Final Project

William Gagne

Faculty of Arts & Science, University of Toronto

STA302: Methods of Data Analysis I

Dr. Katherine Daignault

December 20th, 2022

Introduction

Municipalities and cities across the globe deploy considerable resources into their fire fighting department to minimize the damage of residential fires and prevent the loss of lives. Although residential fire is an important issue, little has been done to research the topic, which is why it will be the focus of this study. The present research will explore whether it is possible to predict the time it takes to control a fire using the extent and ignition source of a fire, the number of responding personnel, the presence of a fire alarm system and the time to arrival of the firefighters. Hopefully, investigating the relationships between the predictors and the response variables will be insightful for firefighting departments.

Past studies showed that firefighters' response time has a linear relationship with the size of a fire, that the source of ignition and fire alarm systems affects the size of a fire, and that the firefighters' crew size impacts the speed at which firefighters complete firefighting tasks (Challands, 2010; Crew Size & Arrival Times influence Fire Outcome, 2010; Holborn, 2004). However, none of the previous research, except the study performed on firefighters' crew size, investigated whether the variables they studied relate to the time it takes to control a fire. Consequently, the current research studies variables that were proven significant in previous research. It then uses these variables to investigate a relationship not undertaken by previous research.

Methods

After randomly splitting the fire incidents dataset from the city of Toronto in half (training and test dataset), we ensure that both datasets have similar statistics (Appendix Table 1). Then, we perform the exploratory data analysis and check the additional conditions to ensure we can interpret the residual plots. Next, we formerly check the assumptions for linear regression using the residual plots. In the residuals versus predictors plots and residuals versus fitted value plots, we check for any discernable patterns indicating a linearity assumption violation. Additionally, we may have correlated errors if we see that large clusters of residuals have an apparent separation from the rest. If we observe a fanning pattern in the residual, constant variance is likely violated. Additionally, we verify the residuals QQ plot. If we observe severe deviations from the standard normal quartiles, normality may be an issue.

In the case of a constant variance violation, we will apply a variance stabilizing transformation on the response by either taking the square root or the natural log of the response. For linearity or normality violations, we will use the Box-Cox transformation method to help decide on a simpler transformation using the lambda value. Once we find the optimal transformation, we recheck the assumptions for linear regression.

We then carry out individual T-tests on each coefficient to know which predictors are not linearly related to the response in the presence of the other predictor. If the T-test's p-value exceeds 5% for a coefficient (not significant), we add the predictors to the set of predictors we want to remove. In the case of our study, only one of the predictors does not appear to be linearly related to the response; hence we remove the predictor from the model. If the set of insignificant

predictors had multiple predictors, we would perform a partial F-test on the set of insignificant predictors and remove the predictors from the model only if the p-value of the partial F-test exceeds 5%. Next, we verify the assumptions for linear regression again.

Subsequently, we check for multicollinearity using the variance inflation factor. Any VIF value exceeding five indicates severe multicollinearity between two variables. We then record our model's leverage points, outliers and influential points.

Lastly, we validate our model by fitting the preferred model to the test dataset. We compare the R-squared value of the training and test models to know whether our model can explain variation in the residuals of other datasets. We then ensure that roughly the same predictors appear significant in both models and that the equivalent predictors of each model are no more than one standard deviation away from each other. If this is not the case, we may be overfitting the training data; thus, we fail to validate the model. Additionally, we fail to validate the model if we observe different assumption violations or if there are any other significant differences between the models.

Results

The exploratory data analysis indicates that some issues may exist with the data (Figure 1). The first problematic observation is that the data is heavily right-skewed for a few variables, including the response, which is likely to cause a normality assumption violation. Furthermore, the variance does not appear constant in the "Number of responding personnel" versus the response plot and the "Time to arrival" versus the response plot. Thus, we may also violate the constant variance assumption. In terms of linearity, there is no clear violation. Lastly, the errors may be correlated, but we cannot confirm this from our exploratory data analysis.

