Matriculation Number: 2614867B

Valérie BLANCH

DMAS ASSESSMENT 3: LAB REPORT

INTRODUCTION

The following analysis uses a dataset containing 23,741 observations about earthquakes over a number of years. It contains the following 12 variables:

Name	Type	Description
id	Numeric	ID of record
lat	Numeric	Latitude of earthquake (degrees)
long	Numeric	Longitude of earthquake (degrees)
dist	Numeric	Distance travelled by earthquake in a particular direction (km)
depth	Numeric	Depth of earthquake (km)
md	Numeric	Magnitude of earthquake (the duration of seismic wave-train (Md)
richter	Numeric	Intensity of earthquake (Richter)
mw	Numeric	Moment magnitude scale value of earthquake (Mw)
ms	Numeric	Surface-wave magnitude scale value of earthquake (Ms)
mb	Numeric	Body-wave magnitude value
country	Character	Country of earthquake
direction	Character	Direction of earthquake

We are particularly interested in the following questions:

a) What is the largest magnitude value for each observation (denoted *xm*)? And is the average value of *xm* different from 4.1?

b) Is there a difference in the moment magnitude scale value of an earthquake (*mw*) between countries in which the earthquakes occurred, on average?

c) How can we build a regression model with *richter* as a response variable?

d) How can we build a logistic regression model from the (modified) *richter* variable?

e) How does a simple logistic regression model from the *richter* variable compare to the previous model?

We will start with an exploratory analysis to gain a deeper insight into the data and verify assumptions, then we will test it with different statistical tools to answer our questions of interest.

EXPLORATORY ANALYSIS

We will focus on 11 variables for the exploration of the dataset, excluding the *id* column, that will not be used.

The variables have been explored according to their data types. First, we will study the categorical variables *country* and *direction*, then the numeric ones.

Categorical variables

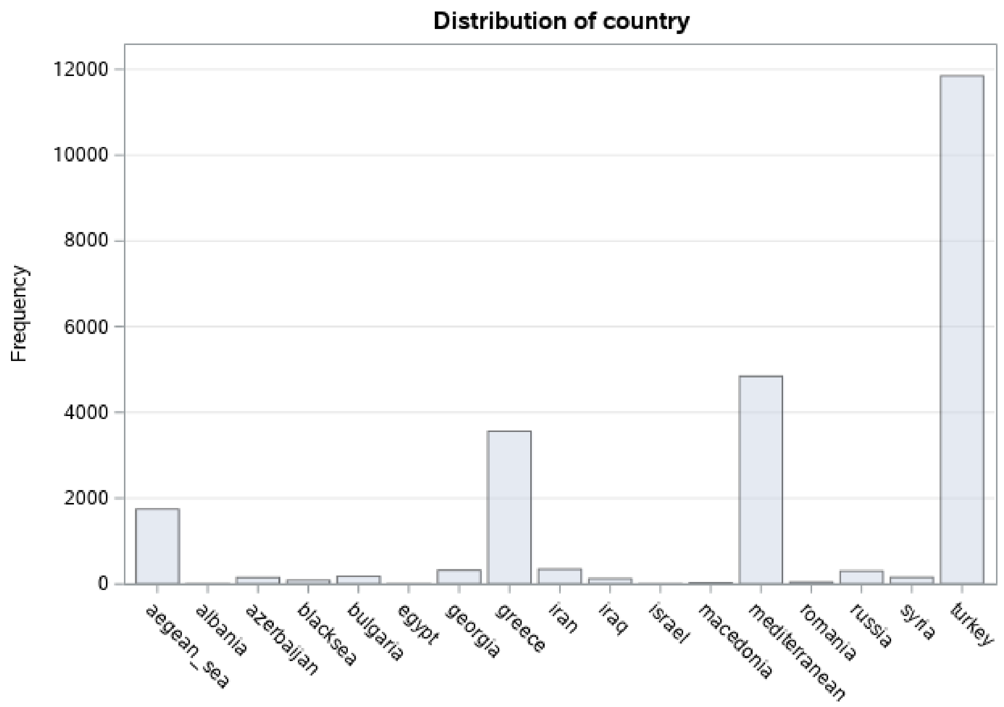
Categorical Variables Frequency Analysis

The FREQ Procedure

country	Frequency	Percent	Cumulative Frequency	Cumulative Percent
aegean_sea	1748	7.36	1748	7.36
albania	2	0.01	1750	7.37
azerbaijan	150	0.63	1900	8.00
blacksea	90	0.38	1990	8.38
bulgaria	176	0.74	2166	9.12
egypt	2	0.01	2168	9.13
georgia	322	1.36	2490	10.49
greece	3560	15.00	6050	25.48
iran	346	1.46	6396	26.94
iraq	122	0.51	6518	27.45
israel	1	0.00	6519	27.46
macedonia	28	0.12	6547	27.58
mediterranean	4843	20.40	11390	47.98
romania	44	0.19	11434	48.16
russia	303	1.28	11737	49.44
syria	154	0.65	11891	50.09
turkey	11850	49.91	23741	100.00

The frequency table above corresponds to the variable *country*. It shows the number of observation per “country” (some levels are not countries), with proportions and percentages. This variable regroups 17 levels.

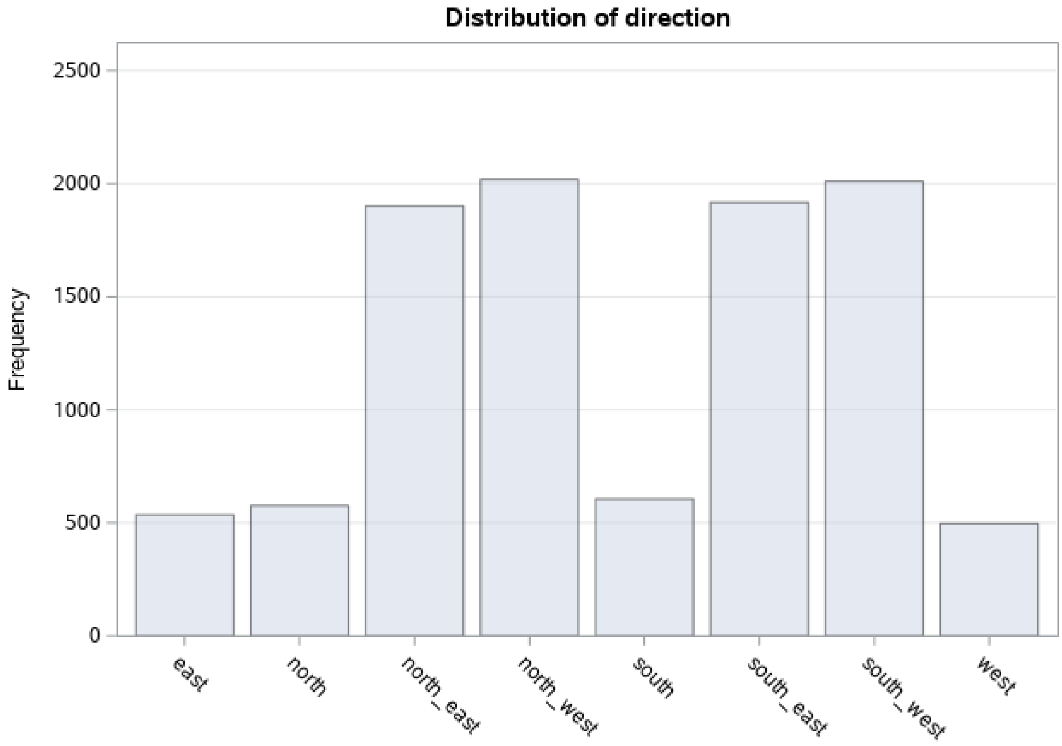
The mode of the distribution is Turkey, meaning it’s the most affected by earthquakes in the dataset, with 11,850 observations regrouping almost half of them (49.91%).



