LAB REPORT

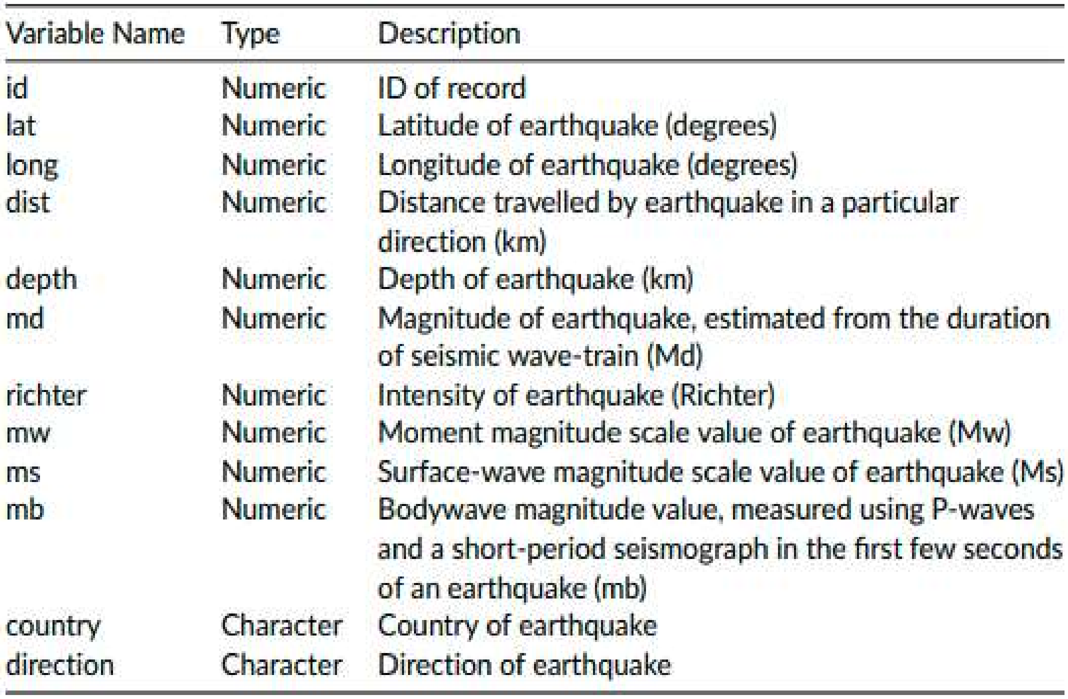
Data Management and Analytics using SAS

VU DUC NGHIA - Student ID: 2611002V

I. Introduction

Earthquakes happen regularly around the world and caused huge damages to humans and properties. Understanding these natural phenomena could help us to prepare and deal better with such disasters. Magnitude of earthquakes can be measured on different scales such as moment magnitude scale, body-wave magnitude scale, surface-wave magnitude scale. However, the popular one, which is often mentioned by news outlets, is intensity of earthquake (Richter scale).

The dataset is used for analysis is **earthquakes.sas7bdat**, which is provided by following path "/courses/dc36fc35ba27fe300/DMASAssessments". The dataset has 23741 independent records about earthquakes in several years. The dataset has 12 variables which provides various information about different kinds of magnitude, the latitude, longitude and depth of earthquakes…Detailed descriptions are given in below table



Variable 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, estimated from 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	Bodywave magnitude value, measured using P-waves and a short-period seismograph in the first few seconds of an earthquake (mb)
country	Character	Country of earthquake
direction	Character	Direction of earthquake

This report is generated to provide data analysis and solution to some questions of interest about earthquakes in different countries in several years. All results are generated from SAS.

