Earthquake Death Toll Linear Regression

MAYA MACIEL-SEIDMAN

2024-03-02

```
knitr::opts_chunk$set(echo = TRUE)
# Load dplyr:
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Load tidyverse:
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.4
## v ggplot2   3.5.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
# Load ggplot2:
library(ggplot2)
# Load modelsummary:
library(modelsummary)
```

```
## Version 2.0.0 of `modelsummary`, to be released soon, will introduce a
##   breaking change: The default table-drawing package will be `tinytable`
##   instead of `kableExtra`. All currently supported table-drawing packages
##   will continue to be supported for the foreseeable future, including
##   ‘kableExtra‘, ‘gt‘, ‘huxtable‘, ‘flextable, and ‘DT‘.
##
##   You can always call the ‘config_modelsummary()‘ function to change the
##   default table-drawing package in persistent fashion. To try ‘tinytable‘
##   now:
##
##   config_modelsummary(factory_default = ’tinytable’)
##
##   To set the default back to ‘kableExtra‘:
##
##   config_modelsummary(factory_default = ’kableExtra’)
```

Motivation

I am interested in exploring the research question: Is there a linear relationship between an earthquake’s magnitude and its death toll? The dependent variable of interest is the earthquake’s death toll. I think that variation in an earthquake’s death toll could be explained by an earthquake’s magnitude since earthquakes with greater magnitude are more severe, and more severe earthquake’s should, in theory, have higher death tolls since they are more destructive. Ideally, the data that would help me examine this is the death toll and magnitude of all recorded global earthquakes. I will be using NOAA’s significant Earthquake Database from 2150 BC to October 16, 2017, which contains this exact data. It was downloaded as a csv from benjiao’s GitHub page at this link: https://github.com/benjiao/significant-earthquakes/blob/master/earthquakes.csv.

Null hypothesis: A linear relationship does not exist between an earthquake’s magnitude and its death toll. Alternate hypothesis: A linear relationship exists between an earthquake’s magnitude and its death toll.