The predominance of Turkey when it comes to earthquakes is even more visible if we draw a frequency bar plot.

direction	Frequency	Percent	Cumulative Frequency	Cumulative Percent
east	536	5.33	536	5.33
north	576	5.72	1112	11.05
north_east	1901	18.89	3013	29.94
north_west	2019	20.07	5032	50.01
south	605	6.01	5637	56.02
south_east	1917	19.05	7554	75.07
south_west	2011	19.99	9565	95.06
west	497	4.94	10062	100.00
Frequency Missing = 13679				

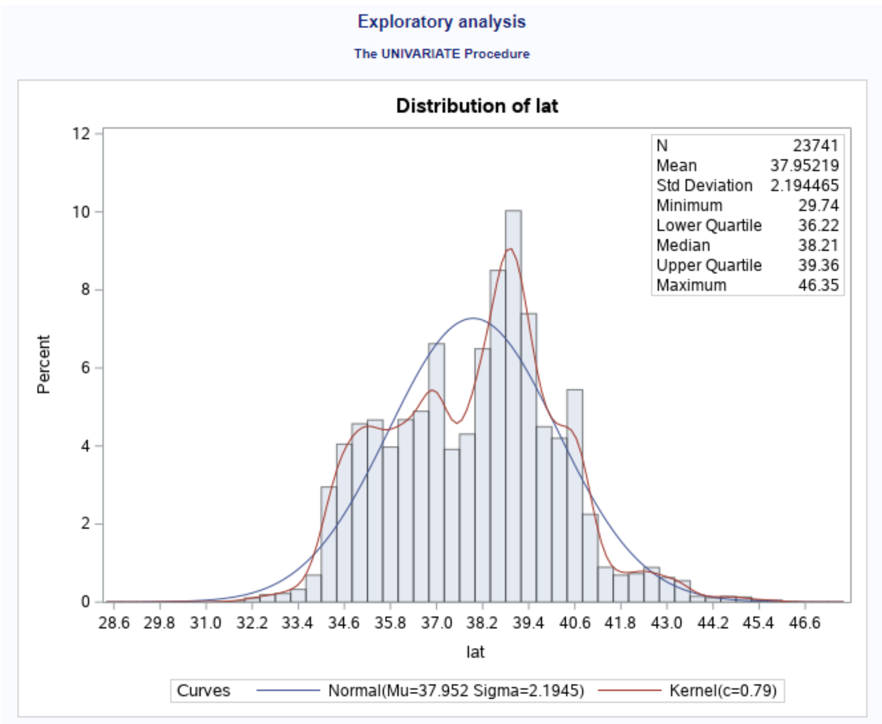
The second categorical variable *direction* records the direction of the earthquakes. Here the software informs us that 13,679 values for that particular variable are missing, which is more than half of the total number of observations.



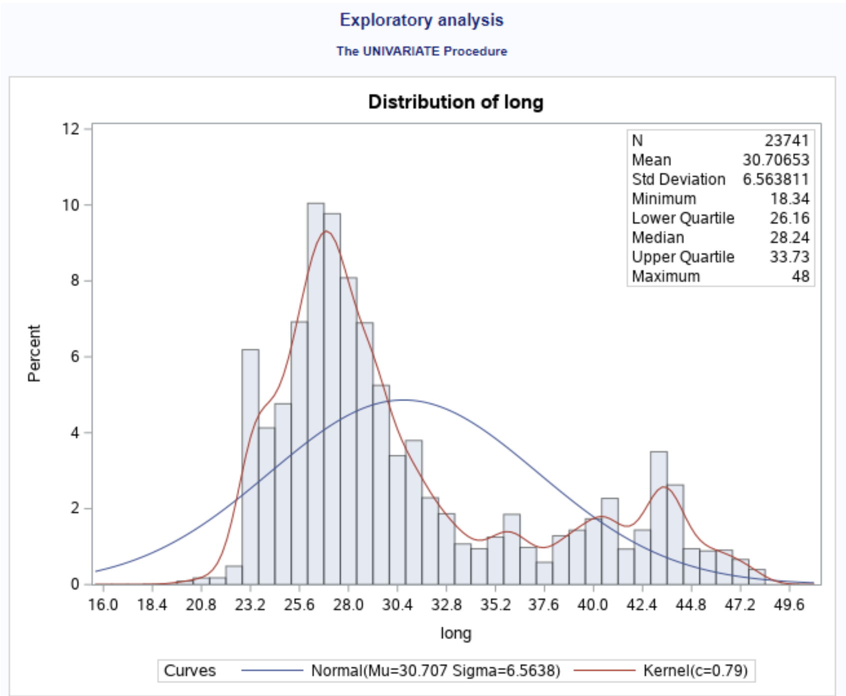
The mode of the distribution is *north_west*, with 2019 observations.

<u>Numeric Variables:</u>

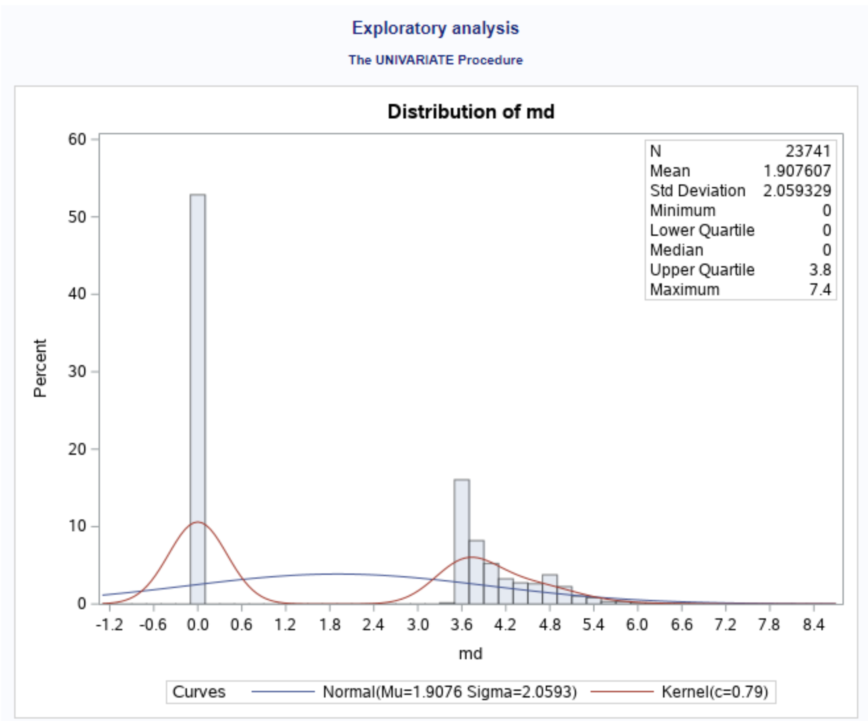
For numeric variables, we have produced histograms and summary statistics to assess their distributions.