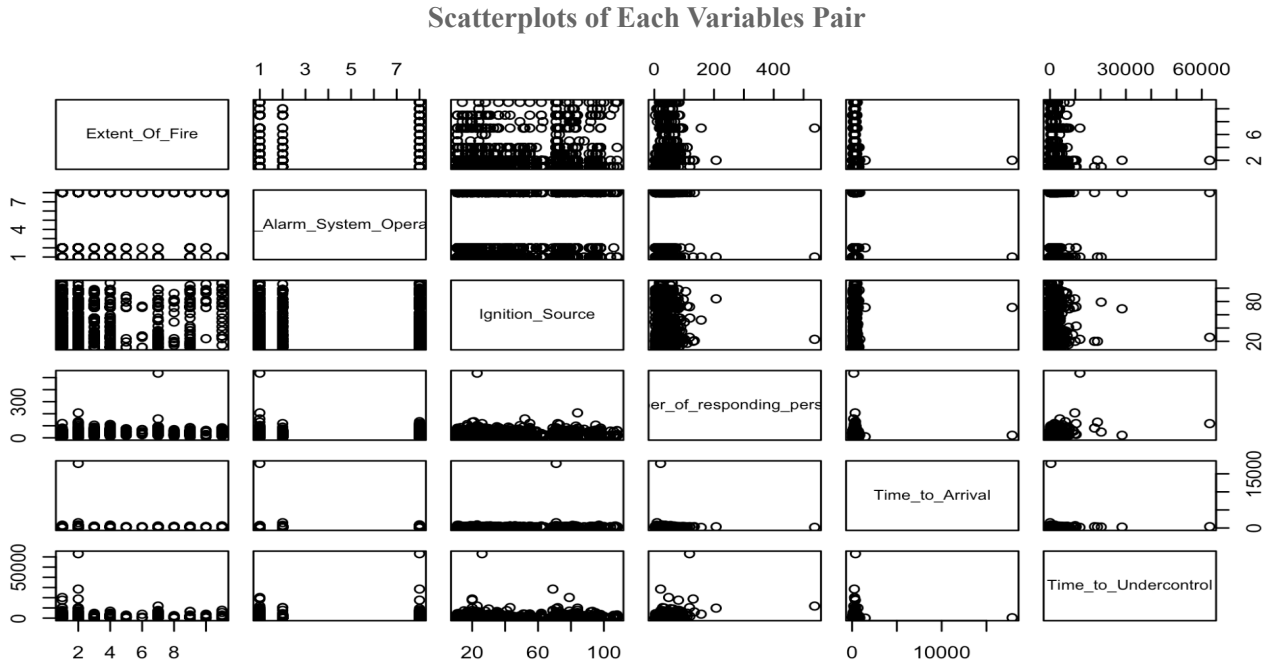

Figure 1: We use the scatterplots of each variable to know which assumption violations might be present in our data.

After fitting the model and analyzing the residual plots, we can confirm that the problems we expected from the EDA are indeed present in the data (Figure 2). The residuals are clustered, and we observe severe deviations between the theoretical and residual quantiles in the QQ plot, indicating that normality is an issue in our data. In addition to the normality violation, variance does not seem constant as we observe a fanning pattern in the residuals, especially in the "Time to arrival" versus residual plot.