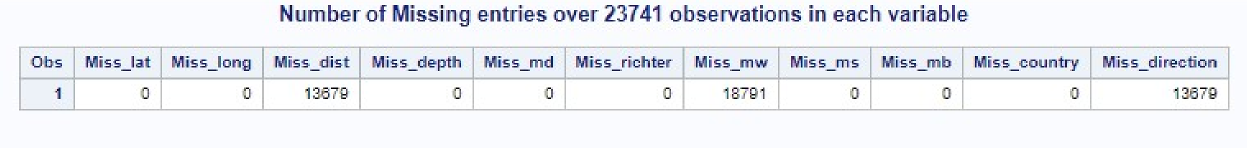
II. Exploratory Analysis:

Check Missing Values:

It is important to be aware how many missing data points for each variable. Following code provides such information (assumed dataset is already loaded to SAS and loaded to work library):

```
proc sql;

create table Missing_Count as
        select count(*)-count(lat) as Miss_lat, count(*)-count(long) as Miss_long,
                count(*)-count(dist) as Miss_dist, count(*)-count(depth) as Miss_depth,
                count(*)-count(md) as Miss_md, count(*)-count(richter) as Miss_richter,
                count(*)-count(mw) as Miss_mw, count(*)-count(ms) as Miss_ms,
                count(*)-count(mb) as Miss_mb, count(*)-count(country) as Miss_country,
                count(*)-count(direction) as Miss_direction from work.earthquakes;
quit;

proc print data=work.Missing_Count;
        title "Number of Missing entries over 23741 observations in each variable";
run;
```

Below table indicates that there is significant amount of data are missed in variables **dist**, **mw** and **direction**. While there are no missing entries in columns **md**, **richter**, **mb**, and **ms**. So we need to be aware of this issue when doing analysis.



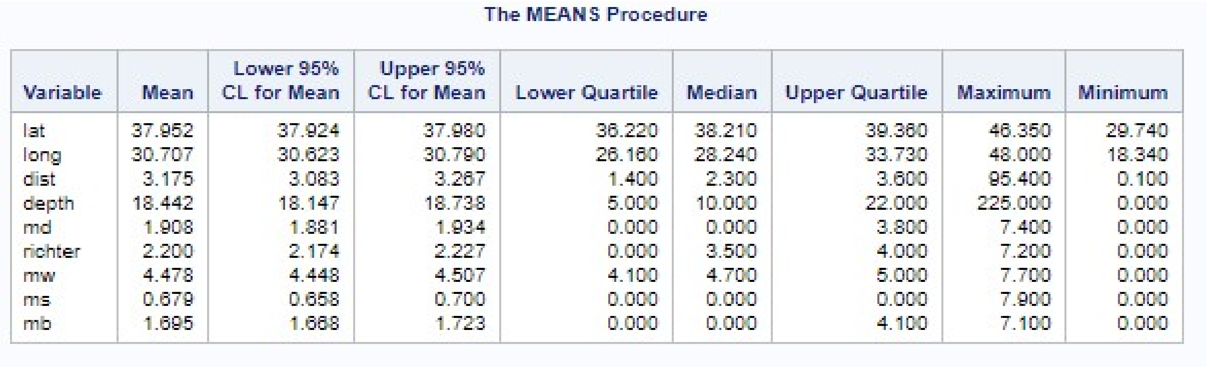
Number of Missing entries over 23741 observations in each variable

Obs	Miss_lat	Miss_long	Miss_dist	Miss_depth	Miss_md	Miss_richter	Miss_mw	Miss_ms	Miss_mb	Miss_country	Miss_direction
1	0	0	13679	0	0	0	18791	0	0	0	13679

Check Distribution of Numerical Variables:

We use **proc means** procedure to summarize numerical variables

```
proc means data=work.earthquakes mean clm q1 median q3 max min maxdec=3;
        var lat long dist depth md richter mw ms mb;
run;
```



