Minor Correlation between Temperature and Accident Severity in US Population

Team Members (apple): Jie Huang, Mustafa Barez
Date: February 23rd, 2020

Abstract:

The following paper seeks to present the relationship between temperature and accident severity from a US dataset containing countrywide accidents from 49 states over a period of 3 years. It was found that there was little correlation between temperature and accident severity based on a regression model. This is important as higher temperatures can be linked to more severe accidents and thus brings the question of how US citizens can be prepared when days of extreme temperatures arrive.

Introduction

According to a report by the world health organization (WHO), the US has traffic fatalities that are about 50 percent higher than similar nations Western Europe, Canada, Australia, and Japan [1]. The total fatalities in the US are about 12.4 deaths out of every 100,000 accidents according to the road and safety report by the WHO [2]. A number of factors were noted the WHO study including intoxication, speeding, and seat belt laws, but there was no mention of the temperatures when most of these collisions happened. Very little research has been presented to show a correlation, if any, between the number of accidents and the temperature of a city which brings the question as to whether or not there is a relationship. In addition, the severity of an accident can also be the result of a drastic change in temperature which is what this present study seeks to find.

The following research paper describes a study conducted through a US population dataset containing the number of accidents from 2016-2019. This research attempts to predict whether or not there is a correlation between temperature and the severity of an accident. With this, readers can get a visual example, through a linear regression model, and see if there is a relationship, and with what temperatures and what severity levels.

Dataset

The following dataset was taken from Kaggle and represents the countrywide accidents in the US. The dataset covers 49 states in the United States and was collected from February 2016 to December 2019 [3]. The data was captured using data providers that included APIs which

provided streaming traffic incident data [4]. The APIs broadcasted traffic data captured by traffic sensors, cameras, and law enforcement agencies. The original dataset contains 49 attributes which include a unique ID, source of accident, city, etc., and contains 3 million records. A subset of this data was taken and contains only the temperature and severity attributes (from 1 – 4 where 1 indicates the least impact on traffic and 4 indicates the most impact on traffic) and contains the first 2000 rows of the original dataset.

Temperature (Degrees Fahrenheit)	Accident Severity (1-4)
36.9	3
37.9	2

Table 1. The following is an example of the table taken from the Kaggle website containing only the Temperature and Accident Severity attributes.

From Table 1 we get an example of the table used for the regression study with only 2 records. Here there is the temperature in degrees Fahrenheit, and the severity of the accident (1 being a short delay in traffic, and 4 being an extreme accident causing many hours in delay).

Discussion

In order to answer the research question, we chose a linear regression model to display the analysis. The motivation for us to choose this model is that linear regression can best indicate the relationship between variable Temperature (F) and variable Severity from our dataset. This method will visually explain our question by fitting a line with variable Temperature(F) on the x-axis and variable Severity on the y-axis. We can then approach the answer by analyzing the linear graph. But before we start making the graph, there are several assumptions we need to make. Firstly, the relationship between Temperature(F) and Severity is linear. Secondly, the variable Severity is measured with continuous values. Thirdly, we assume the values of both variables are normally distributed. Next, we chose Temperature(F) as the x-axis, Severity as the y-axis, and used linear regression function to fit a line between these two variables. We then output a linear model equation: Y = Coefficient * X + Intercept, in which Severity is dependent variable Y, and Temperature(F) as independent variable X. The Coefficient of X suggests how influential variable X is to variable Y. Our Coefficient for the model is 0.0013463, which is very close to zero. The intercept for the model is 2.3398995, which suggests the value of Y when X equals zero.

```
Call:
lm(formula = Severity ~ `Temperature(F)`)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4363 -0.4288 -0.3830  0.5644  1.5904

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.3398995  0.0330773  70.740  < 2e-16 ***
`Temperature(F)` 0.0013463  0.0004982   2.702  0.00695 **
```

Table 2.

