Number of Deaths Caused by

Earthquakes Analysis

Introduction

Earthquakes pose serious risks to infrastructure, economic viability, and human lives across the world. In significant occurrences, buildings could be knocked over, homes could be destroyed, families may end up displaced with their valuables ruined, and hundreds to thousands of lives could be lost. As a result of the unpredictable nature of earthquakes, foretelling their impacts have proven to be difficult, especially pertaining to the death toll in each case. In an attempt to pinpoint which qualities in earthquakes lead to the largest number of lives lost, the questions that will be addressed are:

1. How do magnitude, intensity, focal depth, and houses destroyed impact the number of deaths due to earthquakes?
2. Does geography play a role in the number of deaths due to earthquakes?
3. How have the number of deaths due to earthquakes changed over time (everything being equal, did more or less people die from similar earthquakes in the past compared to the present)?

Determining where the strongest earthquakes occur and how the strength of an earthquake impacts human lives is crucial for many purposes. For the individual buying a house, knowing where these earthquakes are most prevalent is important in ensuring that they can live in a safe place where their risk of being impacted by an earthquake is minimized. If one does live in an area with a high prevalence of earthquakes, it is also critical to understand this so they can prepare for a possible earthquake with insurance and other protections to avoid catastrophic loss. For insurance companies, the answer to these questions would be useful in identifying where they should focus their earthquake insurance resources and which areas they should charge higher premiums to account for the higher probability of an earthquake occurring. From a societal perspective, understanding the impact of these earthquakes is important to ensure that people can work in the future to minimize the damage that these uncontrollable natural disasters cause. By understanding past historical data, society can more accurately predict where these earthquakes will occur in the future, and thus be more prepared for when they happen.

The Data

The data titled the Significant Earthquake Database was gathered by the National Center for Environmental Information (NCEI). This organization is a part of the National Oceanic and Atmospheric Administration of the U.S. Government. This data was collected for public knowledge and reference. The dataset covers earthquakes occurring from 2150 BC to the present, though the data that is being used for this study only incorporates earthquakes that include data on all of the explanatory variables being tested, thus limiting the dataset to only include earthquakes post-1877. However, record keeping has not always been what it is like today, and thus it is important to acknowledge that the data from hundreds of years ago is likely not as accurate as the data collected in modern times. With that being said, it is nevertheless important to include this data to see if and how earthquake trends have changed over time, especially since these types of events occur frequently.

The variables defined in this data set are date of the earthquake, location name, longitude and latitude. The data set defines focal depth of the earthquake, magnitude, modified Mercalli intensity, deaths, injuries, damage in dollars, and number of houses destroyed. Since some earthquakes are missing certain variables, the set is filtered to only include earthquakes with complete data. The missing data is random throughout the dataset, meaning there is no underlying cause for why some earthquakes are missing certain metrics. For the analysis, the number of deaths is the response variable, while magnitude, modified Mercalli intensity, focal depth, and number of houses destroyed are the explanatory variables. The modified Mercalli intensity is a scale that measures the intensity of an earthquake based on its effect on people, property and grounds damaged. This scale is a more useful measure of severity of an earthquake than the traditional magnitude scale, because it measures the impact on human lives and property rather than the intensity of seismic waves.

The Significant Earthquake database was collected from the Significant Earthquake catalog which was an expansion of a file originally created from the world map of significant earthquakes from 1900 to 1979. The data was collected and processed between 1979 and 1981 and it is updated regularly from sources such as: the U.S. Geological Survey; national and government databases and reports; earthquake

and tsunami catalogs; and a number of other sources including post-event reconnaissance reports, journal articles, newspapers, internet pages, email and other written documents.

Regression Analysis

To begin the regression analysis, a linear model was created with deaths as the response variable and magnitude, MMI intensity, houses destroyed, and focal depth as the predictor variables (Model #1). The fitted values vs. residual values were plotted for Model #1 to check its fit. There was an extraneous data point that noticeably affected the linear model (see Figure 1), so the Cook's Distance of the point was calculated to determine if it was an influential observation or an outlier.