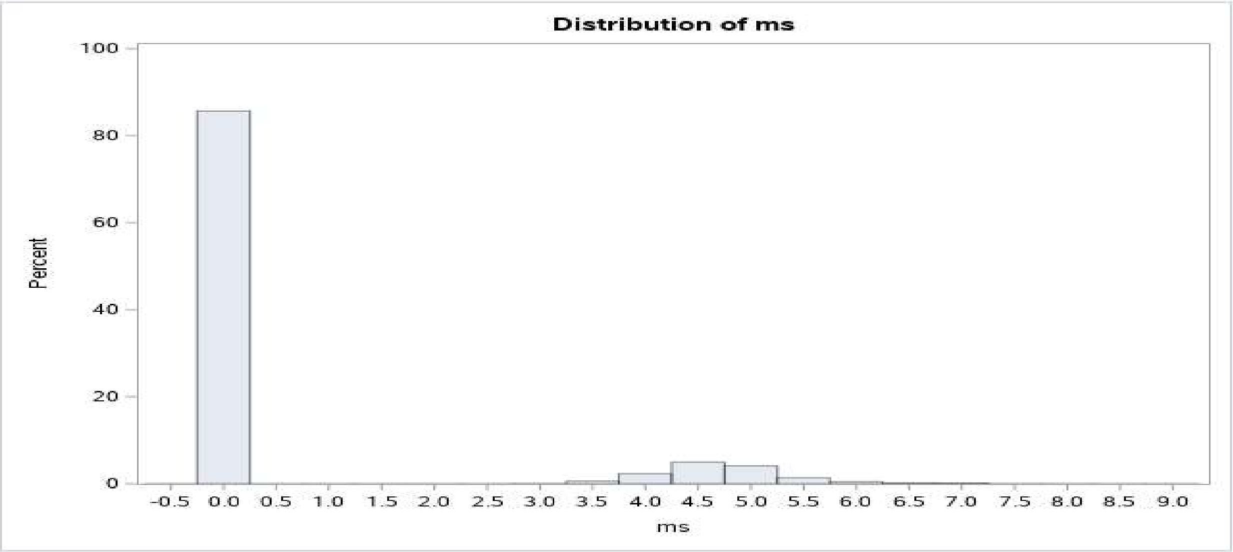
The MEANS Procedure

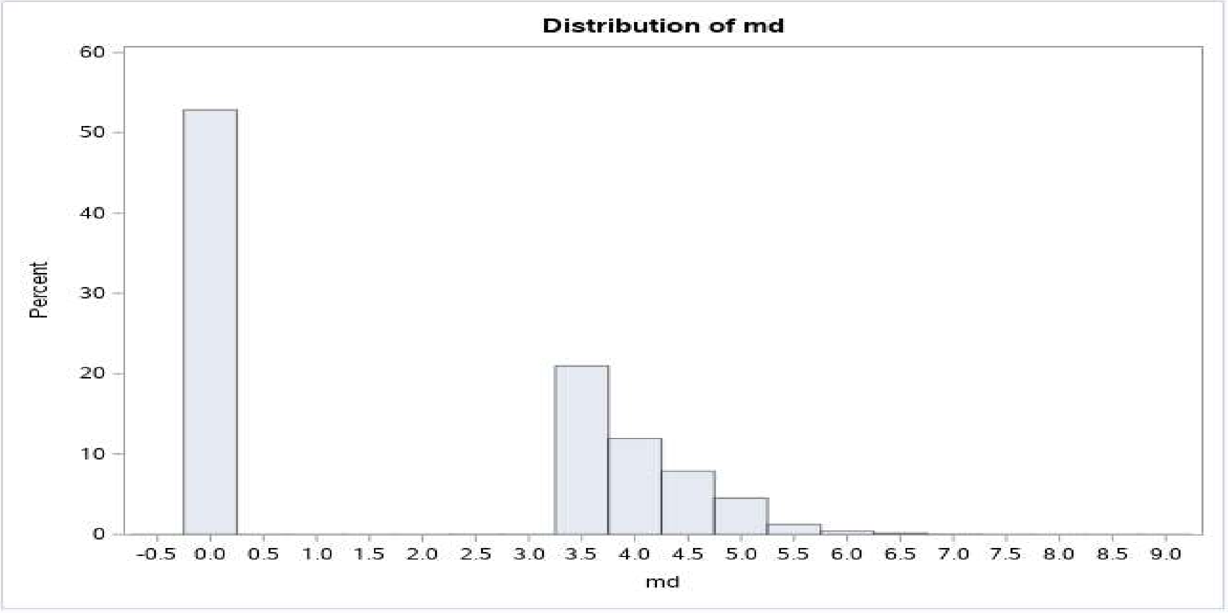
Variable	Mean	Lower 95% CL for Mean	Upper 95% CL for Mean	Lower Quartile	Median	Upper Quartile	Maximum	Minimum
lat	37.952	37.924	37.980	36.220	38.210	39.360	46.350	29.740
long	30.707	30.623	30.790	26.160	28.240	33.730	48.000	18.340
dist	3.175	3.083	3.267	1.400	2.300	3.600	95.400	0.100
depth	18.442	18.147	18.738	5.000	10.000	22.000	225.000	0.000
md	1.908	1.881	1.934	0.000	0.000	3.800	7.400	0.000
richter	2.200	2.174	2.227	0.000	3.500	4.000	7.200	0.000
mw	4.478	4.448	4.507	4.100	4.700	5.000	7.700	0.000
ms	0.679	0.658	0.700	0.000	0.000	0.000	7.900	0.000
mb	1.695	1.668	1.723	0.000	0.000	4.100	7.100	0.000

