Earthquake Death Toll GLM

Maya Maciel-Seidman

2024-03-22

Motivation

I am interested in exploring the research question: Which characteristic out of an earthquake's magnitude, focal depth, and number of houses destroyed has the greatest effect on whether or not an earthquake's fatality?

The dependent variables of interest are the earthquake's total death toll and the earthquake's fatality (whether or not it has a death toll). I think that variation in an earthquake's fatality could be explained by an earthquake's magnitude since earthquakes with greater magnitude are more severe, and more severe earthquake's should, in theory, be more likely to have death tolls since they are more destructive. An earthquake's focal depth could explain variation in an earthquake's fatality since earthquake's with focal depths closer to the surface mean that the earthquake is centered closer to the surface and will have more surface destruction than an earthquake centered deeper in the Earth. The number of houses destroyed could explain variation in an earthquake's fatality because if a house is destroyed with people inside of it, then those people are at greater risk of death.

Ideally, the data that would help me examine this is the presence of a death toll, magnitude, focal depth, and number of houses destroyed 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: There is no significant difference between the effects of an earthquake's magnitude, focal depth, and number of houses destroyed on its fatality.
Alternate hypothesis: There is a significant difference between the effects of an earthquake's magnitude, focal depth, and number of houses destroyed on its fatality.

A logit regression could be a good fit for this problem because the dependent variable, the earthquake's fatality, is binary. This is because the earthquake is either fatal (has a death toll) or not fatal (does not have a death toll). Further, a poisson regression could be a good fit to predict an earthquake's total death toll since this dependent variable is a count variable which starts at 0 and has an unbounded maximum.

Data Preparation

```
# 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
```

```
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
```

```
# Data wrangling:
# Select only the columns with the variables of interest:
earthquakes <- earthquakes %>% select(deaths, magnitude, houses_destroyed, focal_depth)
# Convert NA deaths to 0 since 0 deaths are recorded as NA:
earthquakes[is.na(earthquakes)] <- 0
# Get rid of all other NA values:
earthquakes <- earthquakes %>% drop_na()
# Add fatality binary dummy variable:
earthquakes <- earthquakes %>% mutate(fatal=ifelse(deaths>=1, 1, 0))

# Look at structure of data:
dim(earthquakes)
```

```
## [1] 5963    5
```

```
head(earthquakes)
```

```
##   deaths magnitude houses_destroyed focal_depth fatal
## 1      0       7.3                0           0     0
## 2      0       0.0                0           0     0
## 3      1       7.1                0          18     1
## 4      0       0.0                0           0     0
## 5      0       0.0                0           0     0
## 6      0       0.0                0           0     0
```

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

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

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

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

```
typeof(earthquakes$houses_destroyed)
```

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

```
typeof(earthquakes$focal_depth)
```