The calculation determined that the data point was influential, since its Cook's Distance was in the 99th percentile. However, the point was not an outlier, because its z-score of -0.282 was well below the cutoff of ±3. The data point is not believed to be a mistake, but rather a very unusual value. This data point represents the 2008 Sichuan earthquake in China that had a very large number of houses destroyed. Since it was determined to be an influential point, it was removed from the data set and a new linear model (Model #2) excluding the data point was created.

In an attempt to potentially improve Model #2, a statistical criterion known as the Akaike Information Criterion (AIC) was used in order to determine which model provided the lowest AIC and therefore the best model for predicting the response variable. It was determined that removing focal depth as a predictor variable from the model was the best model for predicting the response variable. This was the method and thought process that led to Model #3.

After the removal of the influential data point from the dataset and focal depth from the model, the regression assumptions for Model #3 needed to be checked. The fitted values vs. residual values were plotted for Model #3. When this new plot was observed, it was once again seen that there was not a random distribution of the data points on the graph (see Figure 2). This indicated that the residual values had unequal variances. To account for this, a log function was used to transform the explanatory variable. After this transformation was applied, a fourth model (Model #4) was created and the fitted vs residual values were plotted (see Figure 3). There is a clear increase in the randomness of the points with the third model compared to the second model. Thus, Model #4 was viewed as a better model for the data.

The next assumption that needed to be checked was whether the errors were normally distributed. To accomplish this, a histogram of the residuals was produced for Model #4 (see Figure 4). When looking at this histogram, it appeared that the residuals are somewhat bimodal in their distribution. Though this is something that was weighed when producing the model, it was concluded that this distribution of the residuals was due to sampling variability. Because the data being analyzed only included earthquakes that had all data available, it could be reasonably assumed that this contributed to the lack of normality in the residuals. Because of this, the analysis continued with Model #4.

As a result of the earthquake data coming from a number of earthquakes that transpired at various points in the past, there was a heightened potential for the issue of correlated residuals. If the residuals from the model turned out to be correlated, any inferences made with the model would be invalidated. To check for this issue, a Durbin-Watson test was performed. After calculating a Durbin-Watson test statistic with a p-value well beyond any reasonable significance level, any potential for the residuals in the model being correlated was negated.

To check for multicollinearity among the predictor variables used in the model, the R-squared values and variance inflation factors (VIFs) of each predictor variable based on the other predictor variables was calculated. Since the R-squared value for each predictor variable was below 0.2, well below the reasonable threshold of 0.9, and the VIF calculated for each predictor variable was below 1.5, below the reasonable threshold of 10, there was no reason to believe that there was any significant correlation between the predictor variables used in the model.

After these steps, a final model was established (see Figure 5).

Discussion of the Analysis

The number of deaths was able to be predicted from earthquake magnitude, intensity, and number of houses destroyed through the formula listed above in Figure 5. It was concluded that focal depth was not a strong enough predictor to be included in the model, thus leaving the model to only include those first three predictors.

The number of deaths due to earthquakes was not able to be predicted based upon geography. Although the data includes the location of where these earthquakes occurred, the technology that would have been necessary to graph these points across a map was unavailable. Given this sophisticated technology, it would have been possible to at least come to a general conclusion for this question.

Regarding the last question, the lack of adequate statistical and technical knowledge prevented a firm conclusion from being reached. However, through the research that has been conducted throughout this analysis and contextual intuition, a reasonable hypothesis is that the number of deaths due to earthquakes of similar magnitude has likely decreased over time. The main factor behind this is founded upon improved house-building technology and regulations. Building regulations have improved over time as structural technology has improved. As these houses are built with improved structural innovations, it can be expected that the number of houses that are destroyed in similarly powerful earthquakes would decline. A great example of this is the influential data point that was determined to be in the data, the 2008 Sichuan Earthquake in China. Reginald DesRoches, a professor of civil and environmental engineering at Georgia Tech, pointed out how China's lack of adequate building regulations caused the widespread destruction of houses found in that data point. "China didn't get an adequate seismic design code until following the big earthquake they had in 1976," DesRoches said. "If the buildings were older and built prior to that [1976 earthquake], chances are they weren't built for adequate earthquake forces" (Bryner). Because of this, it's evident that in comparing earthquake damage over time, a reasonable conclusion to come to would be that the number of deaths should continue to decrease as innovations in structural technology continue to be made.