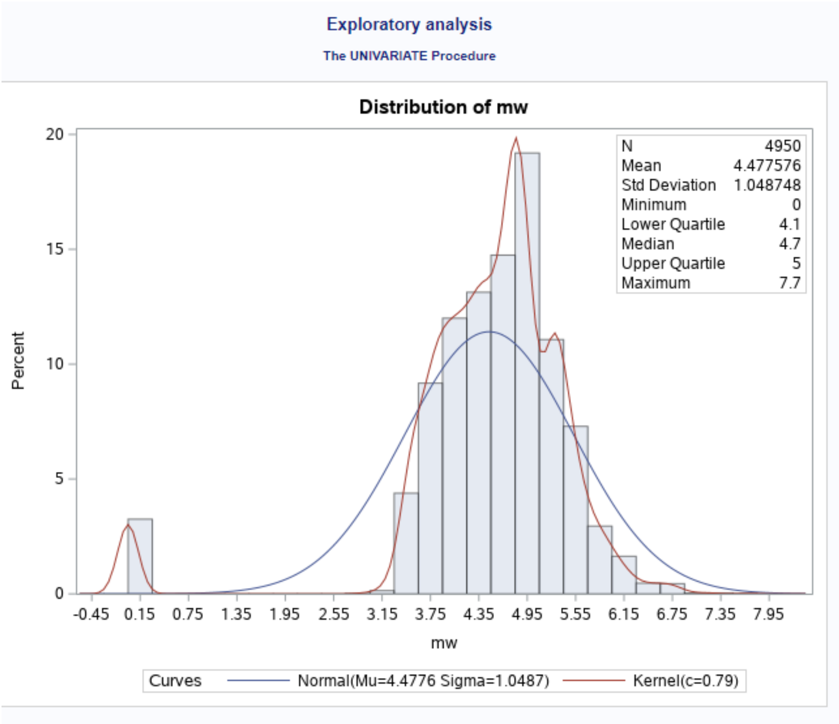
- Shape: the variable *lat* (latitude) does not follow a normal distribution; we can see that the kernel curve is very different from the normal curve.
- Location: since the distribution is not normal, the median seems to be a more robust statistic (38.21), although it is very close to the mean (37.95).
- Spread: domain knowledge would be useful here to determine if the spread of the distribution is significant or not.



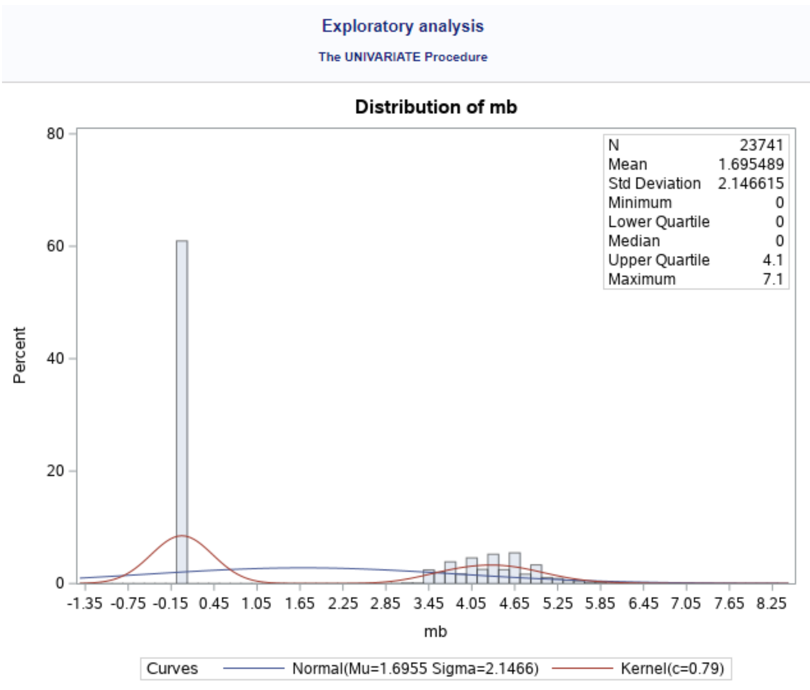
- Shape: just as the *lat* variable, the *long* (longitude) variable does not follow a normal distribution nor is unimodal.
- Location: The median of the distribution is 28.24.



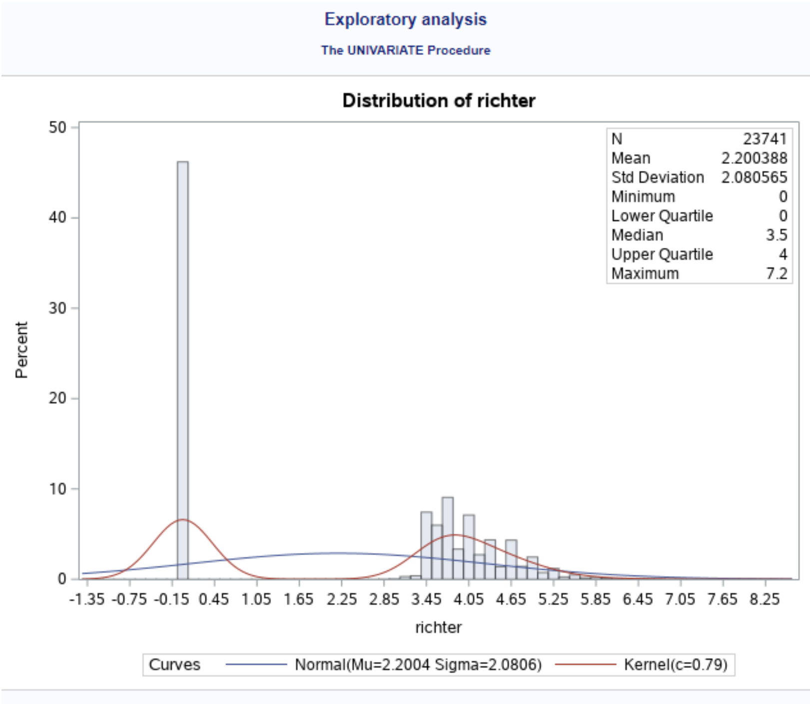
- Spread: we notice a lot of null values in the the *md* column (one of the magnitude scale of the dataset). It is unclear if those values are missing values or corresponds to an actual measurement.
- Shape: the distribution, even if we ignore those null values, is still very skewed to the right, it does not seem to be normal.
- Location: the median reflects the amount of null values and is equal to 0. The mean is 1.91.