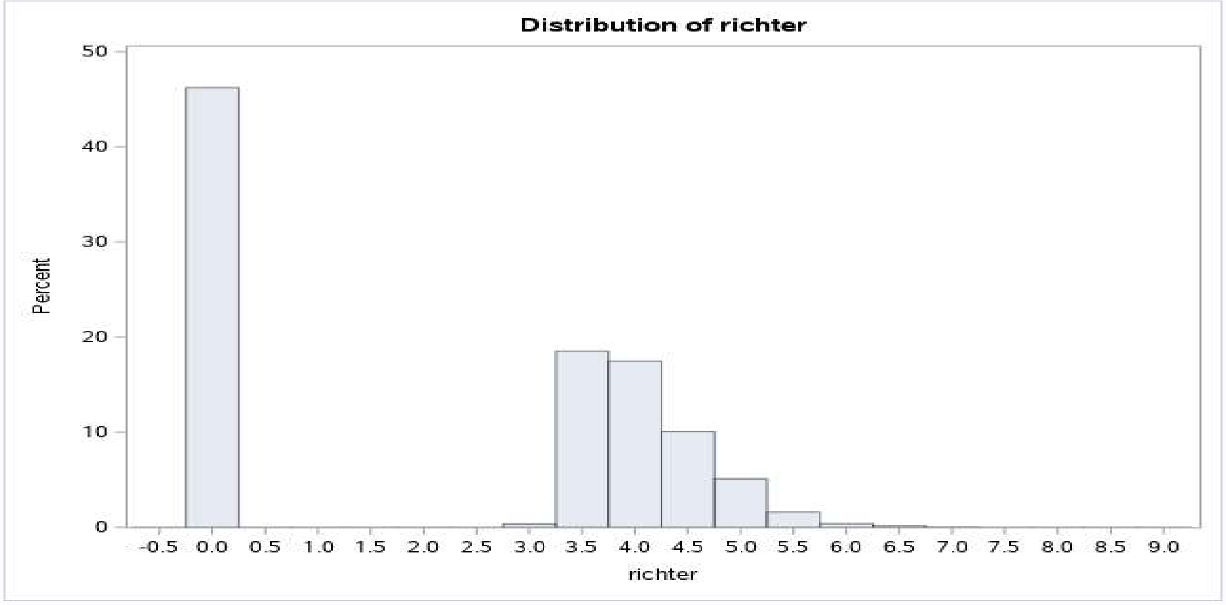
The above table shows a large proportion of data are zeros in variables **md**, **richter**, **ms**, **mb**. We need examine closer to the histograms of those variables and together with mw (note that mw has a lot of missing entries). We use procedure **proc univariate** to achieve that.