For others doing a similar analysis in the future, it would be helpful to have a uniform collection of data. The dataset used for this analysis had many incomplete observations, leading to a majority of the data points being filtered out of the final dataset. However, with a uniform collection of data, one could more accurately perform this analysis with more data points and less variability due to observations being deleted from the set. In addition, one analysis that was not performed in this report but may be helpful is studying the number of deaths based on population density. It was discovered after completing the analysis that population density of the area affected by the earthquake might be an influential factor in the number of deaths. The data set used in this analysis did not contain this information, so it was not feasible to include this in the analysis. However, for others doing a similar analysis, the inclusion of population density in the regression study could be a valuable addition for reliably modeling the number of deaths.

The presence of incomplete data points and the differing standards for data collection in different countries were the main drawbacks to the data collection methods. The difficulty of obtaining accurate data points for earthquakes that transpired years in the past paired with a wide range of geographical locations is most likely the main cause of these drawbacks. While the analysis made here does provide adequate information on the impact that earthquake magnitude, intensity, and number of houses destroyed have on the number of deaths from an earthquake, data collection methods that do not contain these drawbacks would increase the accuracy and utility of the analysis.

Conclusion

Earthquakes have devastated countries all over the world for centuries and will continue to have large impacts on society in the future. Although this analysis helps discern what factors affect the number of deaths due to earthquakes, much of the solution to this is out of society's control beyond improving housing structures. Hopefully, with improvements to both the protection methods against earthquakes and the technology to measure their impacts, analyses of earthquake impact like this will improve. These improvements and increased regulation of building codes will continue to decrease the impact caused by earthquakes, not only causing less damage to property and society, but saving lives as well.

Bibliography

"Magnitude and Risk to Human Life." *FutureLearn*. Cardiff University. Web. 2017.

National Geophysical Data Center / World Data Service (NGDC/WDS): NCEI/WDS Global Significant Earthquake Database. NOAA National Centers for Environmental Information.

Bryner, Jeanna. “Why the China Quake Was So Devastating.” *LiveScience*, 15 May 2008, web.archive.org/web/20080517233751/news.yahoo.com/s/livescience/whythechinaquakewassodevastating.