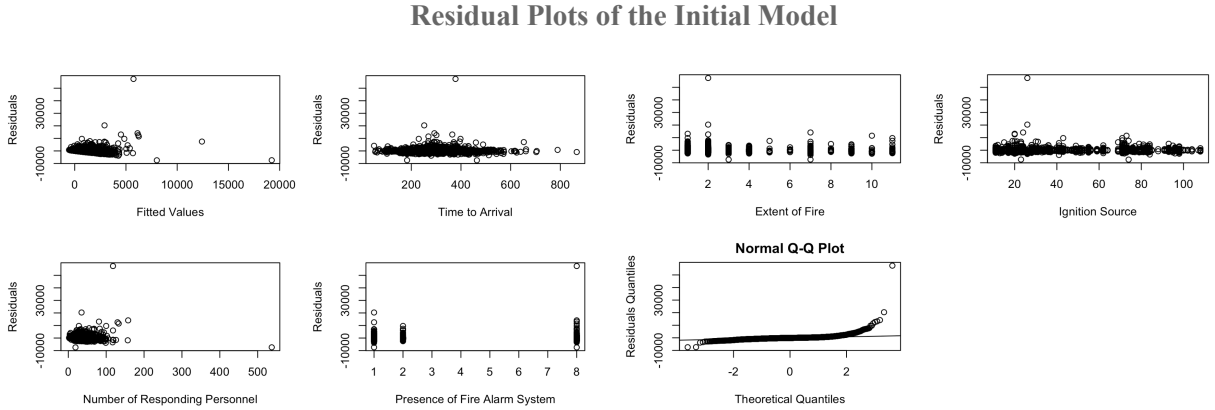

Figure 2: We use the residual plots of the initial model to formerly check for assumption violations. It appears that we are violating the constant variance and normality assumptions.

We use the BoxCox methods to help decide on a simple transformation for the response. The BoxCox lambda value is 20, which signifies that we should attempt a square root transformation on the response. Although normality has improved, a slight constant variance violation persists (Figure 3).

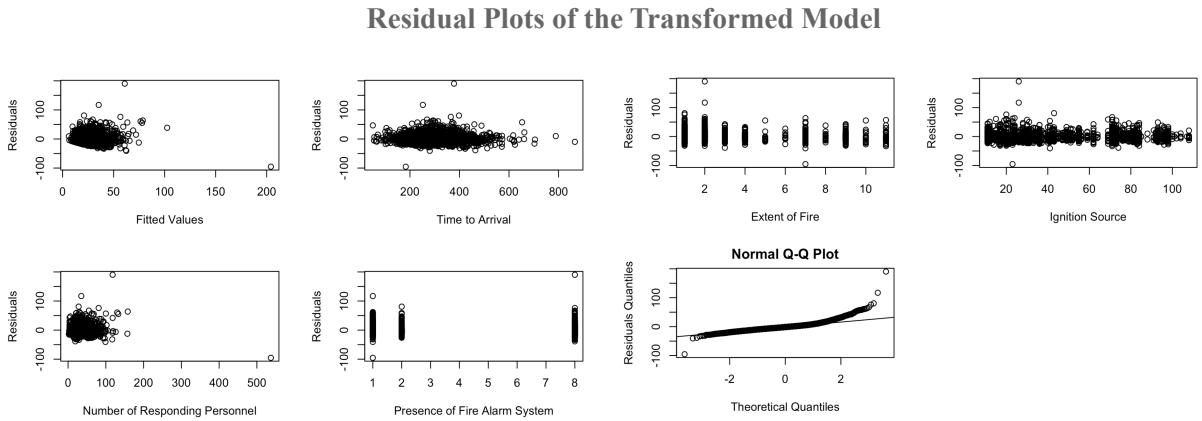

Figure 3: We use the residual plots of the transformed model to formerly check for assumption violations. It appears that we are still violating the constant variance and normality assumptions.

After performing T-tests on each of the predictor's coefficients, we observe that the "Time to arrival" is the only predictor not linearly related to the response. Consequently, we remove that predictor from our model. Next, we perform some tests to find the outlier, leverage and influential points in our model. There are 684 leverage points (a high number probably due to the skew observed earlier), 50 outlier points and notably 420 influential points using DFFITS. Three of these points appear to be errors as their values are non-sensical; hence, we remove them from the dataset.