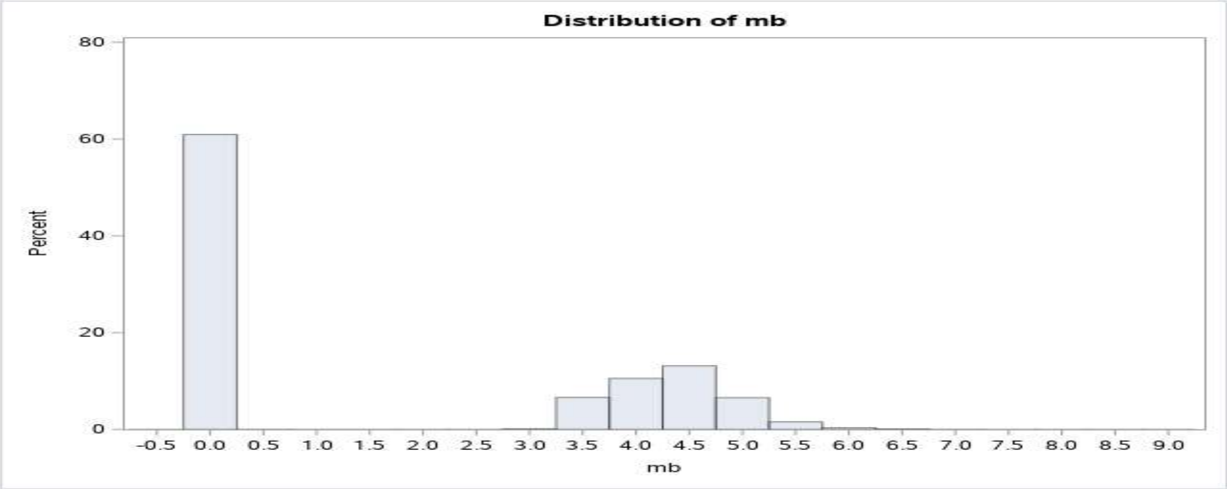
```
proc univariate data=work.earthquakes noprint;
            hist ms mb md richter mw / nmidpoints=20;
run;
```

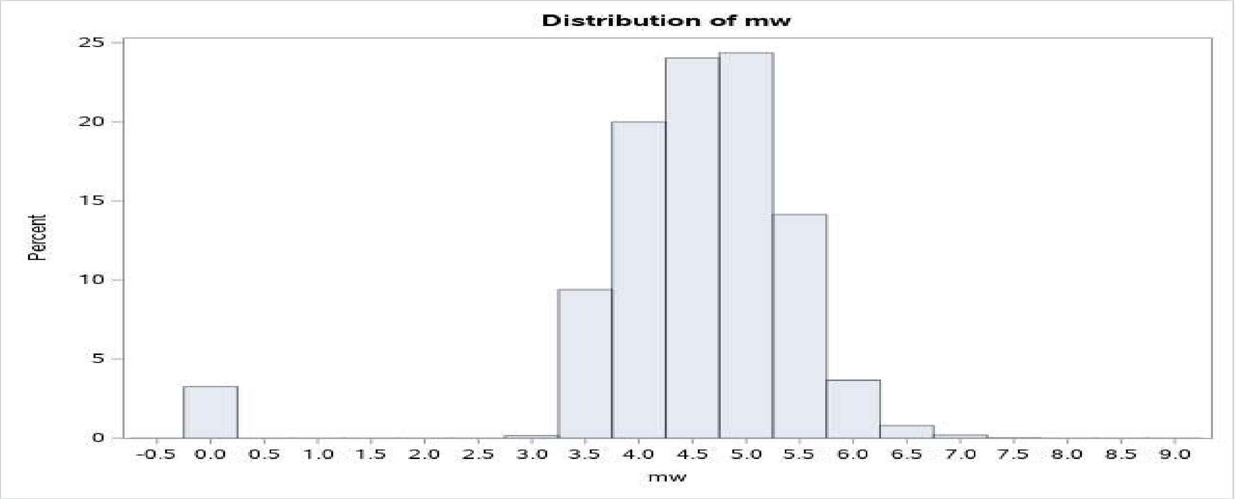
From the histograms, it can be seen that there are about 53%, 46%, 86%, 60% and 4% of data entries are **zeros** in variables **md**, **richter**, **ms**, **mb**, and **mw** respectively.






