-width neighborhood. make span bigger
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, : at
## -0.035
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## radius 0.001225
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, : all
## data on boundary of neighborhood. make span bigger
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## pseudoinverse used at -0.035
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## neighborhood radius 0.035
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## reciprocal condition number 1
```
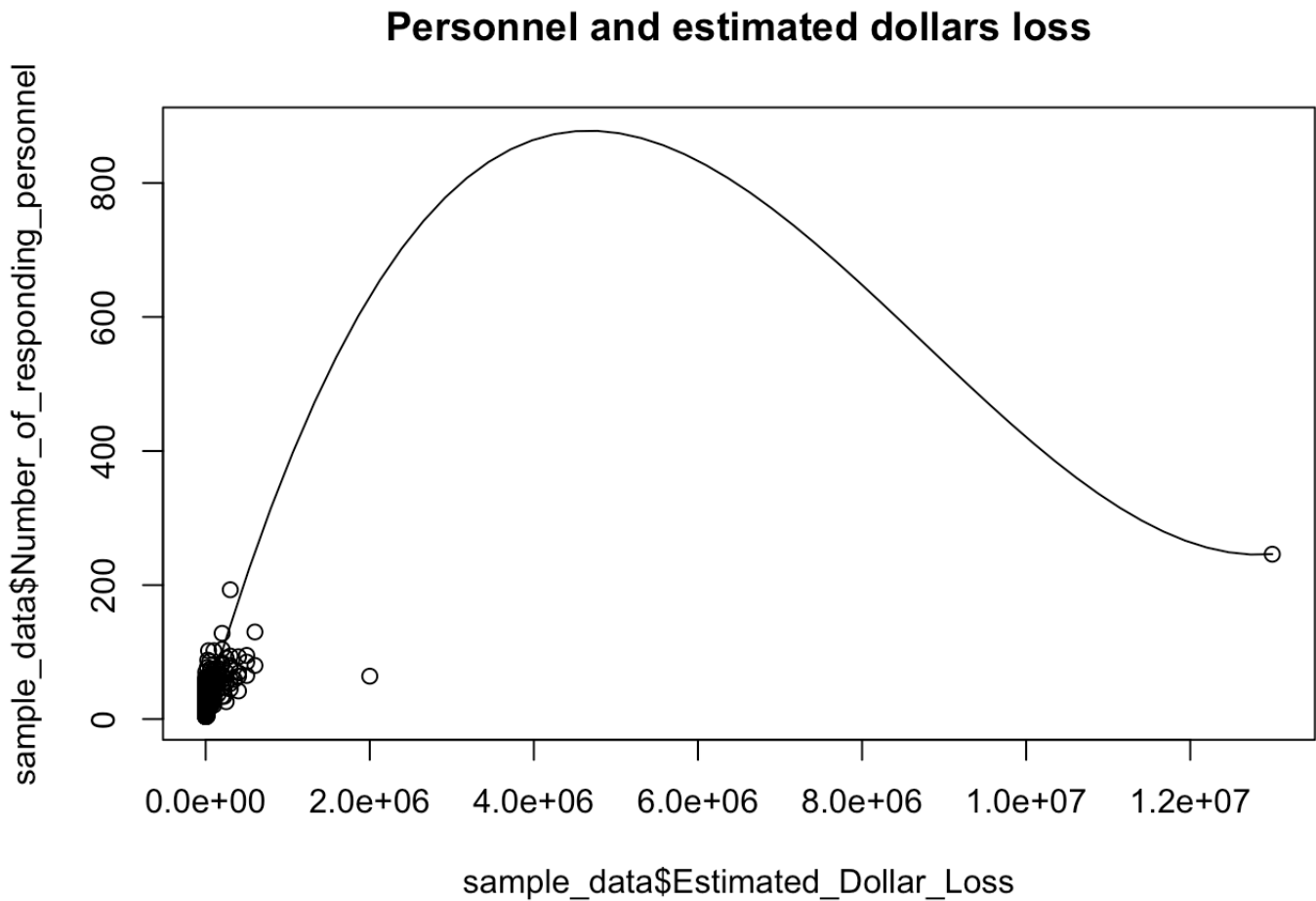
```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = FALSE, :
## zero-width neighborhood. make span bigger
```

Personnel and person rescued
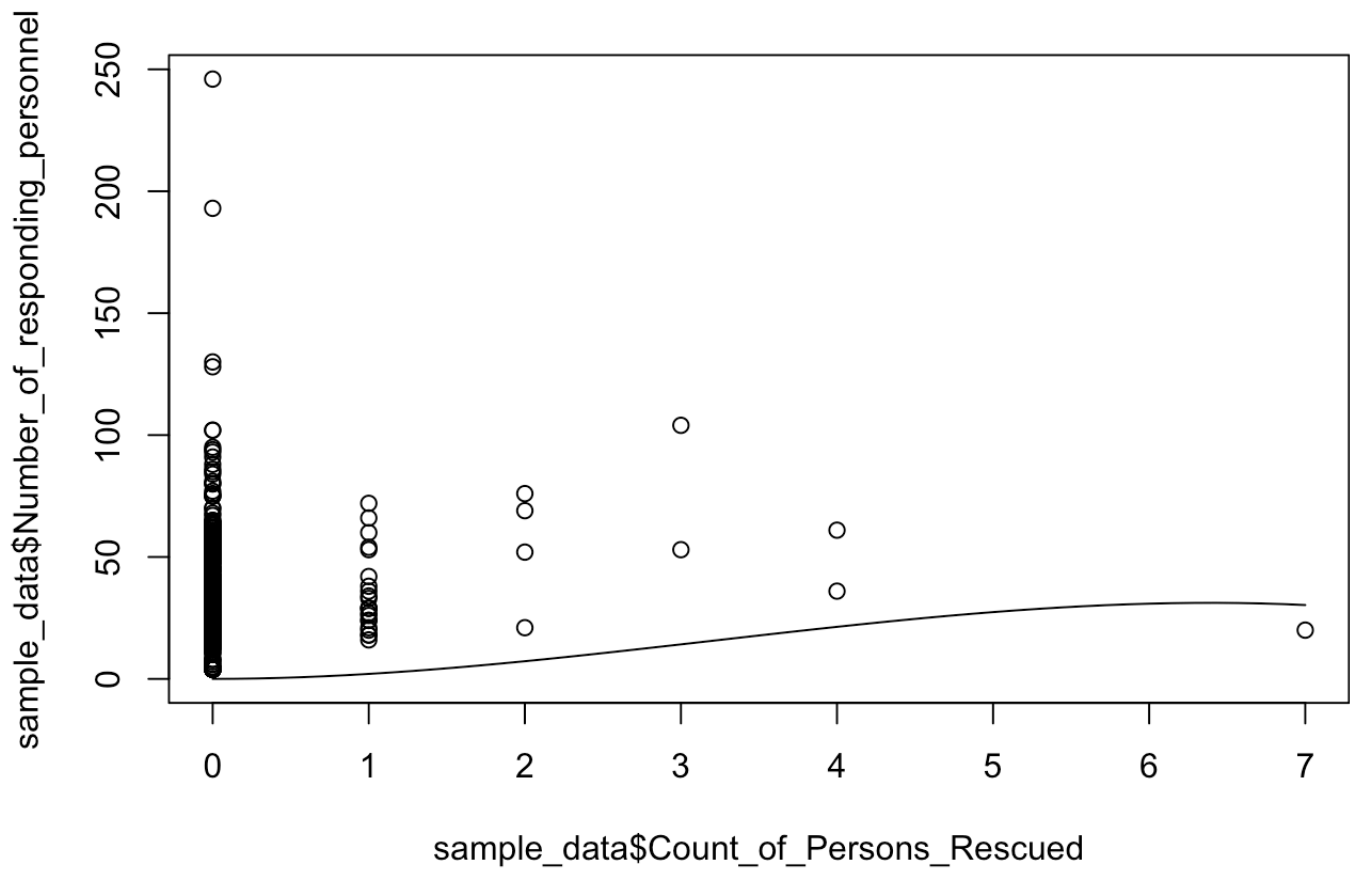
There is a weak positive increasing line in the plot2 with some outliers.