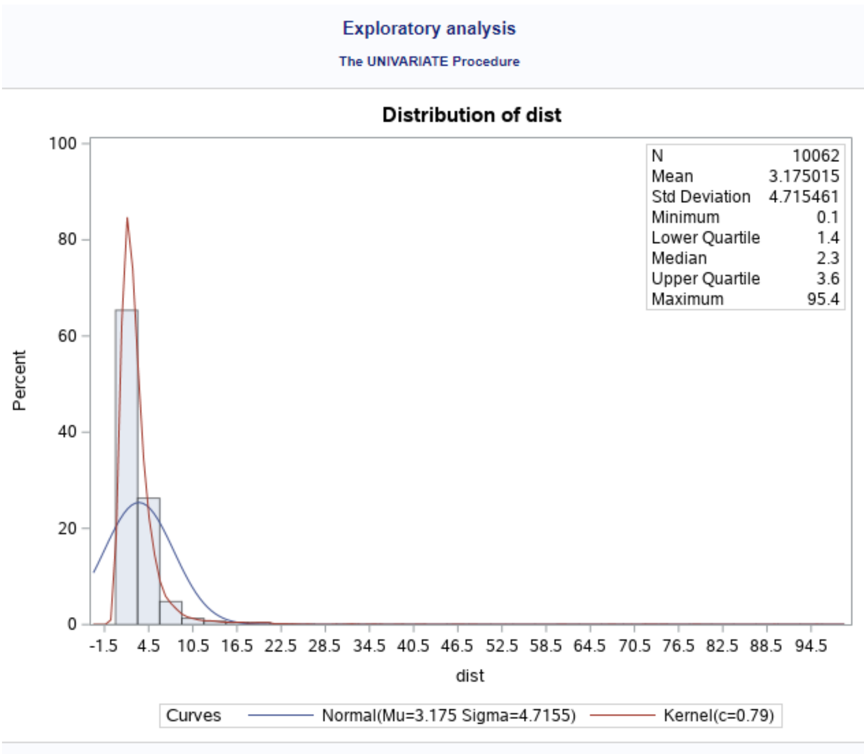
- Spread: the *mw* variable, another magnitude scale, also records null values.
- Shape: the distribution seems too peaked to be normal, the kernel curve does not follow the normal curve.
- Location: the mean (4.48) and the median (4.7) are close.



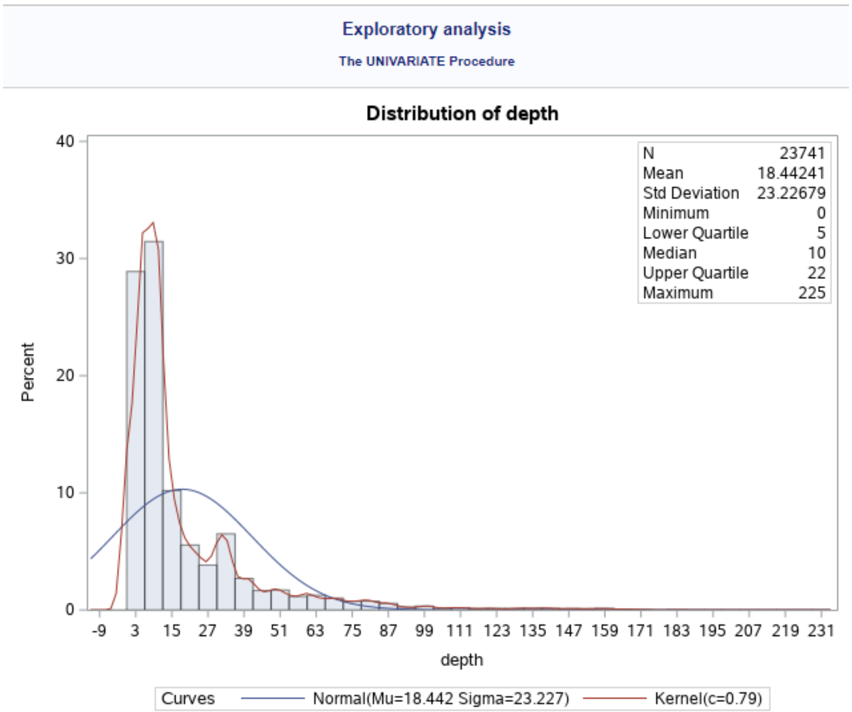
- Spread: the *md* column also contains a lot of null values.
- Shape: without considering the null values, the distribution might be close to normality. It seems to be symmetric and unimodal.
- Location: the median is equal to 0, since more than 50% of the values are 0. The mean is equal to 1.70.



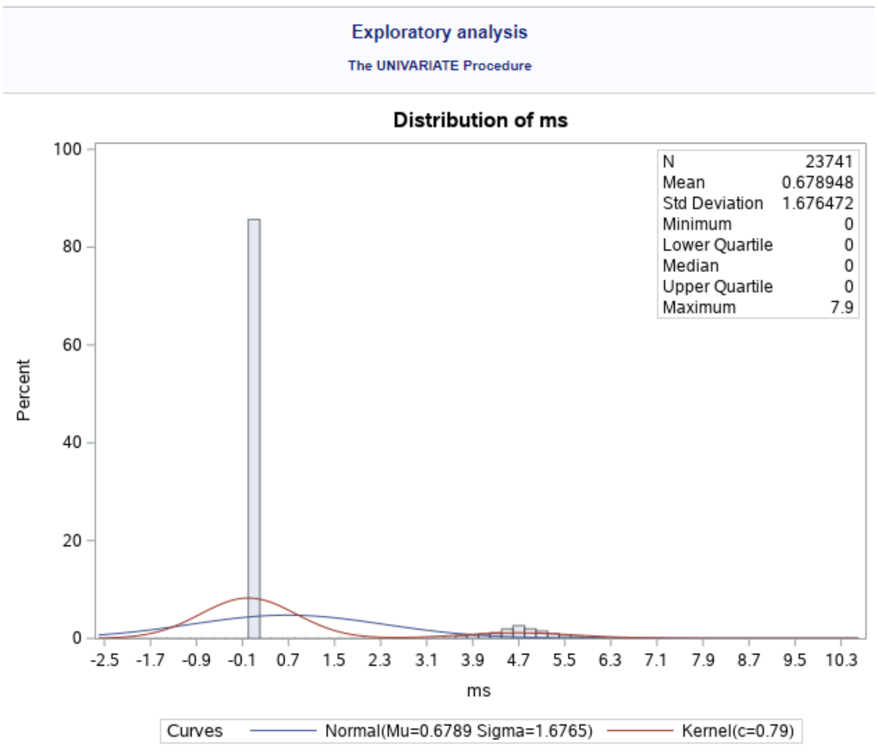
- Spread: more than 40% of the values are null for the *richter* variable.
- Shape: without considering the null values, it seems that the distribution is slightly skewed to the right, therefore not normal.
- Location: The median is 3.5 and the mean 2.20.