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

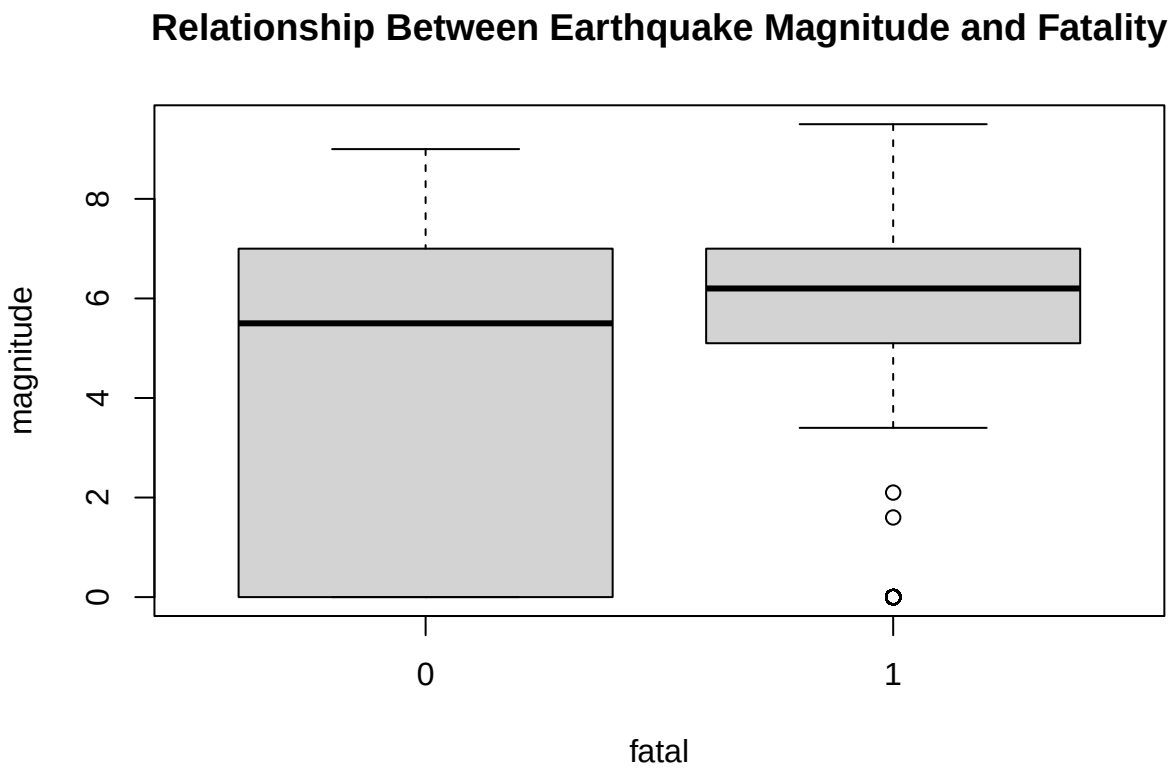
```
typeof(earthquakes$fatal)
```

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

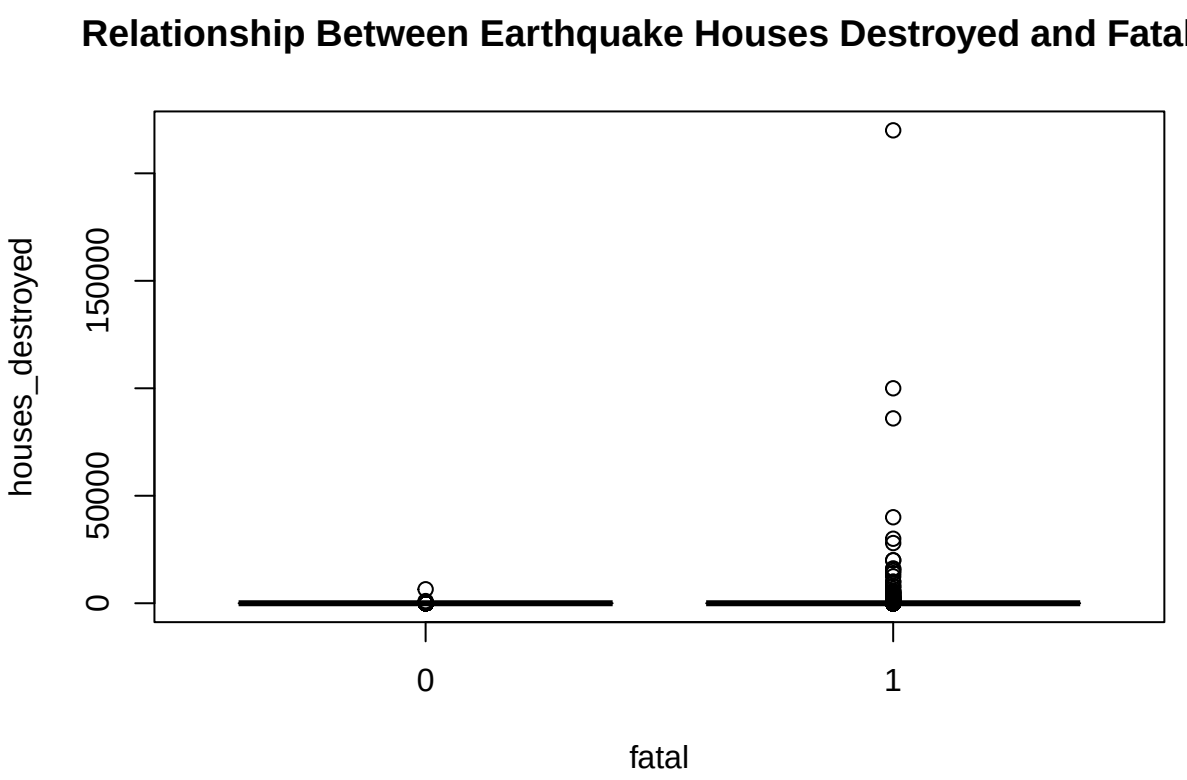
Summary of the data: This data is a record of every known earthquake from 2150 BC to October 16, 2017. Each observation is an earthquake. The variables include earthquake characteristics including location, magnitude, death toll, year, focal depth, number of houses destroyed, and more. Since I am only interested in looking at the relationship between an earthquake's magnitude, number of houses destroyed, focal depth, and death toll, I selected only those columns. Additionally, I created a binary dummy variable `fatal` to describe whether or not the earthquake has a death toll, which will help me evaluate the earthquake's fatality with a logit regression. Whereas the total death toll will stay numeric to be evaluated with a poisson regression. I also converted the NAs in the death toll variable to 0 since earthquakes with 0 deaths were recorded as NA. I omitted the rest of the NAs in other columns since they were not representing 0 counts for other variables. I was left with a dataset containing 5963 observations of earthquakes with the 5 variables. All variables are doubles, which are numeric, allowing me to perform the logit regression and poisson regression, which I discussed above.

Conduct

```
# Create box plots to look at the relationship between features of the
# earthquake and whether or not it is fatal:
boxplot(magnitude~fatal, data=earthquakes, main="Relationship Between Earthquake Magnitude and Fatality
```