Distribution of richter
Percent
0
10
20
30
40
50
-0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0 6.5 7.0 7.5 8.0 8.5 9.0
richter

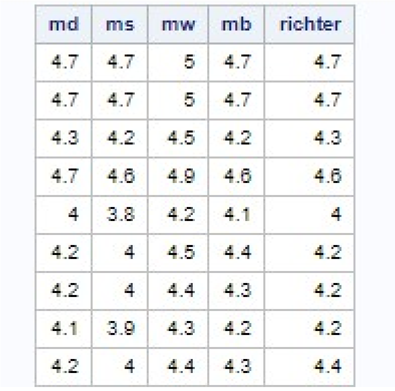

Distribution of mb
Percent
0
20
40
60
80
-0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0 6.5 7.0 7.5 8.0 8.5 9.0
mb


Distribution of mw
Percent
0
5
10
15
20
25
-0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0 6.5 7.0 7.5 8.0 8.5 9.0
mw

Dealing with zero values:

Let examines why there are so many zeros entries in dataset. We first examine a subset of dataset where information about **md**, **ms**, **mb**, **mw**, **richter** are available.

```
proc sql;
        select md,ms,mw,mb,richter,count(*) as number_rows from work.earthquakes
        where md >0 and ms>0 and mw>0 and mb>0 and richter>0;
quit;
```



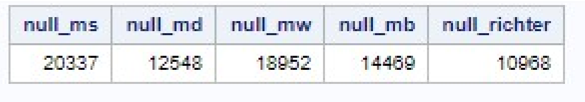
md	ms	mw	mb	richter
4.7	4.7	5	4.7	4.7
4.7	4.7	5	4.7	4.7
4.3	4.2	4.5	4.2	4.3
4.7	4.6	4.9	4.6	4.6
4	3.8	4.2	4.1	4
4.2	4	4.5	4.4	4.2
4.2	4	4.4	4.3	4.2
4.1	3.9	4.3	4.2	4.2
4.2	4	4.4	4.3	4.4

The above table shows that there are not much difference in values between those variables and it is reasonable that magnitude on a specific scale cannot be zero if magnitude on one of other measurements is greater than zero. Therefore, zero values here are not real values. They were filled for observations where data were not available, and we should treat those values as null values, otherwise our analysis would give incorrect conclusion. The following code replaces zeros by nulls

```
data work.earthquakes;
        set work.earthquakes;
        if mw=0 then mw=.;
        if ms=0 then ms=.;
        if md=0 then md=.;
        if mb=0 then mb=.;
        if richter=0 then richter=.;
run;
```

And the code below to count how many null values in each variable (after replacement)

```
proc sql;
        select count(*)-count(ms) as null_ms,
        count(*)-count(md) as null_md,
        count(*)-count(mw) as null_mw,
        count(*)-count(mb) as null_mb,
        count(*)-count(richter) as null_richter
        from work.earthquakes;
quit;
```