- Shape: the *dist* variable (distance travelled by the earthquake) is very skewed to the right.
- Spread: we notice a very wide spread, with a maximum value of 95.4 but with an upper quartile of 3.5. Most of the values (75%) are contained below this last number.
- Location: The mean is 3.18 and the median 2.3.



- Shape: the variable *depth* is very skewed to the right, hence does not follow a normal distribution.
- Spread: the range is wide, from 0 to 225. The null values may or may not be missing values.
- Location: The mean of the distribution is 18.44 and the median 10.



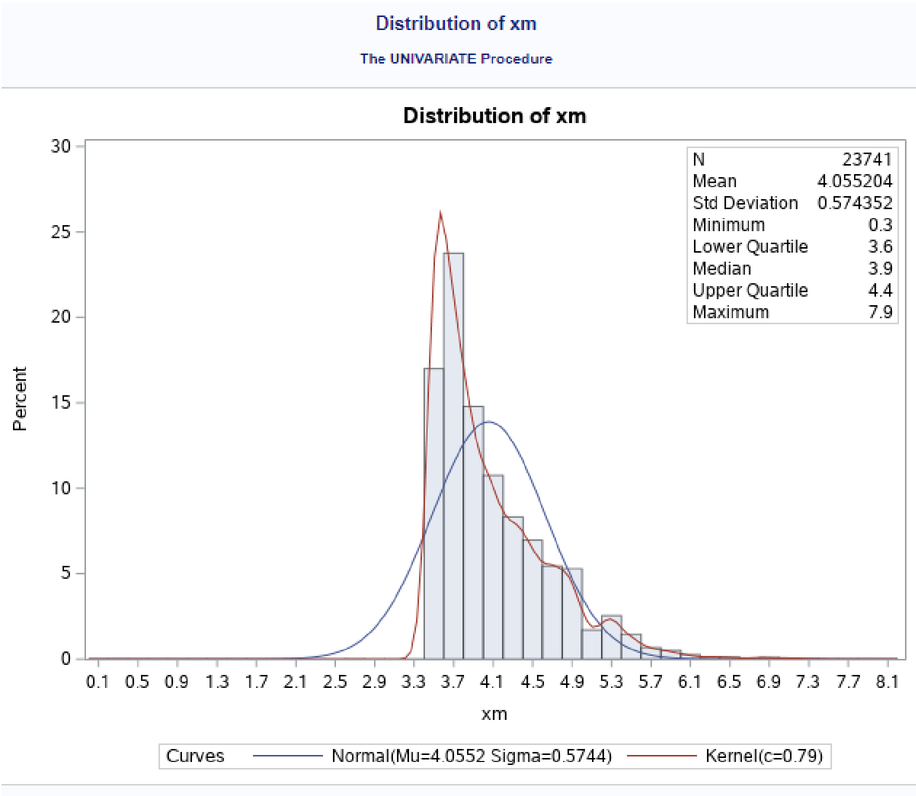
- Spread: more than 80% of the magnitude scale *ms* values are equal to zero. This is further obvious when computing the quartiles, which are all equal to zero, up to the 75% quartile.
- Shape: without considering the null values, which may or may not be missing values, the distribution seems to be symmetric and unimodal. The values seem to be normally distributed.
- Location: the mean is equal to 0.68.

<u>**TESTING AND MODELING**</u>

a) <u>What is the largest magnitude value for each observation (denoted *xm*)? And is the average value of *xm* different from 4.1?</u>

We are going to create a new variable *xm* which is the largest value of the measurements, out of *ms, md, mw, mb* and *richter.* Then we will verify if *xm* is significantly different to 4.1.

- First, a new column *xm* has been created.
- Then we have drawn a histogram to verify the shape of the distribution.



We will use a one-sample t-test to verify the assumption of the mean being different to 4.1. The shape of the distribution is not symmetric and skewed to the right; but in practice, the t-test can provide good results even when the assumption of normality is dubious.

One-Sample t-test:

Here we test the null hypothesis (the mean of *xm* is equal to 4.1) against the alternative hypothesis (the mean of *xm* is significantly different from 4.1).

$$H_0: \mu = 4.1$$

$$H_A: \mu \neq 4.1$$

One Sample T-Test : Mean=4.1 for xm

The TTEST Procedure

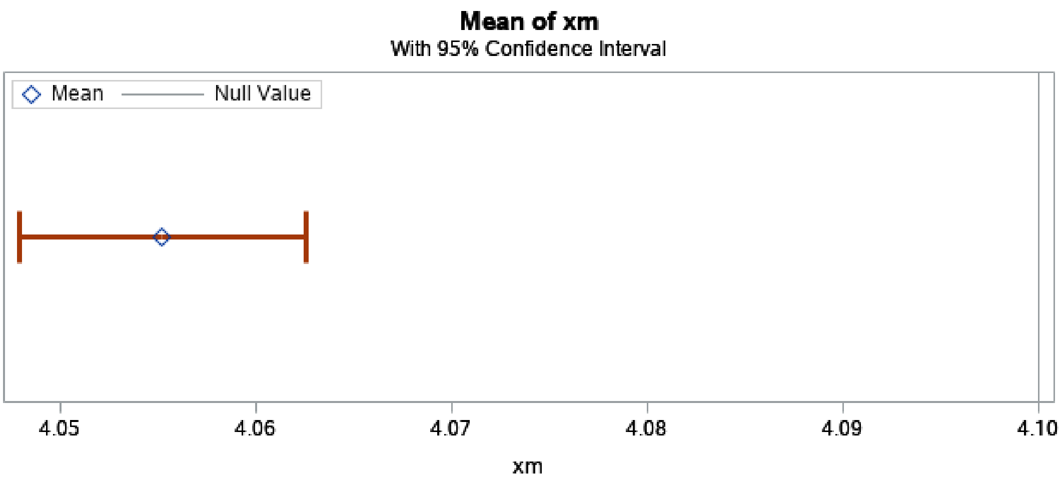
Variable: xm

N	Mean	Std Dev	Std Err	Minimum	Maximum
23741	4.0552	0.5744	0.00373	0.3000	7.9000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
4.0552	4.0479	4.0625	0.5744	0.5692	0.5796

DF	t Value	Pr > |t|
23740	-12.02	<.0001

The p-value being inferior to the significance level of 0.05, we can reject the null hypothesis and confirm that in fact the mean of *xm* is significantly different from 4.1.