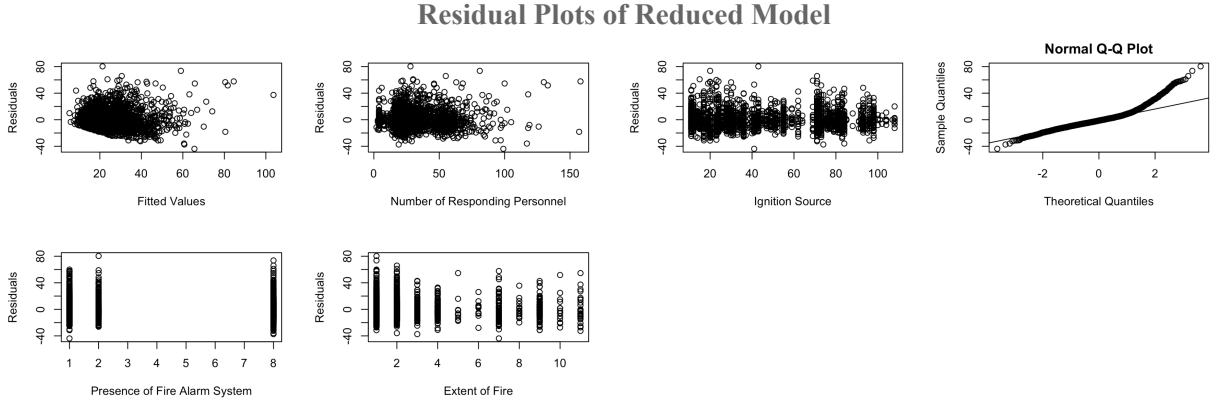

Figure 4: We use the residual plots of the reduced model to formerly check for assumption violations. It appears that we are still violating the normality assumptions. We may also be violating the constant variance assumption, but the violation is less severe than before.

At last, normality is still violated, but the constant variance assumption is slightly less problematic (Figure 4). Note that each change we made to our model increased the adjusted R-squared, decreased the corrected AIC, and decreased the BIC (Appendix Table 2). We even lessened the assumption violations in our final model. Therefore, we chose the final model as our preferred model as it is significantly better than the previous models.

Let us now test our model on the test dataset. By analyzing the residual plots of our preferred model on the test dataset, we observe that normality is violated, and constant variance is also slightly violated (Appendix Figure 5). Looking at Table 3, we see that both models have remarkably similar R-squared and adjusted R-squared values. Furthermore, in both models, the "Time to arrival" predictor did not appear linearly related to the response, while all other predictors were. The main aspect in which the two models differ is each predictor's coefficient. The difference in the predictors' coefficient is due to the dataset having a very high number of categories (over a hundred) and sometimes very few data points per category (one or two points). Hence, the discrepancies in the number of data points per category cause the coefficient to be more than two standard deviations apart for most predictors, which is why we fail to validate the model. The high number of leverage, outlier and influential points in both models is also a possible reason we observe such differences in our models.

Validation of the Preferred Model Table

Table 3: Comparisons between the preferred model on the train and test data. For each coefficient, the first value is the mean and the second value is the standard deviation.