Data Preparation

```r
# Read in the data:
earthquakes <- read.csv("./earthquakes.csv")

# Summarize the data:
summary(earthquakes)
```

```
##        X            damage            day            deaths
##  Min.   :   0   Min.   :     1   Min.   : 1.00   Min.   :     1
##  1st Qu.:1490   1st Qu.:    10   1st Qu.: 8.00   1st Qu.:     4
##  Median :2981   Median :    43   Median :16.00   Median :    26
##  Mean   :2981   Mean   :  2392   Mean   :15.75   Mean   :  3988
##  3rd Qu.:4472   3rd Qu.:   200   3rd Qu.:23.00   3rd Qu.:   400
##  Max.   :5962   Max.   :799000   Max.   :31.00   Max.   :830000
##                 NA’s   :4856     NA’s   :556     NA’s   :4019
##   focal_depth         hour       houses_damaged    houses_destroyed
##  Min.   :  0.00   Min.   : 0.0   Min.   :      1   Min.   :     0.01
##  1st Qu.: 11.00   1st Qu.: 5.0   1st Qu.:     77   1st Qu.:     3.26
##  Median : 27.00   Median :11.0   Median :    600   Median :    20.00
##  Mean   : 41.94   Mean   :11.3   Mean   :  19222   Mean   :  1745.66
##  3rd Qu.: 40.00   3rd Qu.:17.0   3rd Qu.:   4500   3rd Qu.:   170.40
##  Max.   :678.00   Max.   :23.0   Max.   :5360000   Max.   :220000.00
##  NA’s   :2945     NA’s   :2027   NA’s   :5247      NA’s   :5510
##    location           magnitude         minute          mmi_int
##  Length:5963        Min.   :1.600   Min.   : 0.00   Min.   : 2.000
##  Class :character   1st Qu.:5.700   1st Qu.:14.00   1st Qu.: 7.000
##  Mode  :character   Median :6.500   Median :29.00   Median : 8.000
##                     Mean   :6.489   Mean   :28.78   Mean   : 8.447
##                     3rd Qu.:7.300   3rd Qu.:43.00   3rd Qu.:10.000
##                     Max.   :9.500   Max.   :59.00   Max.   :12.000
##                     NA’s   :1789    NA’s   :2232    NA’s   :3325
##      month              name               second            year
##  Min.   : 1.000   Length:5963        Min.   : 0.10   Min.   :-2150
##  1st Qu.: 4.000   Class :character   1st Qu.:15.20   1st Qu.: 1812
##  Median : 7.000   Mode  :character   Median :30.00   Median : 1925
##  Mean   : 6.505                      Mean   :30.16   Mean   : 1798
##  3rd Qu.: 9.000                      3rd Qu.:45.00   3rd Qu.: 1984
##  Max.   :12.000                      Max.   :59.90   Max.   : 2017
##  NA’s   :405                         NA’s   :3335
```

```r
head(earthquakes)
```

```
##   X damage day deaths focal_depth hour houses_damaged houses_destroyed
## 1 0     NA  NA     NA          NA   NA              NA               NA
## 2 1     NA  NA     NA          NA   NA              NA               NA
## 3 2     NA  NA      1          18   NA              NA               NA
## 4 3     NA  NA     NA          NA   NA              NA               NA
## 5 4     NA  NA     NA          NA   NA              NA               NA
## 6 5     NA  NA     NA          NA   NA              NA               NA
##                                         location magnitude minute mmi_int
## 1 POINT (35.5000000000000000 31.1000000000000014)       7.3     NA      NA
## 2 POINT (35.7999999999999972 35.6829999999999998)        NA     NA      10
## 3 POINT (58.2000000000000028 38.0000000000000000)       7.1     NA      10
## 4 POINT (25.3999999999999986 36.3999999999999986)        NA     NA      NA
## 5 POINT (35.2999999999999972 31.5000000000000000)        NA     NA      10
## 6 POINT (25.5000000000000000 35.5000000000000000)        NA     NA      10
##   month                              name second  year
## 1    NA     JORDAN:  BAB-A-DARAA,AL-KARAK     NA -2150
## 2    NA                    SYRIA:  UGARIT     NA -2000
## 3    NA                  TURKMENISTAN:  W     NA -2000
## 4    NA GREECE:  THERA ISLAND (SANTORINI)     NA -1610
## 5    NA          ISRAEL:  ARIHA (JERICHO)     NA -1566
## 6    NA              ITALY:  LACUS CIMINI     NA -1450
```

```r
# Data wrangling:
# Select only the columns with the variables of interest:
earthquakes <- earthquakes %>% select(deaths, magnitude)
# Get rid of NA values:
earthquakes <- earthquakes %>% drop_na()
dim(earthquakes)
```

```
## [1] 1588    2
```

```r
head(earthquakes)
```

```
##   deaths magnitude
## 1      1       7.1
## 2   2500       7.1
## 3    760       7.0
## 4   6000       7.0
## 5     13       5.5
## 6 260000       7.5
```

```r
typeof(earthquakes$deaths)
```

```
## [1] "integer"
```

```r
typeof(earthquakes$magnitude)
```

```
## [1] "double"
```

Summary of the data: This data is a record of every known earthquake from 2150 BC to October 16, 2017. Each observation is an earthquake. The variables include earthquake characteristics including location, magnitude, death toll, year, focal depth, number of houses destroyed, and more. Since I am only interested in looking at the relationship between an earthquake's magnitude and death toll, I selected only those columns. Since this dataset includes very historic earthquakes, it has some NA values since not all earthquake characteristics could be recorded for some historic time frames. I omitted data with NA values so I was left with a dataset containing 1588 observations of earthquakes with the 2 variables magnitude and deaths. Deaths is an integer and magnitude is a double, which are both numeric, allowing me to perfom an OLS regression.