The 95% confidence interval does not contain 4.1 and hence confirms that we can reject the null hypothesis.

b) Is there a difference in the moment magnitude scale value of an earthquake (*mw*) between countries, on average?

We will use a one-way ANOVA test to determine if there is a difference in *mw* between countries. Our null hypothesis is that there is no difference on average between countries for the variable *mw*. Our alternative hypothesis is that there is a significant difference between countries.

One-Way Anova : country vs mw

The GLM Procedure

Class Level Information		
Class	Levels	Values
country	17	aegean_sea albania azerbaijan blacksea bulgaria egypt georgia greece iran iraq israel macedonia mediterranean romania russia syria turkey

Number of Observations Read	23741
Number of Observations Used	4950

One-Way Anova : country vs mw

The GLM Procedure

Dependent Variable: mw

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	218.931125	14.595408	13.78	<.0001
Error	4934	5224.339784	1.058845		
Corrected Total	4949	5443.270909			

R-Square	Coeff Var	Root MSE	mw Mean
0.040221	22.98123	1.029002	4.477576

The p-value is inferior to the significance level, hence we can reject the null hypothesis and conclude that there is a statistically significant difference between countries, on average, for the variable *mw*.

One-Way Anova : country vs mw

The GLM Procedure

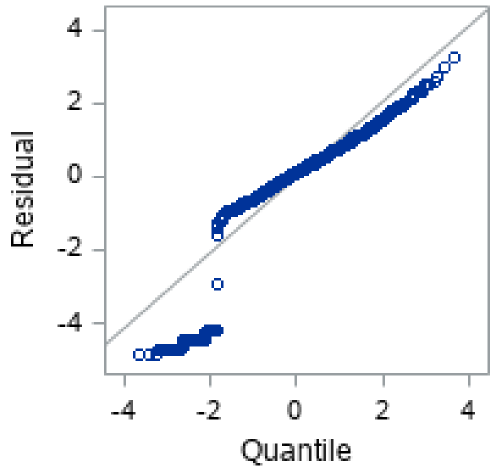
Levene's Test for Homogeneity of mw Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
country	13	766.0	58.9265	5.19	<.0001
Error	4933	55982.5	11.3486		

To verify the assumption of equal variance between the levels of the variable *country*, a Levene's test has been computed. However its p-value is inferior to the significance level, which means the assumption is violated. The results above are therefore biaised.

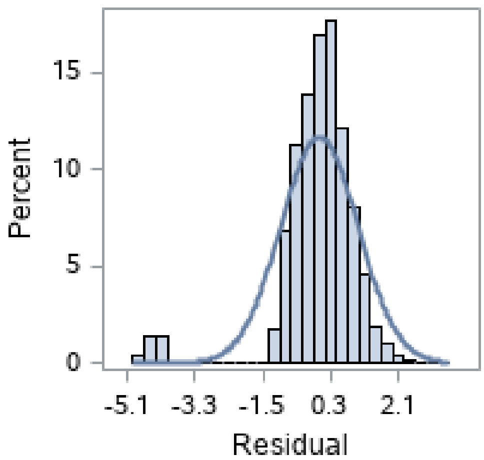
Below is a table that contains the mean and standard deviation for each level of the variable *country.* We can see that the standard deviation is very different from one group to another, which confirms the results of the Levene's Test.

Level of country	N	mw Mean	Std Dev
aegean_sea	242	4.26528926	1.01700975
albania	1	4.60000000	.
azerbaijan	62	4.85000000	0.78651545
blacksea	41	4.66585366	1.46502726
bulgaria	78	4.87435897	0.80088196
egypt	2	4.25000000	0.21213203
georgia	116	4.67672414	0.83783964
greece	987	4.21114488	1.24752794
iran	89	4.87303371	0.81195108
iraq	29	4.84137931	0.47996818
macedonia	15	2.95333333	2.21161436
mediterranean	967	4.69255429	0.85637688
romania	15	4.44666667	1.34263423
russia	123	4.80081301	0.44028604
syria	27	4.44444444	0.53229065
turkey	2156	4.45932282	1.03065375

Residuals plots have been drawn to verify the normality assumption.



In the QQ plot above, the residuals do not really follow the normal line. We notice a lot of outliers as well.



We can see the outliers on the left of the residual histogram. They may correspond to the null values that we have noted in the exploratory analysis of *mw*. Otherwise, the distribution of the residuals seems to be reasonably normal.
Since we do not have further information regarding the outliers (if they are erroneous or relevant), we cannot remove them.

a) How can we build a regression model with *richter* as a response variable?

Test for collinearity:

We need to test for multicollinearity between the predictors, and see if we can remove redundant variables.

Collinearity Diagnostics

The REG Procedure
Model: MODEL1
Dependent Variable: richter

Number of Observations Read	23741
Number of Observations Used	1731
Number of Observations with Missing Values	22010

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	405.88296	50.73537	91.68	<.0001
Error	1722	952.95248	0.55340		
Corrected Total	1730	1358.83544			

Root MSE	0.74391	R-Square	0.2987
Dependent Mean	4.41236	Adj R-Sq	0.2954
Coeff Var	16.85962		

First, the majority of the values tested seems to be missing. The following results only concern 1,731 rows, which is less than 10% of the total number of observations.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.71302	0.58531	1.22	0.2233	0
lat	1	0.04001	0.01470	2.72	0.0066	1.05222
long	1	0.00871	0.00328	2.66	0.0080	1.03701
dist	1	0.00622	0.00434	1.43	0.1517	1.02845
depth	1	-0.00101	0.00099236	-1.02	0.3078	1.36967
md	1	-0.23587	0.02621	-9.00	<.0001	10.64400
mw	1	0.36775	0.02575	14.28	<.0001	1.73521
ms	1	0.35052	0.04061	8.63	<.0001	25.66397
mb	1	-0.06033	0.03639	-1.66	0.0976	20.23008

The VIF of each predictor has been calculated. The variable *ms* has the largest VIF (25.66).
We will remove the variable from the current analysis and rerun the program.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.46049	0.59692	0.77	0.4406	0
lat	1	0.04273	0.01501	2.85	0.0045	1.05173
long	1	0.00813	0.00335	2.43	0.0154	1.03656
dist	1	0.00517	0.00443	1.17	0.2431	1.02764
depth	1	-0.00062670	0.00101	-0.62	0.5359	1.36689
md	1	-0.13486	0.02395	-5.63	<.0001	8.52220
mw	1	0.40819	0.02585	15.79	<.0001	1.67776
mb	1	0.17643	0.02442	7.22	<.0001	8.73810

There is no VIF superior to 10, hence we can conclude there is no redundant information, and all of the variables (plus the categorical ones) can be tested for model selection.

Model selection:

To select a relevant model, we have opted for a stepwise selection.