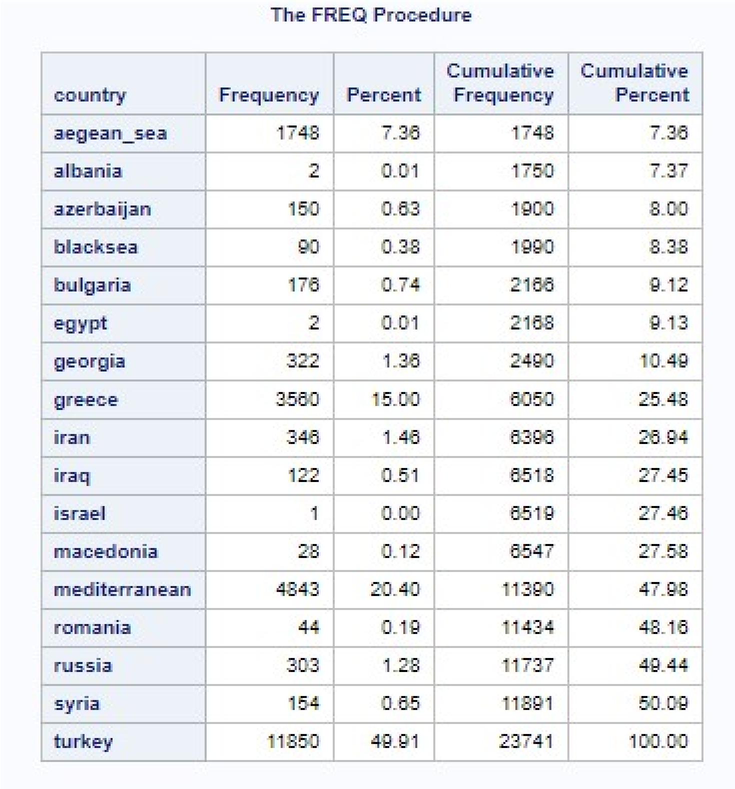


null_ms	null_md	null_mw	null_mb	null_richter
20337	12548	18952	14469	10968

We also should take a quick look on character variable **country**

```
proc freq data=work.earthquakes;
        table country;
run;
```

It can be seen that there are numerous earthquakes in Turkey, Greece and Mediterranean, while there are few in Israel, Albania and Egypt.



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

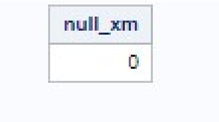
At this stage, we have general understanding about our dataset. It is ready to move on next steps to analyses deeper and answer questions of interest.

III. Formal Analysis

Part A: *Sometimes, the largest value of a series of measurements is used to represent the magnitude of an earthquake. Use "xm" to denote the largest magnitude value out of "md", "mw", "ms", "mb" and "richter" for each record. Is there evidence that the average value of "xm" is different to 4.1?*

We use data step to create **xm** variable, it can be seen from **sql procedure** that number of missing values in new variable **xm** is zero. It means that the column has full coverage.

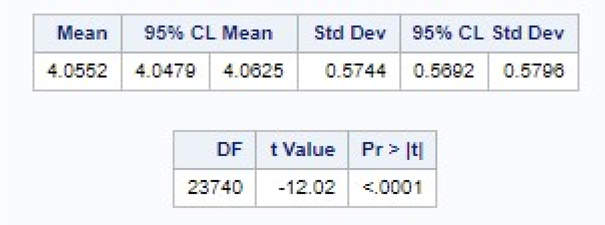
```
data work.earthquakes;
        set work.earthquakes;
        xm=max(md,mw,ms,mb,richter);
run;
proc sql;
        select count(*) as null_xm from work.earthquakes where xm is missing;
quit;
```



null_xm
0

In order to determine whether average of **xm** is different to 4.1, we use **ttest** procedure

```
proc ttest data=work.earthquakes H0=4.1;

        var xm;
run;
```