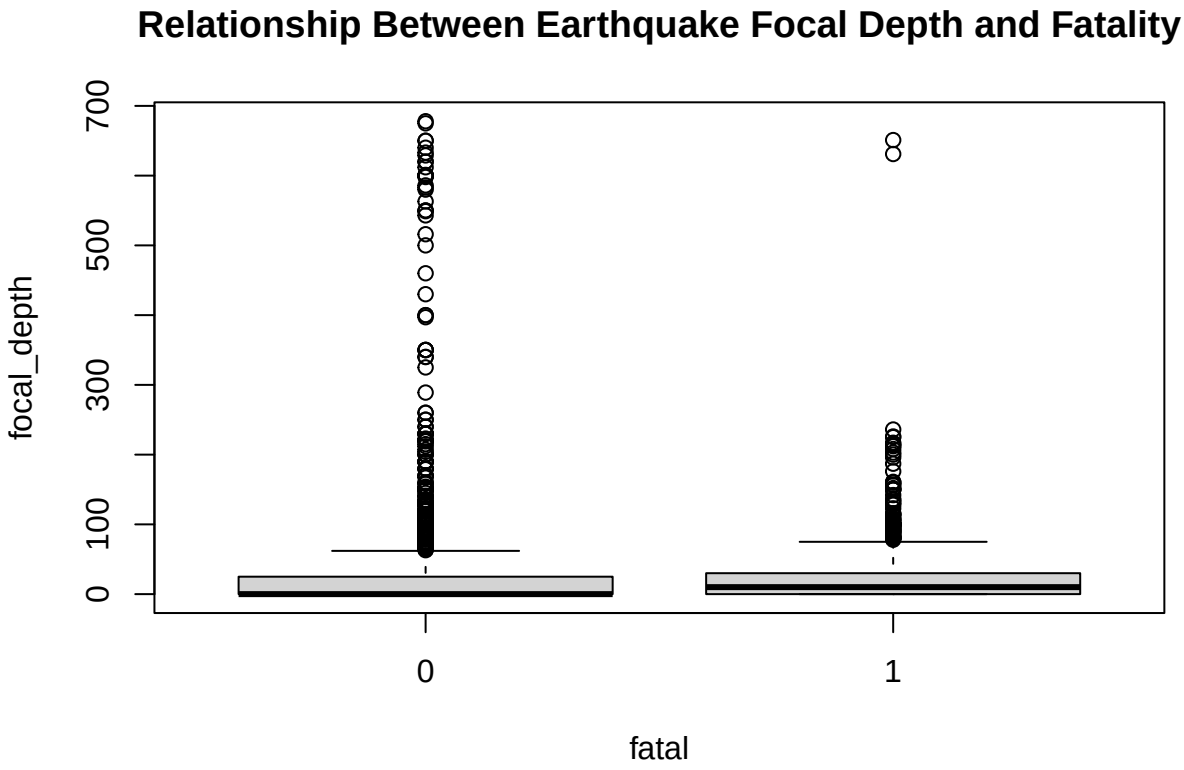

```
boxplot(houses_destroyed~fatal, data=earthquakes, main="Relationship Between Earthquake Houses Destroye
```




```
boxplot(focal_depth~fatal, data=earthquakes, main="Relationship Between Earthquake Focal Depth and Fatal
```

	(1)	(2)
(Intercept)	0.273***	−1.299***
	(0.015)	(0.055)
magnitude	1.128***	0.120***
	(0.011)	(0.010)
Num.Obs.	5963	5963
AIC	7366.3	7366.3
BIC	7379.6	7379.6
Log.Lik.	−3681.128	−3681.128
F	156.763	156.763
RMSE	0.46	0.46