The summary table indicates the statistical significance of the values for both Coefficient and Intercept. We also changed the confidence level to 0.99 to make sure the values of Coefficient and Intercept are significant. The table shows us that the results fall within the range with a confidence level of 0.99.

```
                                  0.5 %      99.5 %
(Intercept)      2.2546162874 2.425182775
`Temperature(F)` 0.0000616556 0.002630883
```

Table 3.

To visually see the model, we plotted the model to see the linear relationship of our variables.

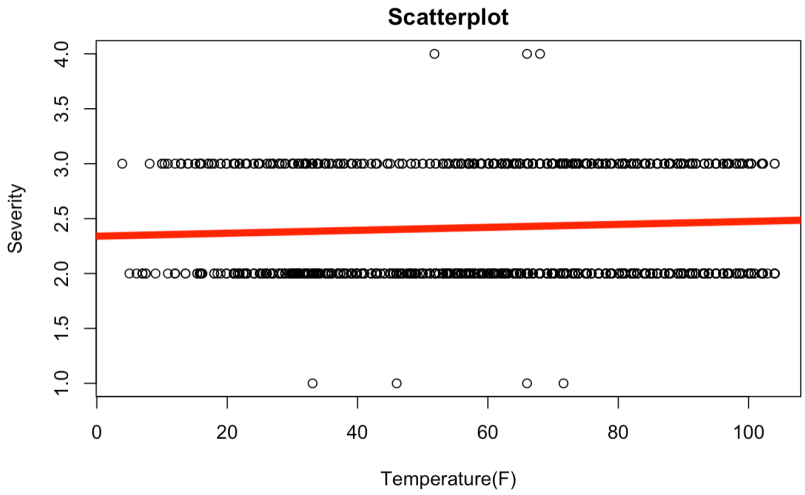

Figure 1.

From the linear regression model, we can see variable Temperature(F) does not have a significant relationship with variable Severity. To further test the result of the relationship, we implemented a correlation test for the variables. The correlation coefficient was 0.06, which is very close to

zero. It suggests that variable Temperature(F) has a very weak positive relationship with variable Severity.

```
df1 <-na.omit((US_Accident_Subset$`Temperature(F)`))
df2 <-na.omit((US_Accident_Subset$Severity))
cor(df1,df2)
```

```
[1] 0.0604453
```

Figure 2.

The reason for this result can be explained by the statistical uncertainty of variable Severity. We assumed that variable Severity is measured with a continuous value and that the values of variable Severity are normally distributed. In reality, the Severity score is mostly distributed unevenly on 2.0 and 3.0.

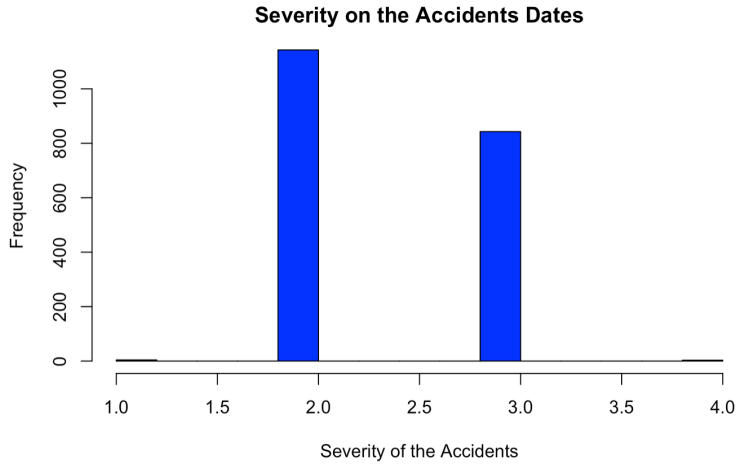

Figure 3.

In other words, the data distribution of variable Severity is not suitable for linear regression modeling. We also plotted a histogram for Temperature(F) to see how the data are distributed.