Plot3: Scatterplot of personnel and estimated dollar loss

Personnel and estimated dollars loss

The line in this plot is incresing at the beginning, then it starts to decrease. There are some outliers in this plot.

Results

By fitting the linear regression model, I find that the p-value in Table3 is < <2e-16, which is significantly smaller than 0.05. Thus I reject the null hypothesis. Then I can get three linear functions of these three models.

Table3: Summary Table for Regression Model

```
##
## Call:
## lm(formula = Number_of_responding_personnel ~ Civilian_Casualties,
##      data = sample_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.991  -8.386  -6.386   4.614 217.614
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           28.386      0.600  47.311   <2e-16 ***
## Civilian_Casualties    3.605      1.500   2.404   0.0164 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.32 on 998 degrees of freedom
## Multiple R-squared:  0.005756,   Adjusted R-squared:  0.004759
## F-statistic: 5.777 on 1 and 998 DF,  p-value: 0.01642
```

For the number of responding personnel and civilian casualties model, the simple linear function is

$$y = 28.386 + 3.605x_{civiliancasualties}$$

. Table 3 shows the important value what I need.

Table4:Summary Table for Regression Model

```
##
## Call:
## lm(formula = Number_of_responding_personnel ~ Count_of_Persons_Rescued,
##     data = sample_data)
##
## Residuals:
##     Min       1Q  Median       3Q      Max
## -46.877   -8.481   -6.481    4.519  217.519
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  28.4813     0.5831  48.846  < 2e-16 ***
## Count_of_Persons_Rescued      5.4851     1.5753   3.482  0.00052 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.26 on 998 degrees of freedom
## Multiple R-squared:  0.012,  Adjusted R-squared:  0.01101
## F-statistic: 12.12 on 1 and 998 DF,  p-value: 0.0005195
```

For the number of responding personnel and count of persons rescued model, the simple linear function is

$$y = 28.386 + 3.605 x_{countofpersonsrescued}$$

. Table 4 shows the important values what I need.

Table5:Summary Table for Regression Model

```
##
## Call:
## lm(formula = Number_of_responding_personnel ~ Estimated_Dollar_Loss,
##     data = sample_data)
##
## Residuals:
##     Min      1Q   Median      3Q      Max
## -43.197  -8.044  -6.064   4.976  158.949
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2.802e+01  5.181e-01   54.09   <2e-16 ***
## Estimated_Dollar_Loss 2.009e-05  1.230e-06   16.33   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.32 on 998 degrees of freedom
## Multiple R-squared:  0.2108, Adjusted R-squared:   0.21
## F-statistic: 266.6 on 1 and 998 DF,  p-value: < 2.2e-16
```

For the number of responding personnel and estimated dollar loss model, the simple linear function is

$$y = (2.802e + 01) + 2.009e - 05x_{cestimateddollarloss}$$

. Table 5 shows the important values what I need.