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

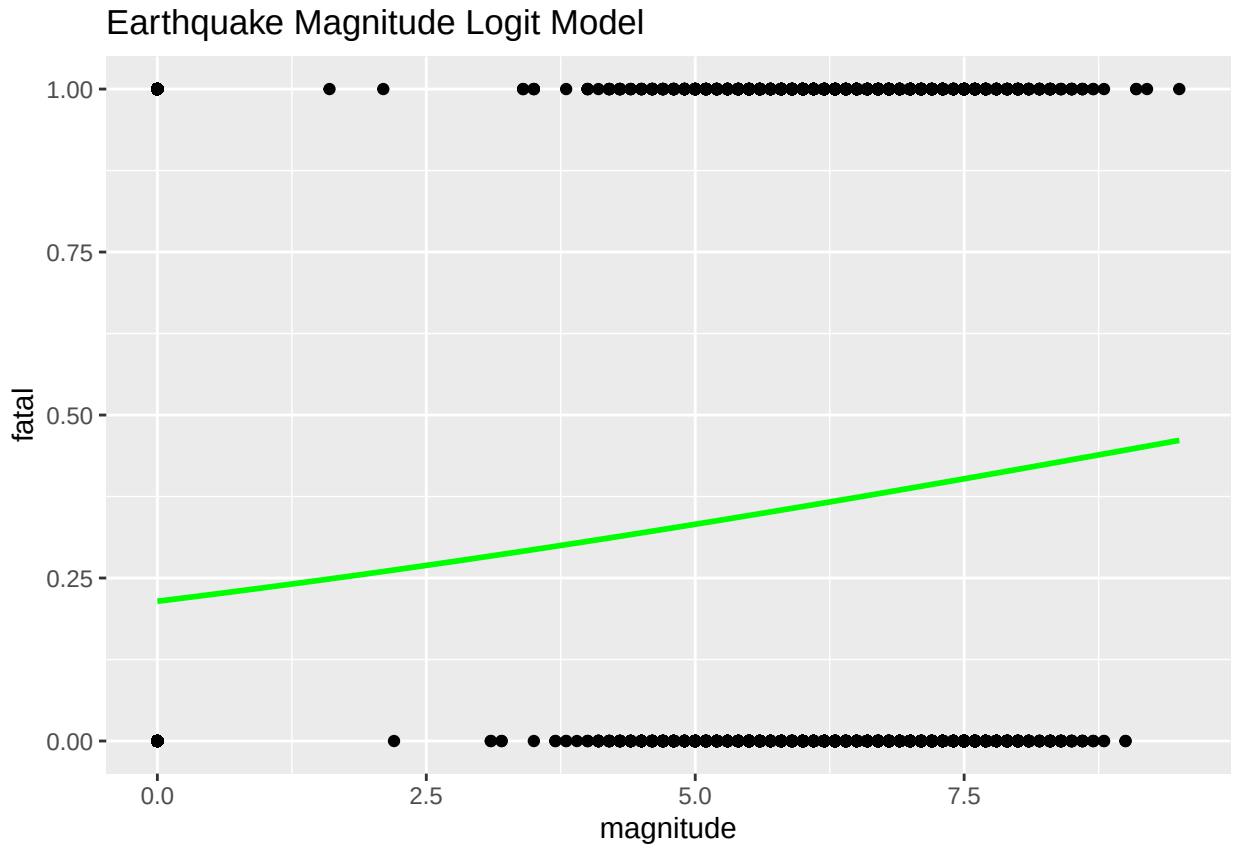


Based on the box plots, it seems to be a slight difference in magnitude between fatal and nonfatal earthquakes. However, there does not seem to be a difference in the number of houses destroyed between fatal and nonfatal earthquakes. Additionally, there seems to be a very small (almost trivial) difference in the focal depth between fatal and nonfatal earthquakes.

```
# Run logit regression for magnitude:
earthquake_model_magnitude <- glm(fatal~magnitude, data=earthquakes, family="binomial")
magnitude_logit <- list(earthquake_model_magnitude, earthquake_model_magnitude)
# Obtain model summary for magnitude:
modelsummary::modelsummary(magnitude_logit, exponentiate=c(TRUE, FALSE), stars=TRUE)
```

```
# Create visualization for magnitude logit model:
ggplot(earthquakes, aes(x=magnitude, y=fatal)) + geom_point() + stat_smooth(method="glm", color="green"
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# Run logit regression for houses_destroyed:
earthquake_model_houses <- glm(fatal~houses_destroyed, data=earthquakes, family="binomial")
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
houses_logit <- list(earthquake_model_houses, earthquake_model_houses)
# Obtain model summary for houses_destroyed:
modelsummary::modelsummary(houses_logit, exponentiate=c(TRUE, FALSE), stars=TRUE)
```