Characteristic	Preferred Model (Train)	Preferred Model (Test)
R-squared	0.3831	0.3858
Adjusted R-squared	0.3662	0.3695
Largest VIF value	1.45	1.35
Number of Cook's D	0	0
Number of DFFITS	229	215
Violations	Constant variance, Normality	Constant variance, Normality
Intercept	5.956 ± 0.676	7.956 ± 0.654
Number_of_responding_personnel	0.368 ± 0.015	0.32 ± 0.013
Fire_Alarm_System_Operation-2	0.396 ± 0.65	-1.617 ± 0.613
Fire_Alarm_System_Operation-8	-1.115 ± 0.477	-2.027 ± 0.456
Extent_Of_Fire-2	2.23 ± 0.492	2.185 ± 0.473
Extent_Of_Fire-3	1.491 ± 1.088	4.32 ± 1.021
Extent_Of_Fire-4	5.507 ± 1.195	8.666 ± 1.083
Extent_Of_Fire-5	14.088 ± 3.911	10.217 ± 3.291
Extent_Of_Fire-6	2.132 ± 3.663	4.799 ± 4.442
Extent_Of_Fire-7	13.282 ± 1.653	11.475 ± 1.51
Extent_Of_Fire-8	5.948 ± 3.055	9.594 ± 2.834
Extent_Of_Fire-9	4.114 ± 1.292	4.376 ± 1.365
Extent_Of_Fire-10	14.619 ± 3.297	11.502 ± 3.314
Extent_Of_Fire-11	20.867 ± 2.456	9.094 ± 2.522
Ignition_Source-12	2.058 ± 1.057	0.274 ± 1.023
Ignition_Source-13	-3.181 ± 3.126	-1.35 ± 2.782
Ignition_Source-14	1.651 ± 1.711	0.074 ± 1.792
Ignition_Source-15	5.56 ± 2.655	11.626 ± 2.86
Ignition_Source-16	2.927 ± 2.171	1.955 ± 1.987
Ignition_Source-17	17.075 ± 1.908	18.895 ± 2.186
Ignition_Source-19	1.344 ± 1.721	-0.107 ± 1.626
...	...	...

Discussion

By building a linear model, we found that a linear relationship exists between the time it takes to control a fire and all of the predictors except the "Time to arrival" variable. Hence it is possible to roughly predict the time it takes to control a fire using the extent and ignition source of a fire, the number of responding personnel and the presence of a fire alarm system. However, the "Time to arrival" variable cannot be used to predict the time it takes to control a fire.

Analyzing the coefficient of the predictor variable helps us acquire insights about how the predictor affects the response. For example, if a fire alarm system is present, the ignition source is a stove, and the fire is confined, then for a one-unit increase in the number of responding personnel (while holding the other predictors fixed), the square root of the time to control the fire increases by 0.368 seconds. By looking at the relationship between the predictors and the response, we also see that having a fire alarm system reduces the time it takes to control a fire. Moreover, salamanders (a cooking appliance found in restaurants) account for the greatest increase in the time it takes to control the fire, and fires that spread beyond the building of origin take the longest to control.

Regarding the limitation of our model, the most worrying assumption violation is normality due to the response and the "number of responding personnel" variables being heavily right-skewed. Consequently, our model does not accurately capture the mean of the distribution, which causes the model to be biased. To correct the normality violation, we need to acquire a bigger sample of fire incidents and hope that CLT applies. Unfortunately, this is not something we can do in the current study. We tried a transformation on the response instead, which improved the normality violation but only partially corrected it. Lastly, the small sample size in some categories is another important limitation of our models that causes biased predictor coefficients, ultimately preventing us from validating the model.

References

Challands, N. (2010). Relationships Between Fire Service Response Time and Fire Outcomes. Fire Technology, 46(3), 665–676. https://doi.org/10.1007/s10694-009-0111-y

Crew Size & Arrival Times Influence Fire Outcome. (2010). Professional Safety, 55(6), 6–.

Holborn, P. ., Nolan, P. ., & Golt, J. (2004). An analysis of fire sizes, fire growth rates and times between events using data from fire investigations. Fire Safety Journal, 39(6), 481–524. https://doi.org/10.1016/j.firesaf.2004.05.002

Appendix

Comparison Between Training and Test Dataset Table

Table 1: The values in the "Time to arrival" row, the "Time to control the fire" row and the "Extent of the fire" row represent the mean and the value in parentheses is the standard deviation. The values in the "Extent of the fire" row, the "Presence of fire alarm system" row and the "Ignition source" row represent the counts for each category.