–Diagonostic check: Plot4: 4 differnt plots
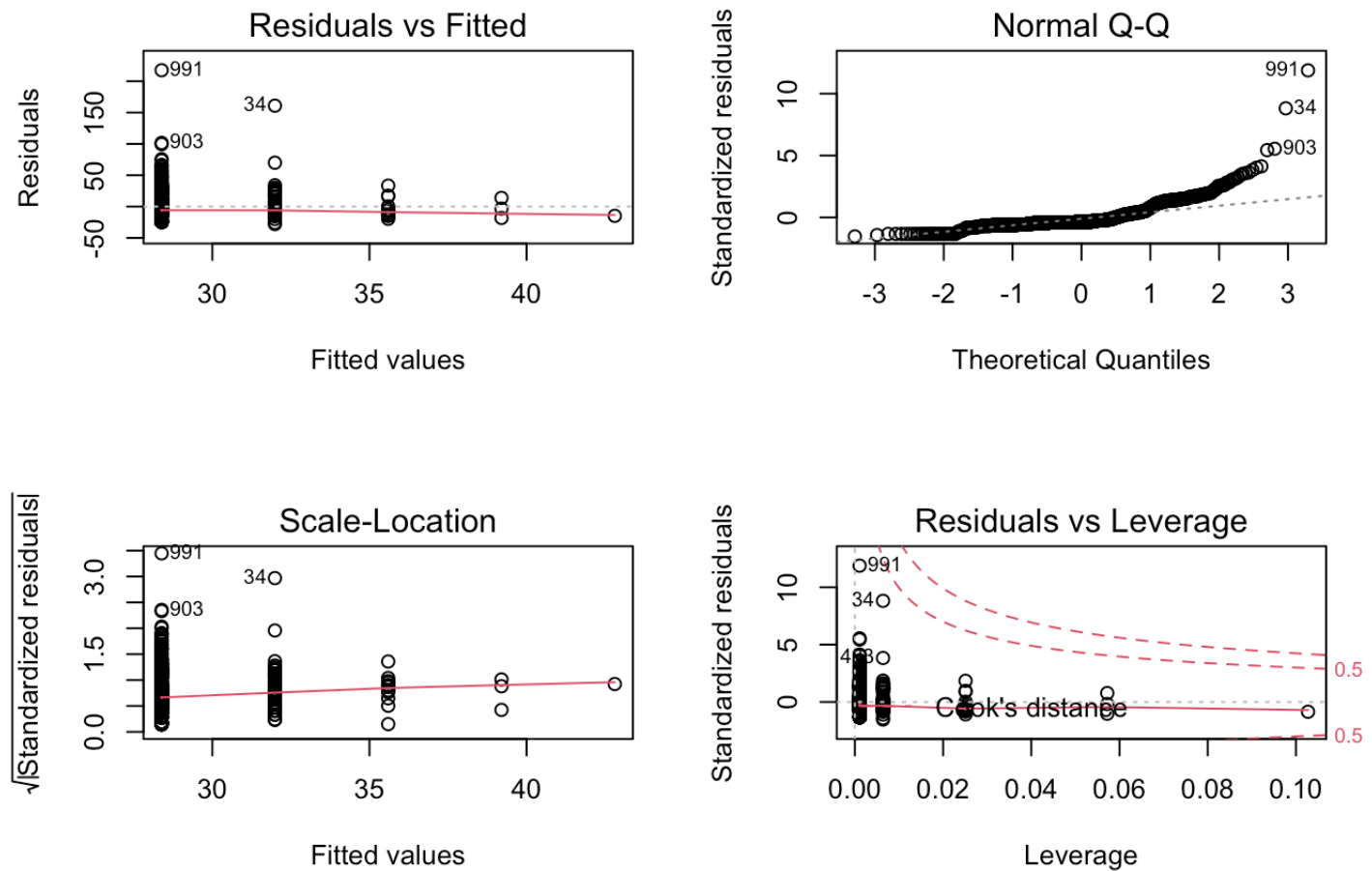
In the residuals vs. fitted plot, the horizontal line in this plot indicates a linear relationship between the independent variable and the dependent variable.In the normal Q-Q plot, we can find that there is a right-skewed line. In the third plot, the line with a slightly increasing trend shows the relationship between the root of standardized residuals and fitted values. Some residuals do not appear randomly spread. This last plot helps us to find influential cases. This plot is the typical look when there is no significant case or cases. It is barely to see Cook's distance lines (a red dashed line) because all cases are not well inside Cook's distance lines, meaning they have high Cook's distance scores, the cases are influential to the regression results.