Stepwise Model Selection for richter - SL 0.05

The GLMSELECT Procedure

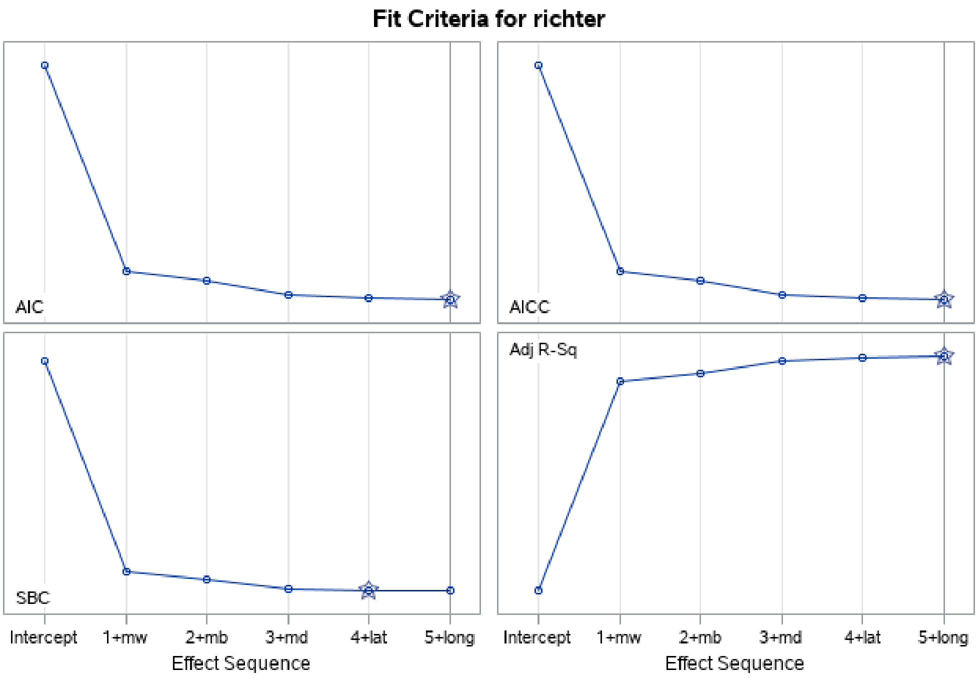
Data Set	WORK.EARTHQUAKES
Dependent Variable	richter
Selection Method	Stepwise
Select Criterion	Significance Level
Stop Criterion	Significance Level
Entry Significance Level (SLE)	0.05
Stay Significance Level (SLS)	0.05
Effect Hierarchy Enforced	None

Number of Observations Read	23741
Number of Observations Used	1731

Class Level Information		
Class	Levels	Values
country	1	turkey
direction	8	east north north_east north_west south south_east south_west west

Dimensions	
Number of Effects	11
Number of Parameters	18

Once again, the program has only used a small fraction of the dataset, due to missing values. Furthermore, the 17 levels of the *country* variable have not been recognized by the program (only Turkey).



The AIC, AICC and Adjusted R-Square all detected the same optimal model.
The last step of the program can be seen below.

Stepwise Model Selection for richter - SL 0.05

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 5).

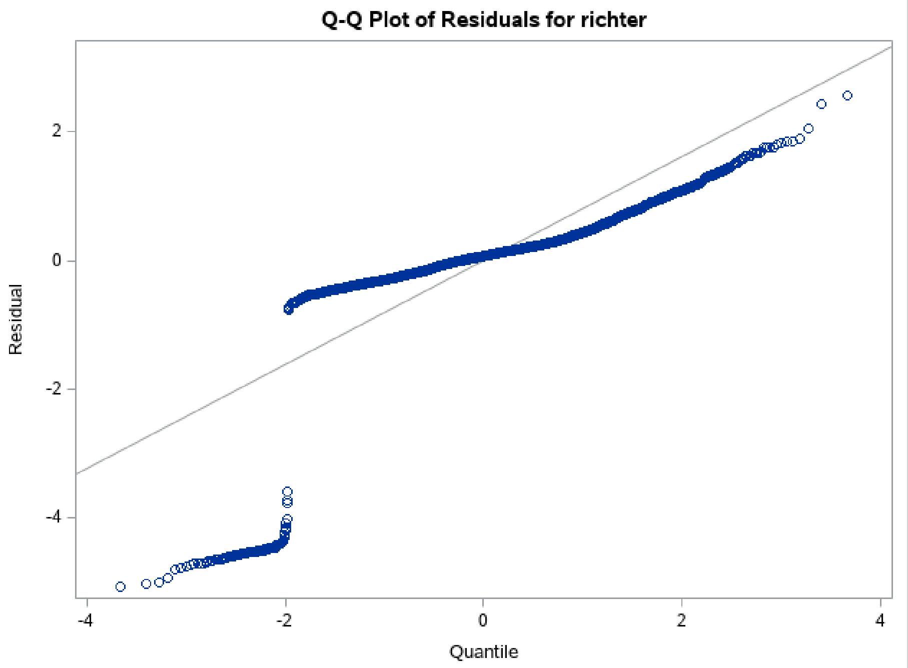
Effects:	Intercept lat long md mw mb

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	5	363.63902	72.72780	126.06
Error	1725	995.19641	0.57693	
Corrected Total	1730	1358.83544		

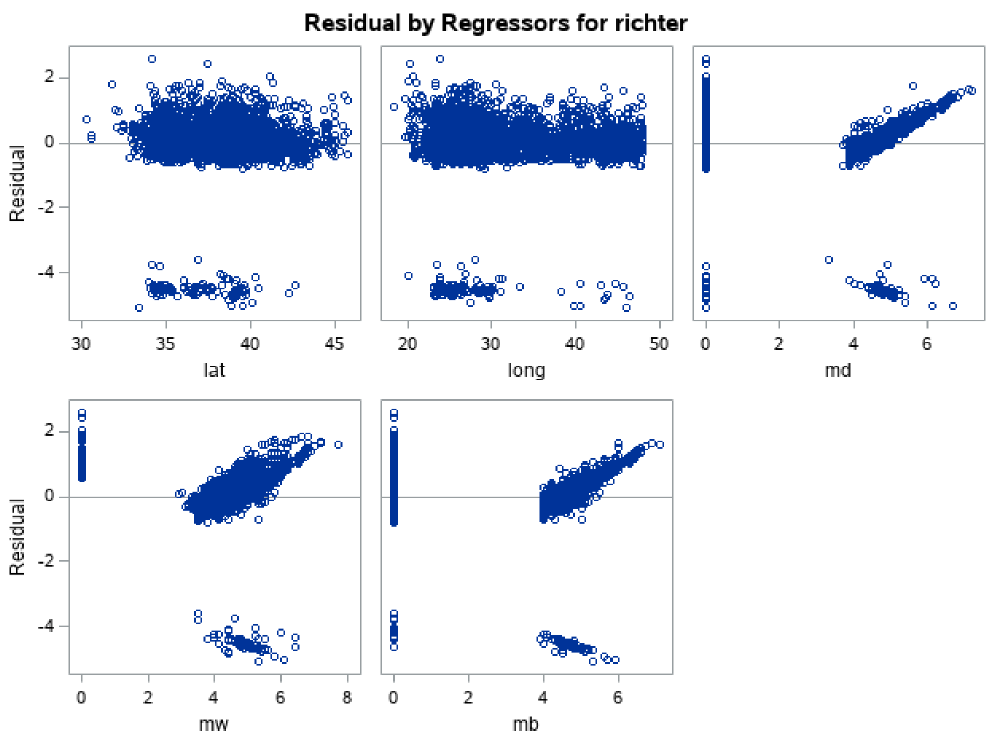
Root MSE	0.75956
Dependent Mean	4.41236
R-Square	0.2676
Adj R-Sq	0.2655
AIC	786.86651
AICC	786.93151
SBC	-913.39476