Variable	Training Set	Test Set
Number of firefighters	29.3583578 (18.3526302)	29.4403401 (17.1243906)
Time to arrival	294.1055718 (80.547576)	300.5464673 (311.9717444)
Time to control the fire	784.2384164 (1730.6070971)	744.9091176 (1276.0644579)
Extent of the fire	1486, 1396, 155, 129, 10, 11, 65, 16, 100, 14, 28	1470, 1395, 166, 148, 13, 7, 74, 18, 84, 13, 23
Presence of fire alarm system	1525, 473, 1412	1584, 493, 1334
Ignition source	739, 164, 15, 54, 21, 32, 43, 52, 37, 22, 1, 52, 104, 5, 42, 3, 95, 33, 64, 32, 8, 18, 20, 8, 53, 7, 4, 3, 13, 33, 16, 122, 3, 12, 1, 17, 7, 34, 34, 14, 2, 1, 110, 9, 14, 1, 16, 1, 3, 13, 534, 75, 75, 2, 21, 30, 30, 51, 4, 59, 30, 63, 31, 3, 2, 5, 15, 46, 1, 6, 33, 5, 62, 4, 1, 2, 5, 1, 7	744, 164, 18, 46, 17, 36, 30, 55, 38, 21, 6, 49, 113, 2, 33, 2, 73, 41, 52, 28, 10, 13, 33, 2, 47, 11, 3, 5, 12, 18, 14, 132, 2, 2, 1, 16, 6, 50, 43, 11, 1, 5, 107, 10, 9, 1, 17, 10, 21, 554, 57, 84, 15, 24, 31, 53, 4, 58, 26, 73, 41, 6, 10, 11, 32, 3, 6, 43, 1, 84, 3, 2, 2, 4, 4

Comparison Between Fitted Models Table

Table 2: Comparison between the different fitted models so that we can chose the preferred model. We chose the corrected AIC because the sampple size over the number of predictors plus two is smaller than 40.

Model	Adjusted R^2	Corrected AIC	BIC
Full model	0.23	5.0009172×10^4	5.058434×10^4
Transformed model	0.3464181	1.7436118×10^4	1.8011286×10^4
Reduced model	0.366172	1.7009027×10^4	1.7578177×10^4

Residual Plots of the Preferred Model on the Test Data

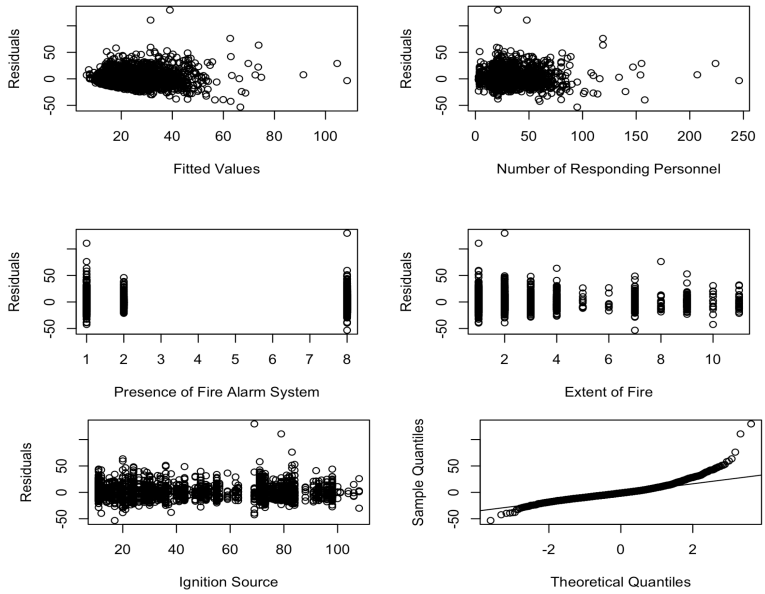

Figure 5: Overall, the residual plot of the preferred model on the test dataset is very similar to the training dataset's residual plots. We observe a violation of the normality assumption and a slight violation of the constant variance assumption.