Conduct

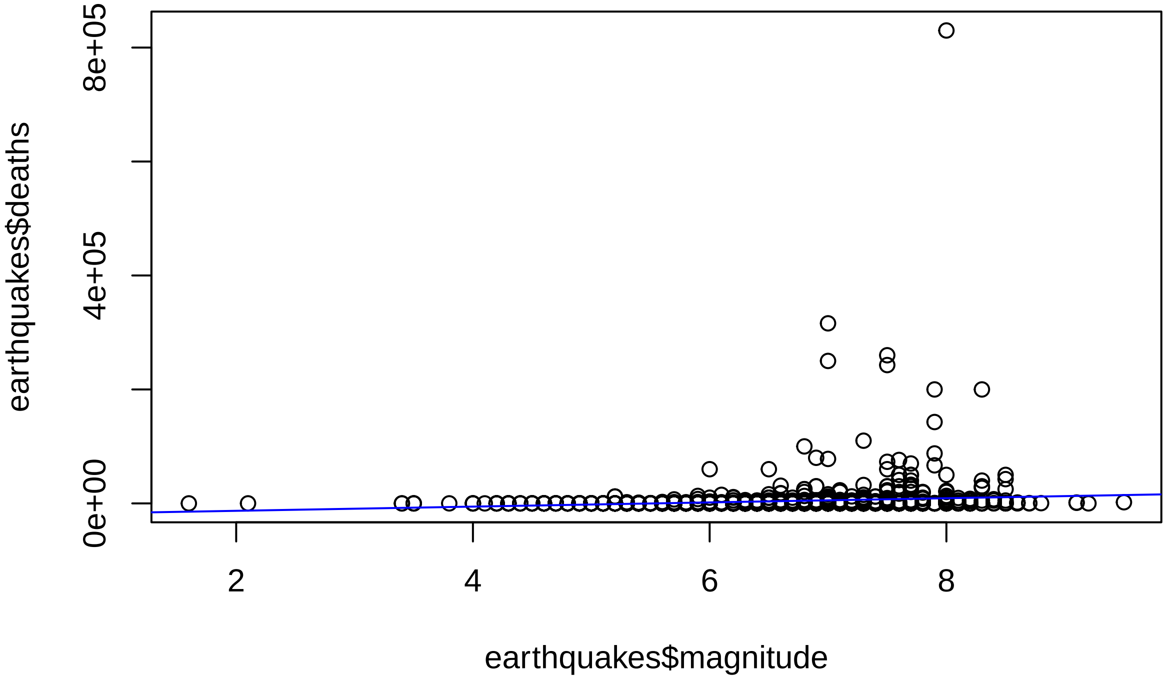
```r
# Run OLS regression:
lm_earthquakes <- lm(formula = deaths~magnitude, data=earthquakes)
```

```r
# Table of model results:
modelsummary(lm_earthquakes, stars = TRUE)
```

```r
# Figure of model results:
# Plot scatter plot of original data with the regression line:
plot(earthquakes$magnitude,earthquakes$deaths)
abline(lm(formula=deaths~magnitude, data=earthquakes), col="blue")
```

	(1)
(Intercept)	−20 404.556***
	(4568.996)
magnitude	3686.261***
	(699.860)
Num.Obs.	1588
R2	0.017
R2 Adj.	0.017
AIC	36 917.8
BIC	36 933.9
Log.Lik.	−18 455.886
F	27.743
RMSE	26 988.23

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001



Interpretations

According to my regression table, an increase of 1 in an earthquake's magnitude, on average, is associated with a 3,686.261 increase in an earthquake's death toll. The stars next to this number on my regression table indicate that this result is statistically significant, meaning it is unlikely that we would observe an association between an earthquake's magnitude and death toll due to chance. However, when we look at the goodness of fit statistics from the regression table, it is clear that this association might not be linear. The

R2 and R2 Adj. values, both 0.017, show that only 1.7% of the variation in an earthquake's death toll can be explained by an OLS regression model. The AIC and BIC are 36917.8 and 36933.9, respectively. These are large values, which means that the OLS regression model is not the best fit for this data. Finally, the RMSE value is 26988.23, indicating a high average error in the model predicting an earthquake's death toll given its magnitude. Additionally, looking at the scatter plot of the original data with the regression line, the regression line does not seem to fit the data well. There are many points far above the line.

These results mean that while there is a statistically significant association between an earthquake's magnitude and death toll, an OLS regression is not the best model to explain the variation in an earthquake's death toll.

Based on these results, I would want to conduct future analysis with a different type of model. While there seems to not be a linear relationship between an earthquake's magnitude and its death toll, there might be another type of relationship; perhaps an exponential relationship. This would help me advance future research on this topic because if there is an exponential relationship between magnitude and death toll, then we would be able to understand a threshold of earthquake magnitude where the death toll is much higher than other magnitudes. We also then could have a different scale for earthquake death toll which is separate from the magnitude scale, which only describes severity in terms of destruction.

Diagnostics

Based on the results from my model, OLS seems to be not appropriate. The data is appropriate to use to investigate my research question, but it does not follow a linear relationship. My main cause for concern about using an OLS regression model include the very low R2 and R2 Adj. values and very high AIC, BIC, and RMSE values from my regression table. These goodness of fit statistics for my model convey to me that my model is mis-specified for this data because it does not fit the data well for the reasons discussed in my Interpretations section above.