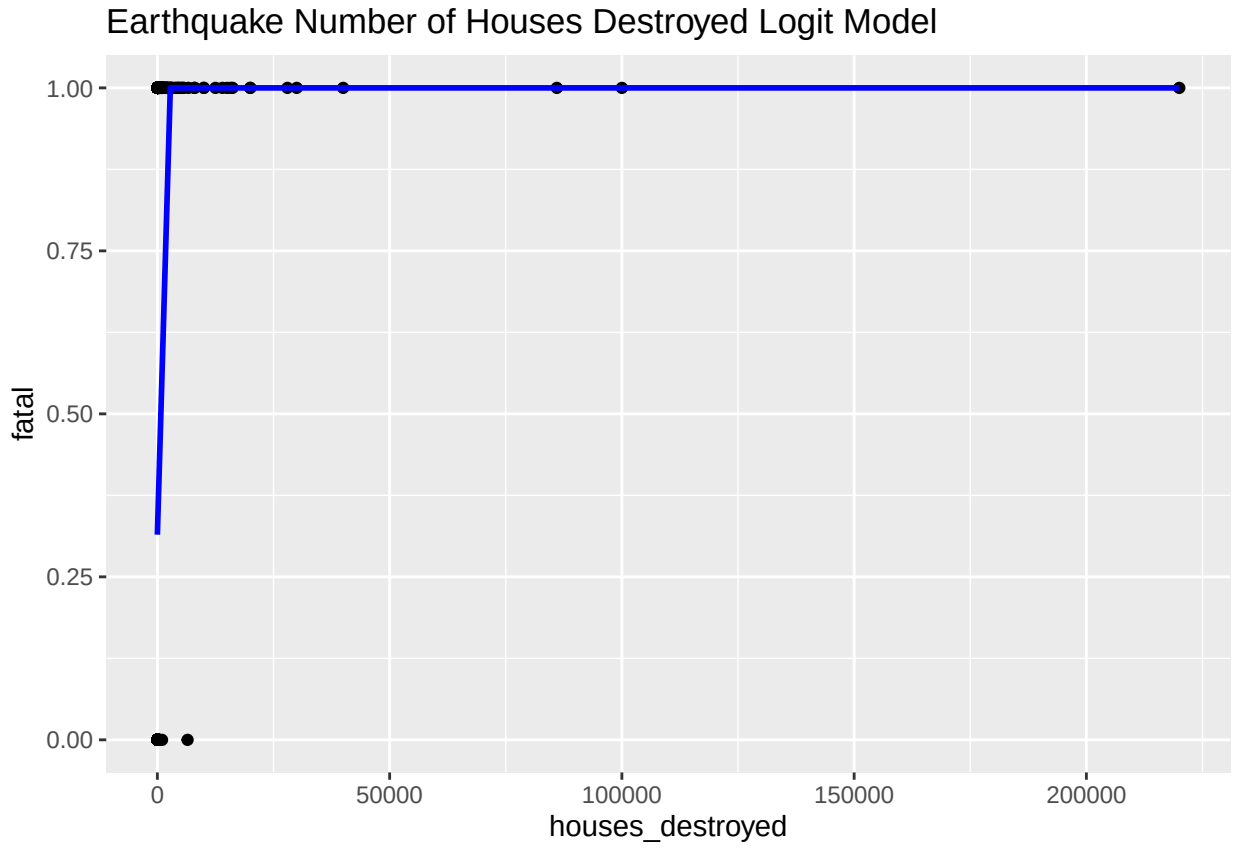
```
# Create visualization for houses_destroyed logit model:
ggplot(earthquakes, aes(x=houses_destroyed, y=fatal)) + geom_point() + stat_smooth(method="glm", color=
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

	(1)	(2)
(Intercept)	0.459***	−0.778***
	(0.013)	(0.028)
houses_destroyed	1.004***	0.004***
	(0.001)	(0.001)
Num.Obs.	5963	5963
AIC	7364.6	7364.6
BIC	7378.0	7378.0
Log.Lik.	−3680.322	−3680.322
F	36.793	36.793
RMSE	0.46	0.46

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



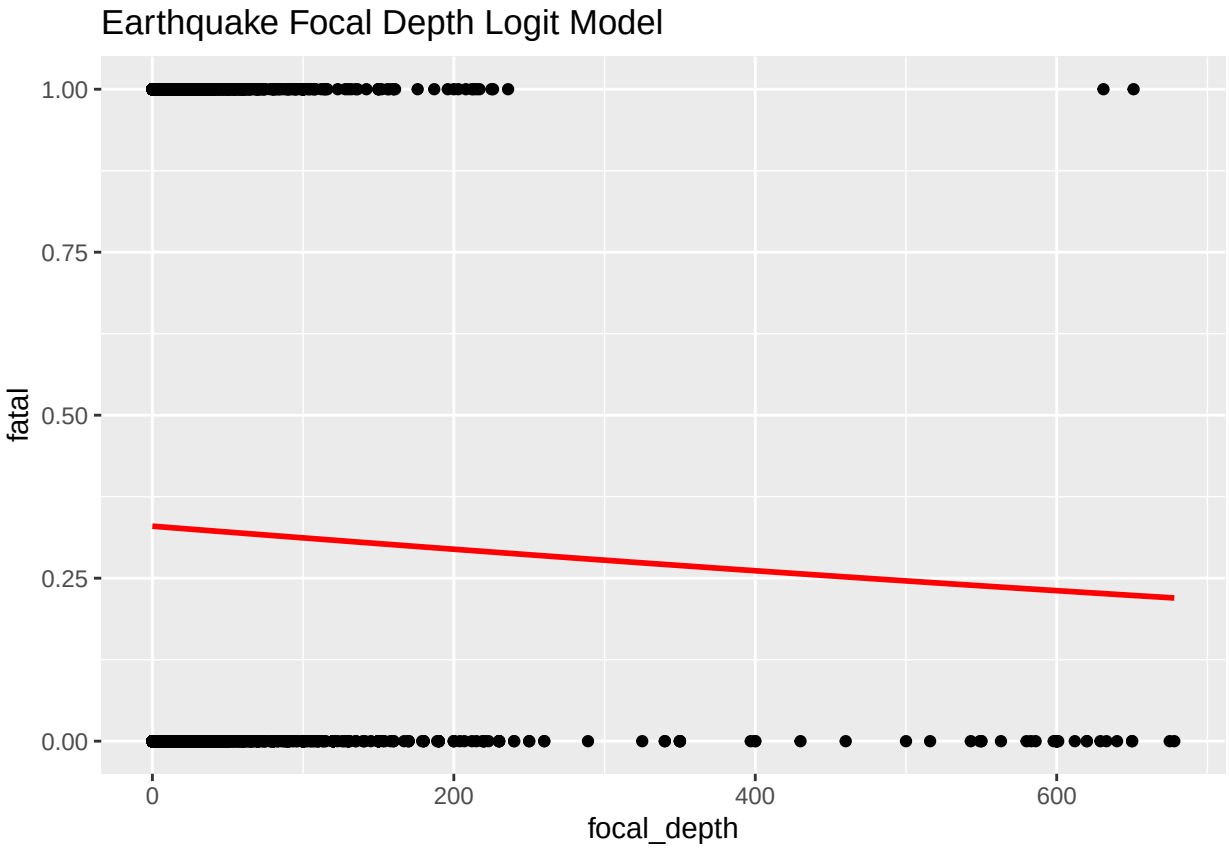

```
# Run logit regression for focal_depth:
earthquake_model_focal_depth <- glm(fatal~focal_depth, data=earthquakes, family="binomial")
focal_depth_logit <- list(earthquake_model_focal_depth, earthquake_model_focal_depth)
# Obtain model summary for focal_depth:
modelsummary::modelsummary(focal_depth_logit, exponentiate=c(TRUE, FALSE), stars=TRUE)
```

```
# Create visualization for focal_depth logit model:
ggplot(earthquakes, aes(x=focal_depth, y=fatal)) + geom_point() + stat_smooth(method="glm", color="red"
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

	(1)	(2)
(Intercept)	0.492***	−0.709***
	(0.015)	(0.030)
focal_depth	0.999	−0.001
	(0.001)	(0.001)
Num.Obs.	5963	5963
AIC	7530.7	7530.7
BIC	7544.1	7544.1
Log.Lik.	−3763.341	−3763.341
F	2.283	2.283
RMSE	0.47	0.47

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



Formula for Magnitude Logit Regression: $P(fatal = 1) = \frac{1}{1+e^{-(-1.299+0.120X_m)}}$

Formula for Houses Destroyed Logit Regression: $P(fatal = 1) = \frac{1}{1+e^{-(-0.778+0.004X_h)}}$

Formula for Focal Depth Logit Regression: $P(fatal = 1) = \frac{1}{1+e^{-(-0.709-0.001X_f)}}$