The 5 variables *lat, long, md, mw* and *mb* have been selected for the model.
However, we notice a very low Adjusted R-Square: the model only explains 26.55% of the variability of the data, which is a mediocre performance.

Residual plots:



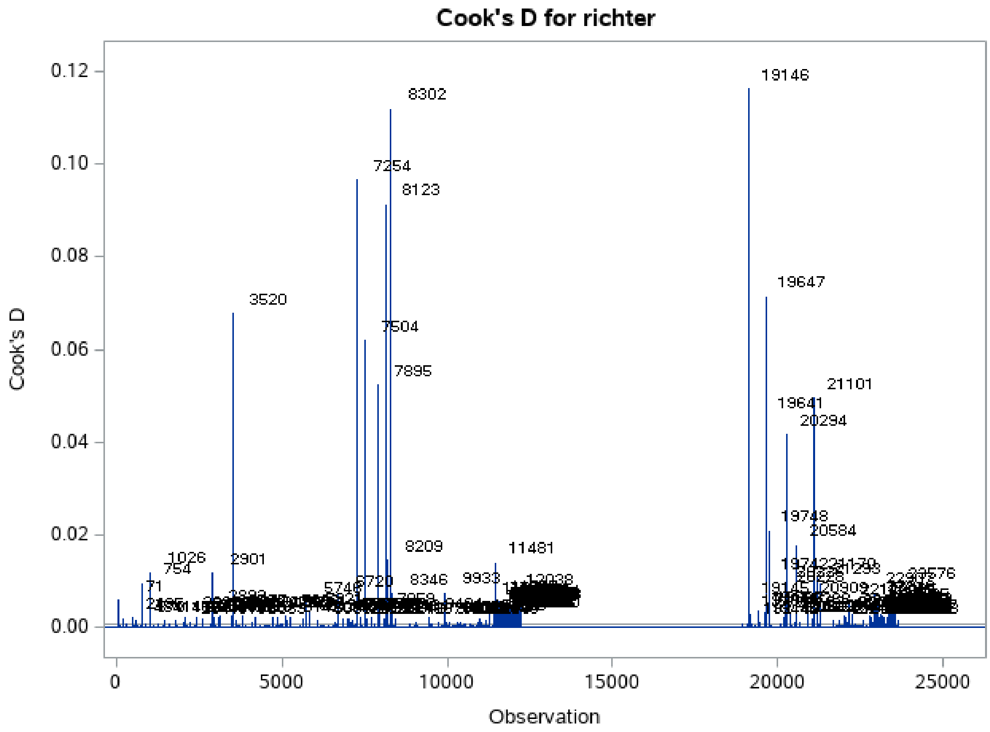
The QQ plot above shows that the residuals do not follow the normal line, and hence the assumption of normality for the residual distribution is dubious. Furthermore, we can notice a lot of outliers on the bottom left of the plot, which might influence those results.



The residuals vs fits plots show outliers, and for several variables (*mb, mw* and *md*), an ascending pattern that violates the assumption of linearity.

Influential Observations Diagnostics

The REG Procedure
Model: siglimit
Dependent Variable: richter



This plot shows the great number of outliers that are in fact influential observations, according to calculations (Cook's distance). Those observations explain in part the poor performance of the model.

d) How can we build a logistic model with *richter* as a response variable?

A new variable called *serious* has been added to the dataset. If the variable *richter* is superior or equal to 5, the variable *serious* is equal to 1, otherwise to 0 ("not serious").

That way we can compute a logistic regression with response *serious.* All variables have been used, except for *id, richter, xm* and *mw*.

Logistic Regression - Richter

The LOGISTIC Procedure

Model Information	
Data Set	WORK.EARTHQUAKES
Response Variable	serious
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	23741
Number of Observations Used	10062

Response Profile		
Ordered Value	serious	Total Frequency
1	0	9712
2	1	350

Probability modeled is serious=0.

Note: 13679 observations were deleted due to missing values for the response or explanatory variables.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	3040.693	1999.737
SC	3047.910	2107.984
-2 Log L	3038.693	1969.737

We have obtained a new model with an AIC of 1,999.737.

e) How does a simple logistic regression model from the *richter* variable compare to the previous model?

Simple Logistic Regression - Richter vs xm

The LOGISTIC Procedure

Model Information	
Data Set	WORK.EARTHQUAKES
Response Variable	serious
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	23741
Number of Observations Used	23741

Response Profile		
Ordered Value	serious	Total Frequency
1	0	22752
2	1	989

Probability modeled is serious=0.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	8224.823	2243.667
SC	8232.898	2259.817
-2 Log L	8222.823	2239.667

If we compute a simple logistic regression with *serious* as a response variable and *xm* as the only predictor, we obtain a model with an AIC of 2,243.667 which is higher than the AIC of the previous model (1,999.737). Hence, we can conclude that the model with multiple predictors is more efficient.

CONCLUSION

Throughout this analysis, we have answered 5 different questions of interest:

a) What is the largest magnitude value for each observation (denoted *xm*)? And is the average value of *xm* different from 4.1?

With a one-sample t-test, we have concluded that the average value of *xm* was indeed different from 4.1 and more likely to lie between the 95% CI [4.05, 4.06].

b) Is there a difference in the moment magnitude scale value of an earthquake (*mw*) between countries, on average?

According to the one-way ANOVA test that was computed, there is a difference in the moment magnitude scale value between countries. However, we have seen that assumptions had been violated and the results are therefore biased.

c) How can we build a regression model with *richter* as a response variable?

After testing for collinearity and using stepwise selection, the most efficient model was found with 5 variables for the response variable *richter.* However, assumptions were violated and hence the performance of the model is very mediocre.

d) How can we build a logistic regression model from the (modified) *richter* variable?

A new binary variable has been created to build a logistic model with several predictors. The same assumptions were violated, hence it would need further testing to provide relevant predictions.

e) How does a simple logistic regression model from the *richter* variable compare to the previous model?

We have computed a simple logistic regression and compared it with the previous one: using the AIC metric, we have determined than the multiple logistic model was the most efficient.