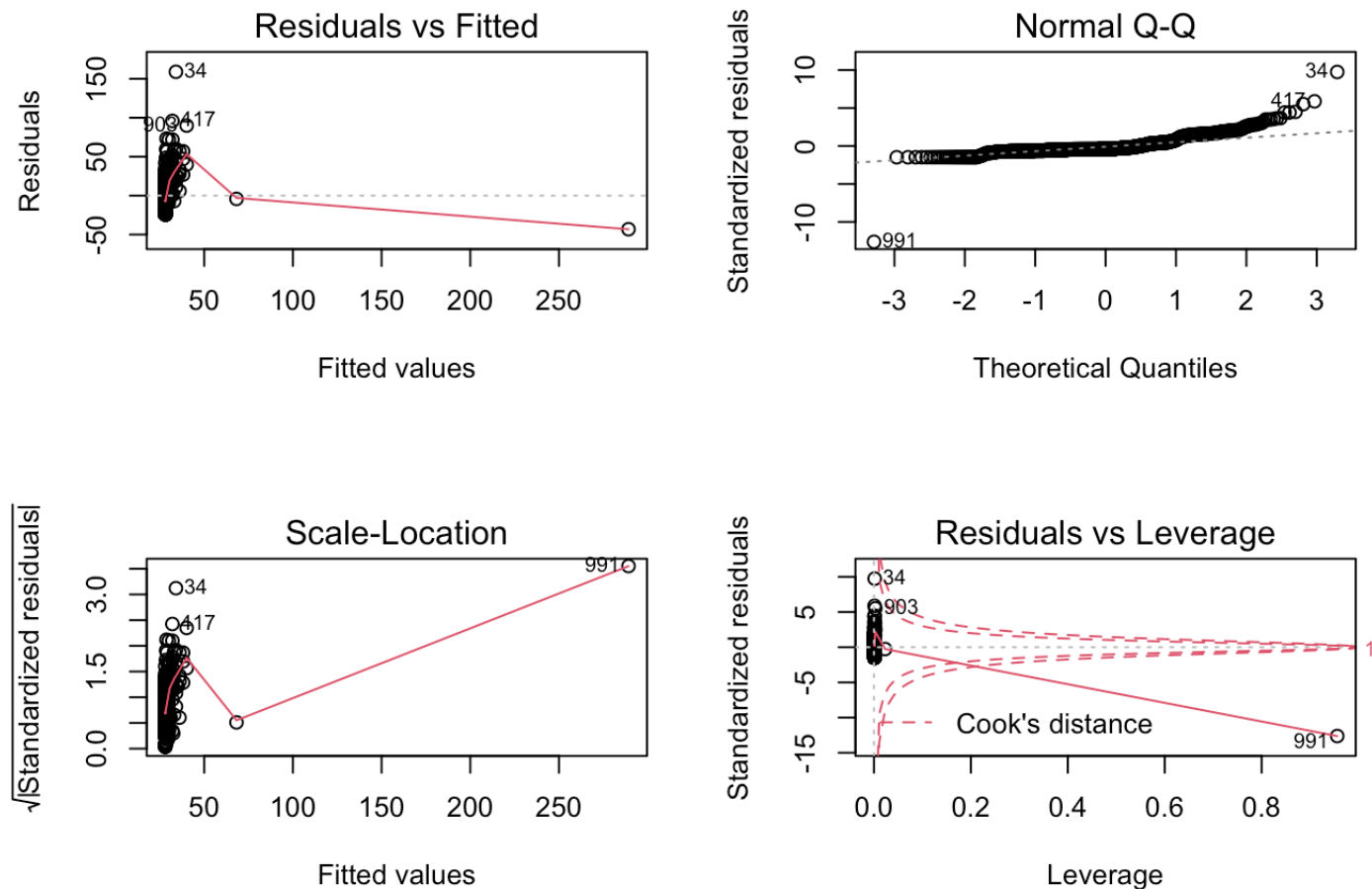
Plot5: 4 different plots
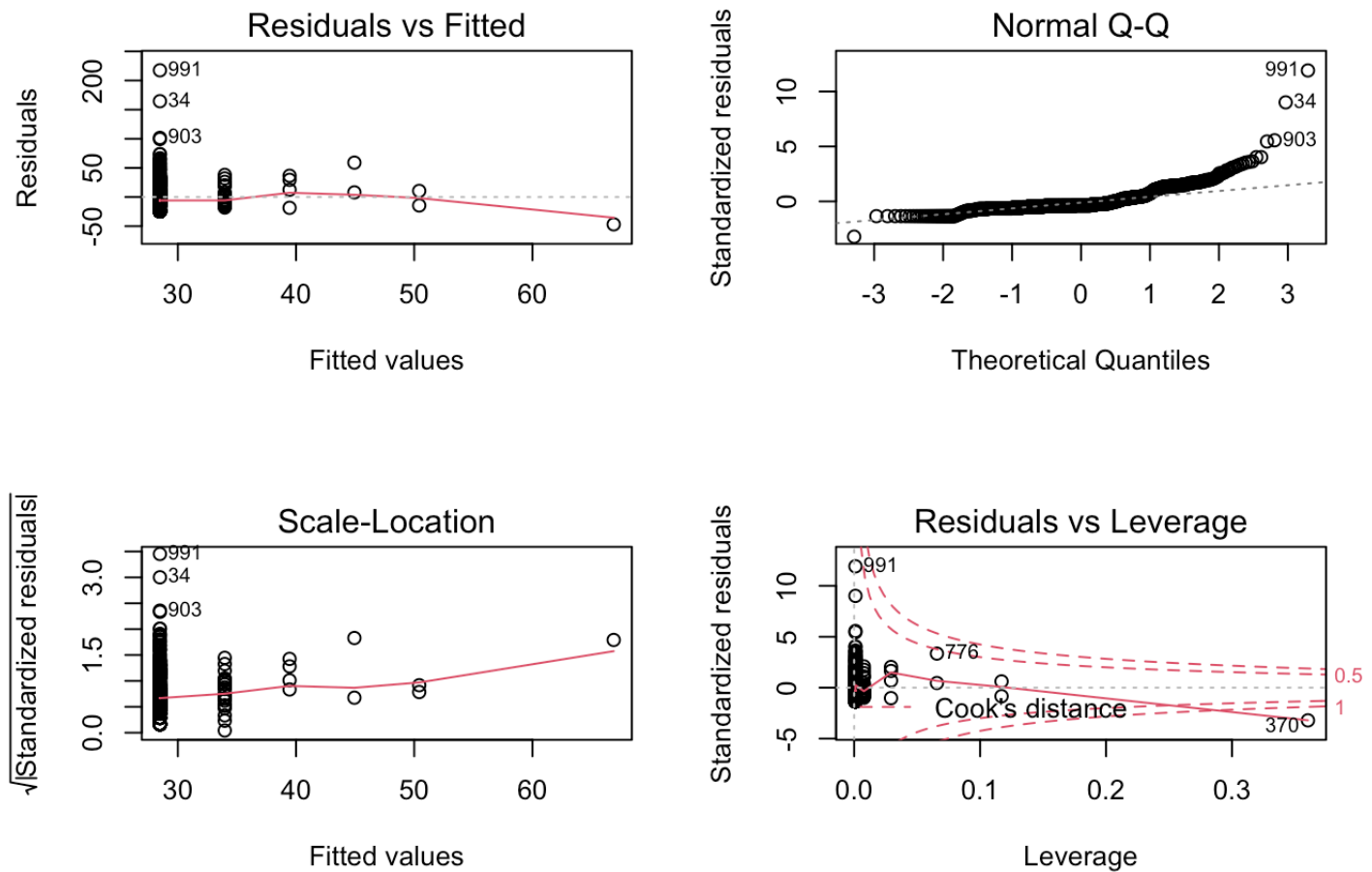
In the residuals vs. fitted plot, the line is not a horizontal line and the trend of it is decreasing. In the normal Q-Q plot, we can find that there is a right-skewed line. In the third plot, the line with a strongly increasing trend. Some residuals do not appear randomly spread. This last plot helps us to find influential cases. It is barely to see Cook's distance lines because all cases are well inside Cook's distance lines, meaning they do not have high Cook's distance scores, the cases are not influential to the regression results.

Plot6: 4 differnt plots

In the residuals vs. fitted plot, the line is not a horizontal line and the trend of it is decreasing. In the normal Q-Q plot, we can find that there is a right-skewed line. In the third plot, the line with a strongly increasing trend. Some residuals do not appear randomly spread. This last plot is barely to see a red dashed line, which contains all the cases.