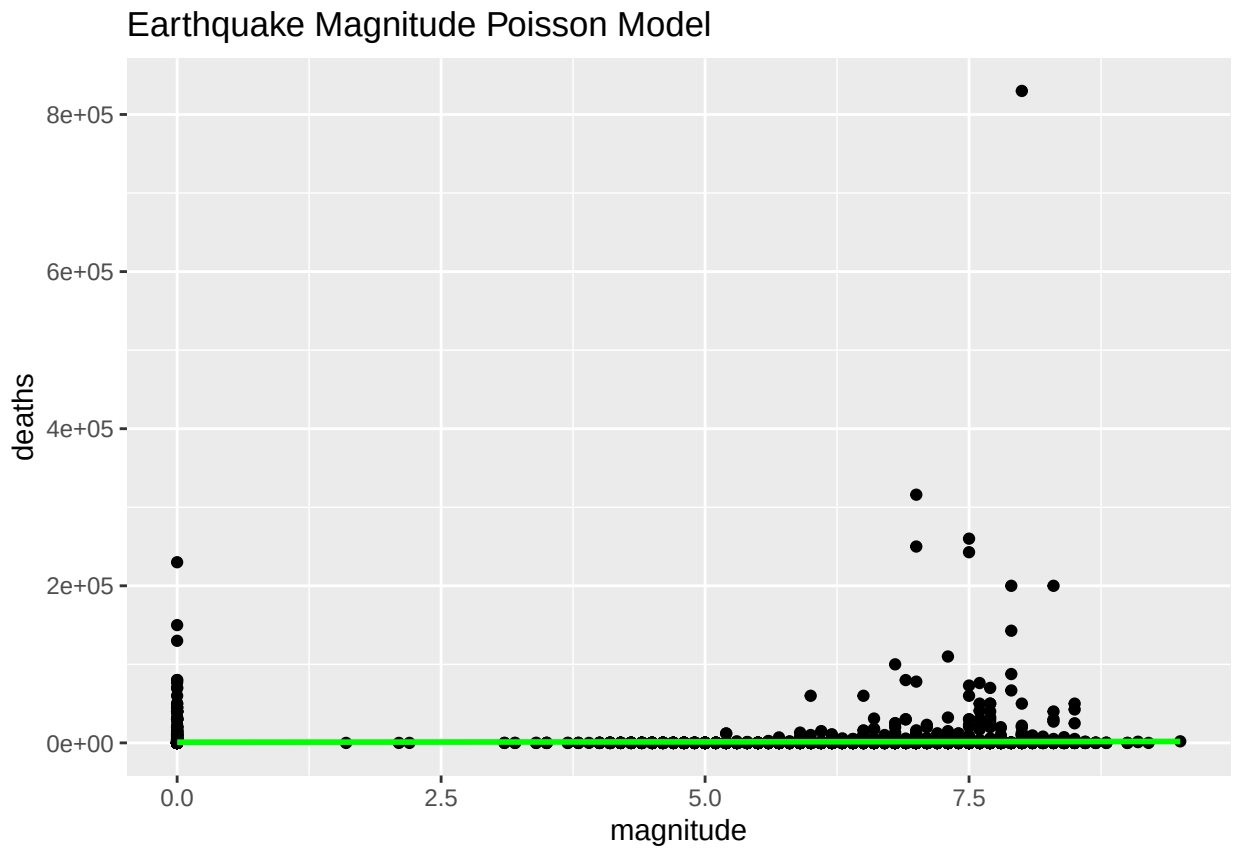
```
# Run poisson regression for magnitude:
magnitude_poisson <- glm(deaths~magnitude, family=poisson, data=earthquakes)
# Obtain model summary for magnitude:
modelsummary::modelsummary(magnitude_poisson, stars=TRUE, exponentiate=TRUE)
```

	(1)
(Intercept)	904.708***
	(0.681)
magnitude	1.077***
	(0.000)
Num.Obs.	5963
AIC	60 021 705.5
BIC	60 021 718.9
Log.Lik.	−30 010 850.771
F	340 811.946
RMSE	15 026.40