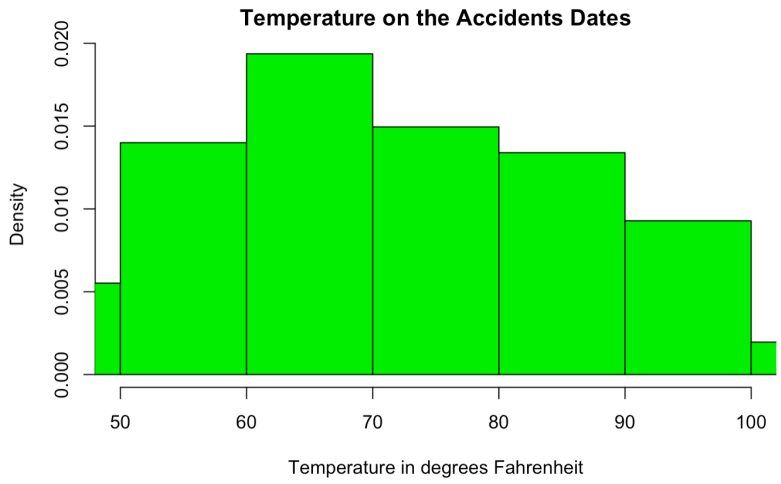

Figure 4.

As shown in the graph, variable Temperature(F) is a continuous variable with a right-skewed data distribution. It is a suitable variable for linear regression modeling. The major weakness of linear regression is that we assume data are normally distributed before we proceed, and also the assumption that two variables will have a linear relationship. Therefore, to better proceed with linear regression modeling, we should first check the correlation between the variables using a correlation function. Once we determine the correlation of the variables, we can then decide whether to use linear regression or other modeling algorithms for the research question. In addition, using a scatter plot and histogram will also improve the understanding of the dataset.

Ethical Issues

A number of ethical issues should be noted for the study. According to Davies [6] honesty in the data plays a critical role when evaluating research. This requires researchers to undersell instead of over-exaggerating findings. The present research does not show areas of dishonesty in the collection of data. A separate area that presents ethical issues is objectivity. This is where biases in the collection and design of the experiment occur [6]. The following dataset presented certain areas of bias in the severity and did not mention the individuals responsible for confirming the extremity of the accidents. Future studies should include not only the individual claiming the severity but the reasoning behind the evaluation. Other ethical issues include confidentiality and legality. The research did not comprise any of these ethical problems since the accidents did not contain any details of the individuals involved and the data was collected according to the laws and regulations of the state. Overall the present study did not comprise any major ethical issues.

Weaknesses of Research Question

A variety of weaknesses are present in the dataset. The first is the bias that is represented in the severity attribute from the dataset. The severity attribute describes how severe the impact on traffic is and presents some biases as to the decision of the severity. Some individuals may classify major traffic accidents as a 3 instead of a 4. Although the severity descriptions are clearly noted in the dataset, the ambiguity of extreme accidents may cause problems in the analysis. Another weakness in the research is the temperature range. With many different temperatures being present for each accident, it can be hard to pinpoint a specific temperature correlating to a specific level of severity. This problem is more evident in the analysis of the regression model. In terms of the dataset, one weakness is the specifics of the accident which are not noted. Here individuals can examine the severity by things such as the hours of the delay and injuries which play critical roles in understanding the accident. Future datasets should contain these attributes in order to provide evidence of the severity.

Appendix

The name of the person who obtained approval for a different dataset was Mustafa Barez (email: Mustafa.barez@mail.utoronto.ca) who confirmed this with Rohan Alexander (email: rohan.alexander@utoronto.ca) and the email for the approval was sent on February 21st, 2020.

References

[1] Demyttenaere, K., Bruffaerts, R., Posada-Villa, J., Gasquet, I., Kovess, V., Lepine, J., ... & Kikkawa, T. (2004). Prevalence, severity, and unmet need for treatment of mental disorders in the World Health Organization World Mental Health Surveys. *Jama*, *291*(21), 2581-2590.

[2] Peden, M. (2004). World report on road traffic injury prevention.

[3] Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, 2019.

[4] Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.

[5] RStudio Team (2019). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/

[6] Davies, A. (2017). Change in Ethics Violation Reporting: Honesty Is the Best Policy. *Radiologic technology*, *88*(3), 342-343.