Appendix

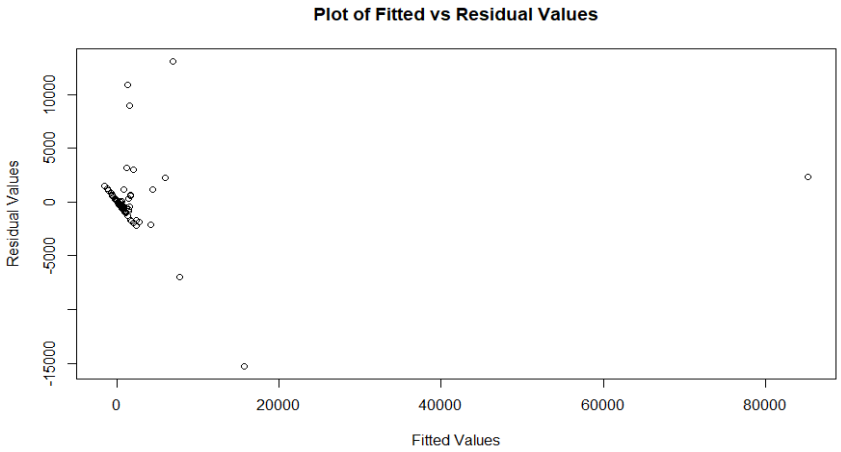

Figure 1: Plot of Fitted vs Residual Values for Model #1

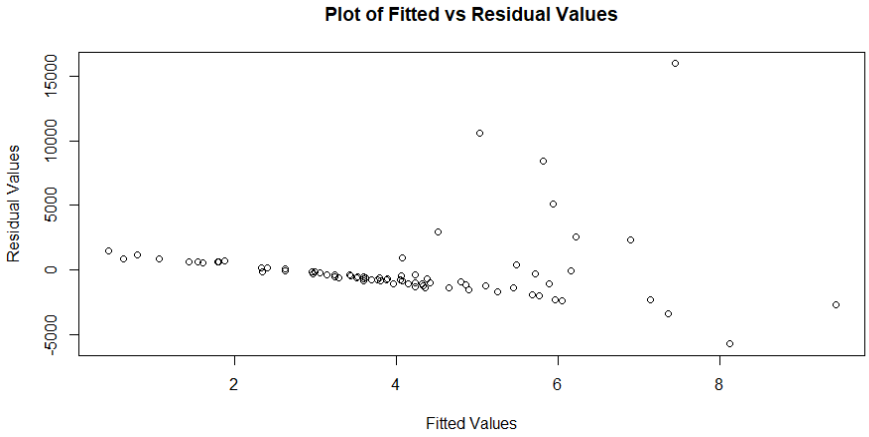

Figure 2: Plot of Fitted vs Residual Values for Model #3

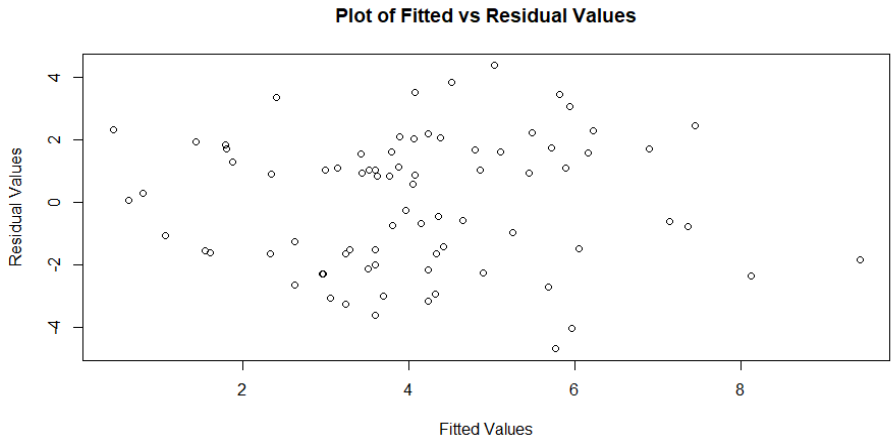

Figure 3: Plot of Fitted vs Residual Values for Model #4

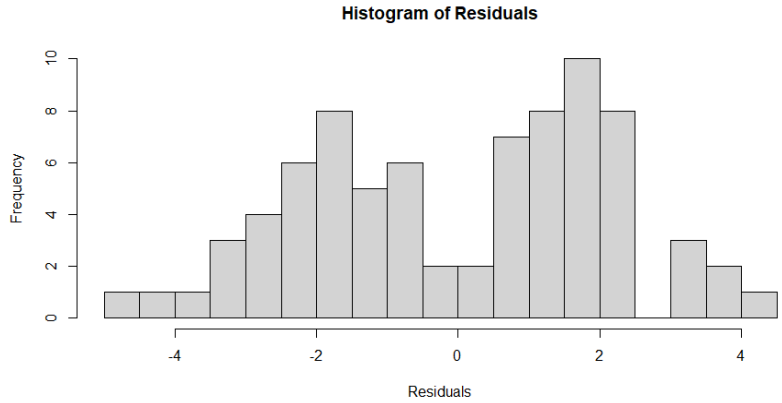

Figure 4: Histogram of the Residuals for Model #4

Number of Deaths = -8.078 + (.9076 * Magnitude) + (.7222 * MMI Intensity) +

(3.878e-06 * Houses Destroyed)

Figure 5: Model #4