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

```
# Create visualization for magnitude poisson model:
ggplot(earthquakes, aes(x=magnitude, y=deaths)) + geom_point() + stat_smooth(method="glm", color="green
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



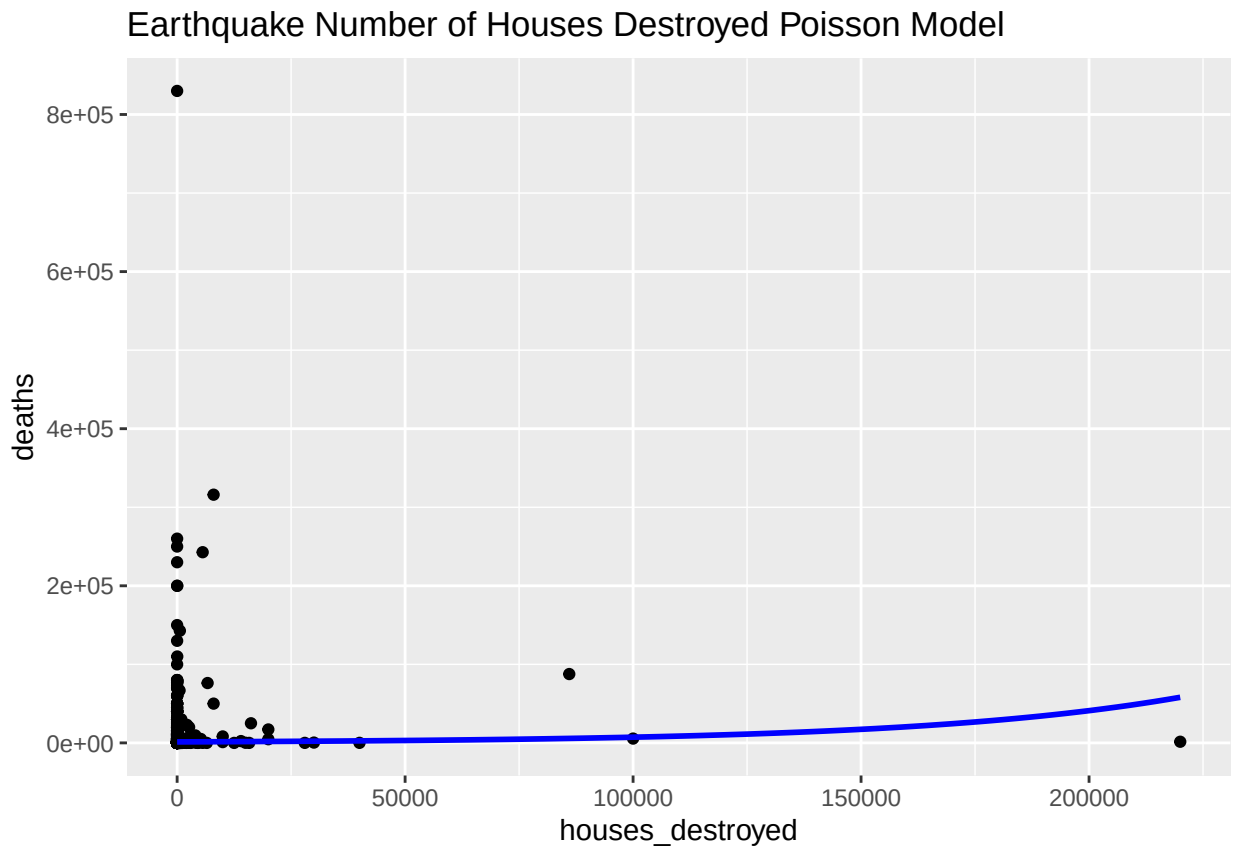

```
# Run poisson regression for houses_destroyed:
houses_destroyed_poisson <- glm(deaths~houses_destroyed, family=poisson, data=earthquakes)
# Obtain model summary for houses_destroyed:
modelsummary::modelsummary(houses_destroyed_poisson, stars=TRUE, exponentiate=TRUE)
```

	(1)
(Intercept)	1287.193***
	(0.464)
houses_destroyed	1.000***
	(0.000)
Num.Obs.	5963
AIC	60 034 967.0
BIC	60 034 980.4
Log.Lik.	−30 017 481.510
F	871 576.144
RMSE	15 042.29

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

```
# Create visualization for houses_destroyed poisson model:
ggplot(earthquakes, aes(x=houses_destroyed, y=deaths)) + geom_point() + stat_smooth(method="glm", color=
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



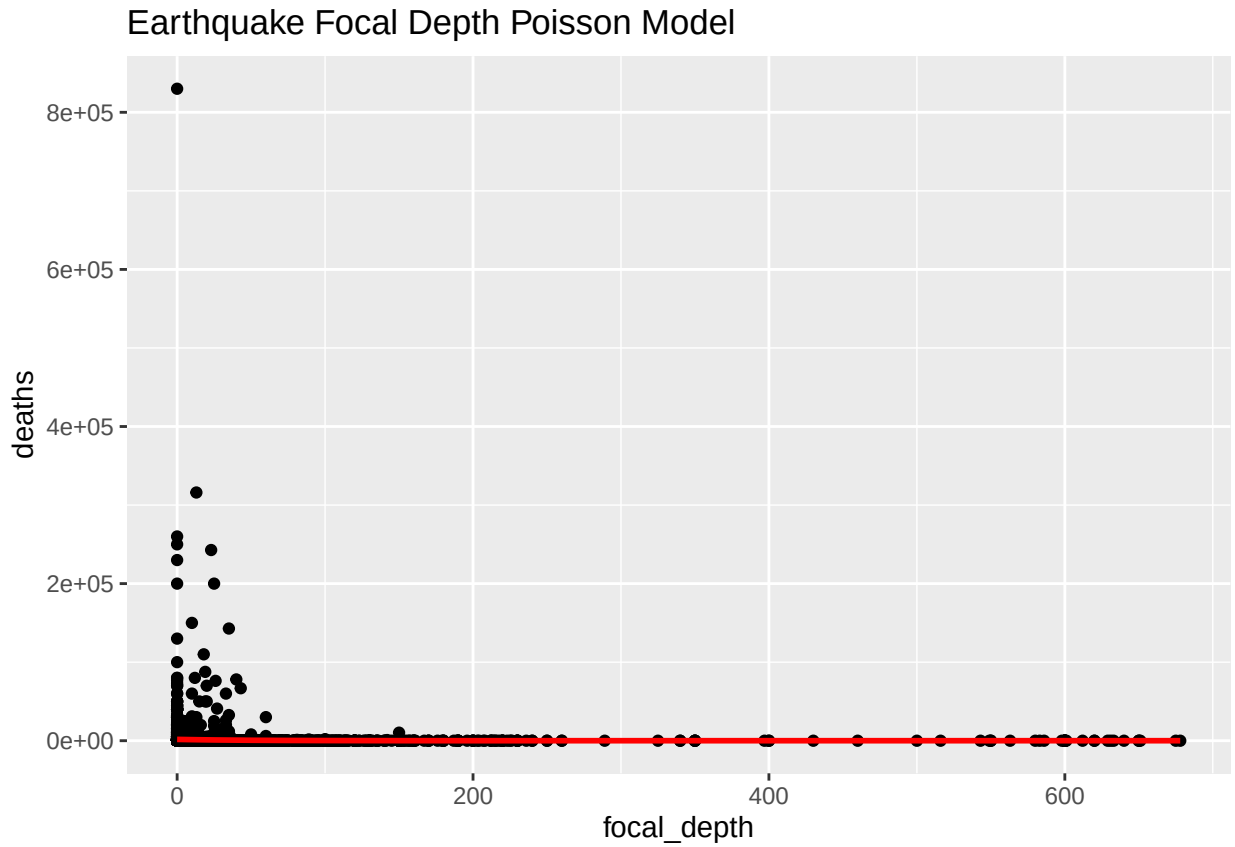

```
# Run poisson regression for focal_depth:
focal_depth_poisson <- glm(deaths~focal_depth, family=poisson, data=earthquakes)
# Obtain model summary for focal_depth:
modelsummary::modelsummary(focal_depth_poisson, stars=TRUE, exponentiate=TRUE)
```