Disscussion

Summary: At the beginning, I try to clean the data and create a new data frame to analyze. After cleaning the data, I try to build the multiplier linear regression model to classify the relationship between the number of responding personnel and civilian casualties, count of persons rescued, estimated dollar loss. However, I find that there is no relationship between these variables. So that I try to make three simple linear models to analyze the correlation between the dependent and independent variables. Then I build 4 plots for each variable to do some diagnostic checks to justify the correlation is strong and valid. Finally, I got the results and make a conclusion.

The p-value in this question is absolutely smaller than 0.05, but I cannot get the values of each variable in the same situation. Therefore, I cannot classify that the people will affect the fire incidence in someway.

Then I create three linear models to analyse. I find that the number of responding personnel has the weakly positive relationship with the civilian casualties and the count of persons rescued. When the number of responding people increase, the civilian casualties and persons rescued will increase softly. It is possible that these people who want help the firemen to distinguish the fire, however, they do not have the related professional knowledge to protect themselves. Then they may become the injured person from the rescuers. The correlation between the responding people and estimated dollar loss has is very specific. The line increases at the beginning and then it decreases.The relationship between the dependent variable and the independent variable cannot get a clear conclusion. The responding people can call the 911 to find the help from the firemen. The extent factors can be determined as the responding people, responding equipment, weather ect. And there are many unpredictable variable, we cannot conclude that all the external factors will affect the fire loss. At the same time, the fire loss is also related with the the fire size and current situation. We cannot make a standard conclusion to estimate the loss in the fire. Number of deaths, injuries, and property damage connect with the fire where it happened. The relationship between these variables are very weak.

Weakness of Model and Possible Future Improvements

The multiplier linear regression is fail. I cannot get the formula of the linear regression model. I try to find the correlations between the predictor and the three variables, however, I find that the relationship between is very week, meaning that we can ignore the relationship between them. And there are lots of debates after I complete the project. the number of people who would like to help the people in fire will affect the people rescued. But not every time someone is willing to help you. And the data has problems as well. The people rescued and civilian casualties are very similar, we cannot define these two very clearly.

Then, the data we are using is not perfectly accurate, since I do not select a large sample set. Another weakness of the model is that the relationship between the predictor variable and response variable is real small, which means that they may do not have any connection. And for these data, I do not have a good survey to help analyze. To fix this problem, I think I need to change the predictor variable and response variable, finding some variables with strong relationship. Besides, to make the model more accurate, I can change the sample size and the add more useful predictors. If it is possible, I will try to find a survey about the fire incidence to help estimate in the future.

Reference

[1]Ahrens, M., et al. "Environmental Impact of Fire." Fire Science Reviews, SpringerOpen, 1 Jan. 1970, firesciencereviews.springeropen.com/articles/10.1186/s40038-016-0014-1.

[2]"Conduct and Interpret a Linear Regression." Statistics Solutions, 1 May 2020, www.statisticssolutions.com/conduct-interpret-linear-regression/.

[3]data_hacks, and Harshita_Dudhe. "What Is the Singularity Error in Linear Regression." Data Science, Analytics and Big Data Discussions, 8 Sept. 2015, discuss.analyticsvidhya.com/t/what-is-the-singularity-error-in-linear-regression/3924.

[4]Hadley Wickham, Romain François, Lionel Henry and Kirill Müller(2020). dplyr: A Grammar of Data Manipulation. R package version1.0.2. https://CRAN.R-project.org/package=dplyr (https://CRAN.R-project.org/package=dplyr)

[5]"Healthy Working Lives - Fire." Common Fire Related Hazards, www.healthyworkinglives.scot/workplace-guidance/safety/fire/Pages/fire-related-hazards.aspx.

[6]JohnKJohnK 16.8k88 gold badges5353 silver badges9898 bronze badges, et al. "How to Interpret a QQ Plot." Cross Validated, 1 July 1963, stats.stackexchange.com/questions/101274/how-to-interpret-a-qq-plot.

[7]"Open Data Dataset." City of Toronto Open Data Portal, open.toronto.ca/dataset/fire-incidents/.

[8]R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/ (https://www.R-project.org/).