	(1)
(Intercept)	1668.538***
	(0.680)
focal_depth	0.979***
	(0.000)
Num.Obs.	5963
AIC	59 114 726.6
BIC	59 114 740.0
Log.Lik.	−29 557 361.310
F	754 606.880
RMSE	15 023.60

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

```
# Create visualization for focal_depth poisson model:
ggplot(earthquakes, aes(x=focal_depth, y=deaths)) + geom_point() + stat_smooth(method="glm", color="red
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Interpretations

Magnitude logit regression:
Based on the model summary table, magnitude is statistically significant with an alpha of 0.001. Every

one-unit increase in an earthquake's magnitude is associated with the odds of an earthquake being fatal increasing by 12%. The RMSE for this model is 0.46, which means that the model is not very accurate since an average error of 0.46 is not good when the values of the dependent variable are either 0 or 1. Additionally, this is confirmed by looking at the plot for this model compared to the actual data. We can see that the model does not fit the data well.

Houses destroyed logit regression:
Looking at the model summary table, the number of houses destroyed by an earthquake is statistically significant with an alpha of 0.001. Every one-unit increase in the number of houses that an earthquake destroys is associated with the odds of an earthquake being fatal increasing by 0.4%. The RMSE for this model is also 0.46, meaning that this model is also not very accurate since an average error of 0.46 is not good when the values of the dependent variable are either 0 or 1. Additionally, we also see the model not fitting the data very well in the graph.

Focal depth logit regression:
The model summary table for this model shows that an earthquake's focal depth is not statistically significant. However, every one-unit increase in an earthquake's focal depth is associated with the odds of an earthquake being fatal decreasing by 0.1%. The RMSe for this model is 0.47, also indicating that this model is not very accurate. Looking at the plot also confirms that this model does not fit the data very well.

Magnitude poisson regression:
The model summary table for this model shows that every one-unit increase in an earthquake's magnitude is associated with the expected death toll being 1.077 times more than otherwise, all else equal. This is is statistically significant with an alpha of 0.001. The RMSE for this model is 15026.40, which is a bit high for an average error of deaths. This means that this model is not very accurate. Looking at the plot, we can also see that this model does not fit the data very well.

Houses destroyed poisson regression:
Looking at the model summary table for this model, every one-unit increase in the number of houses destroyed by an earthquake is associated with the expected death toll being 1.000 times more than otherwise, all else equal. This means that the number of houses destroyed is associated with neither an increase nor decrease in an earthquake's death toll. However, this is statistically significant with an alpha of 0.001. The RMSE for this model is similarly 15042.29, which is also high for an average error of deaths and not very accurate. The plot for this model also does not fit the data very well.

Focal depth poisson regression:
The model summary for this table shows that every one-unit increase in an earthquake's focal depth is associated with the death toll being 0.979 times less than otherwise, all else equal. This is statistically significant with an alpha of 0.001. The RMSE for this model is 15023.60, similar to the other Poisson models. This is also high for an average number of deaths and not very accurate. The plot for this model also shows this model not fitting the data very well.

Diagnostics

The results of the models mean that an earthquake's magnitude has the greatest effect on whether or not an earthquake is fatal, when compared to an earthquake's number of houses destroyed and focal depth. In terms of death toll, magnitude also has the greatest effect. All of these predictors were statistically significant in all of the models, meaning that they all have statistically significant effects on an earthquake's fatality.

Neither the logit regression nor the poisson regression fit the data for any of the predictors very well. This leads me to believe that either some other type of model could be used for future analysis or that there isn't enough variation in the data between nonfatal and fatal earthquakes for these predictors to have large enough impacts on an earthquake's fatality. To investigate this further in the future, I would be interested in gathering data on where the earthquakes occurred and how far they were from major population centers because if most of these earthquakes occurred far from large populations, then they couldn't possibly be

fatal. If this is in fact the case, that could explain why the models don't fit the data very well. Taking location into account would further research on the topic of earthquake fatality.

I gained new knowledge that I didn't have during problem set 3. In problem set 3, I only investigated whether or not there was a linear relationship between an earthquake's magnitude and its death toll. In this problem set, I was able to analyze relationships between a binary variable (earthquake fatality) and continuous variables, which I was unable to do in problem set 3. Additionally, I was able to use a poisson regression in this problem set to predict the effects of predictor variables on a count variable (earthquake death toll), which allowed me to go beyond my investigation in problem set 3.

In problem set 3, I came to the conclusion that an OLS regression was not the most appropriate model to evaluate my research question. I now know that other regressions (logit and poisson) were more suited based on the natures of the dependent variables that I investigated in this problem set.