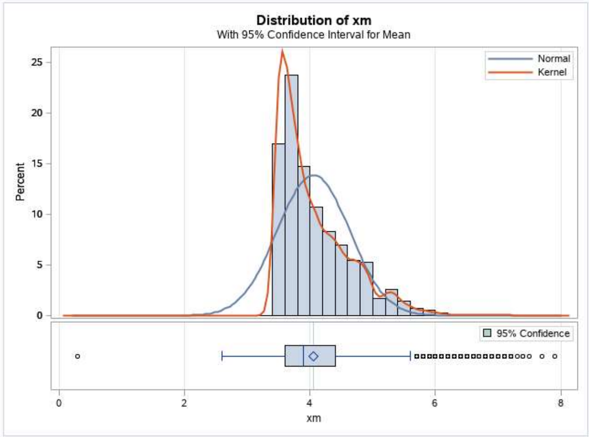


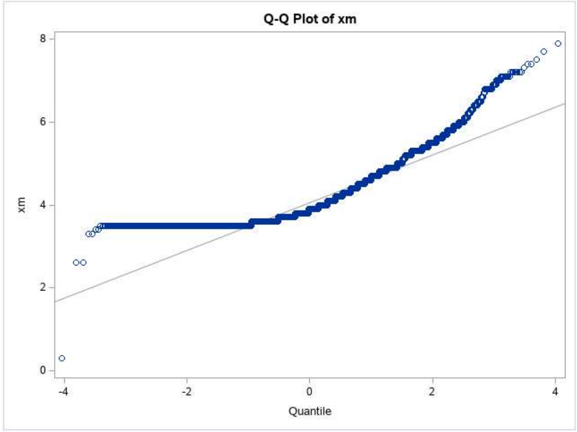
Mean	95% CL Mean		Std Dev	95% CL Std Dev	
4.0552	4.0479	4.0625	0.5744	0.5692	0.5796

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

The confidence interval does not contain 4.1. and p-value <0.001. It seems we can reject the null hypothesis that average of xm is equal to 4.1 and support alternative hypothesis that average of xm is different to 4.1. However, we need to check whether model assumptions are met. The independent assumption is satisfied as provided in the data, and earthquakes are natural phenomena, observations are independent.

The assumption of normal distribution seems do not hold as the histogram is highly skewed and a large portion of data points do not follow the diagonal line of QQ plot. Therefore, we cannot use ttest procedure to confirm whether the average of xm is different to 4.1.

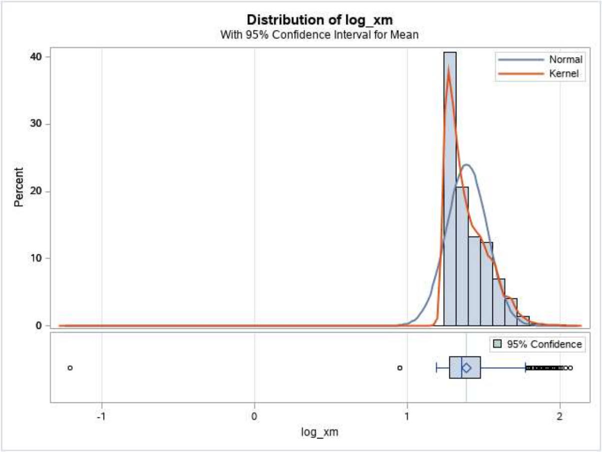


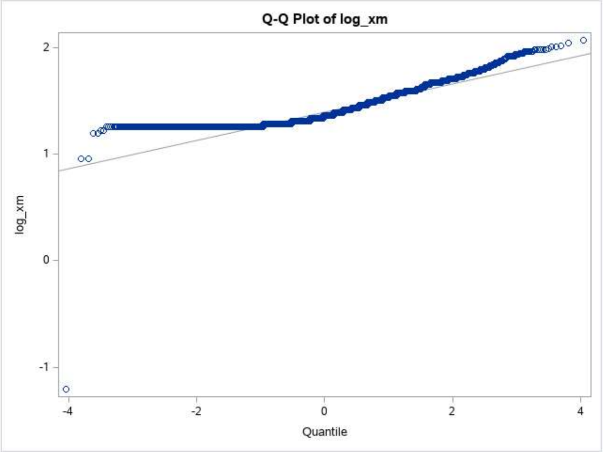


We may try to find some transformation of **xm** in order to satisfy the normal distribution assumption. Some proposed transformations are logarithm, exponential, square, square-root of **xm** but none of them work. One example (log transformation) is shown below.

Therefore, the final conclusion is there are no evidence/method to confirm whether the average of **xm** is different to 4.1. However, the mean of xm in **our sample** is very close to 4.1, which is at 4.055 and standard deviation is also small at 0.57.

```
data work.earthquakes_trans;
        set work.earthquakes;
        log_xm=log(xm);
        expo_xm=exp(xm);
        sqrt_xm=sqrt(xm);
        square_xm=xm**2;
run;

proc ttest data=work.earthquakes_trans;
        var log_xm expo_xm sqrt_xm square_xm;
run;
```

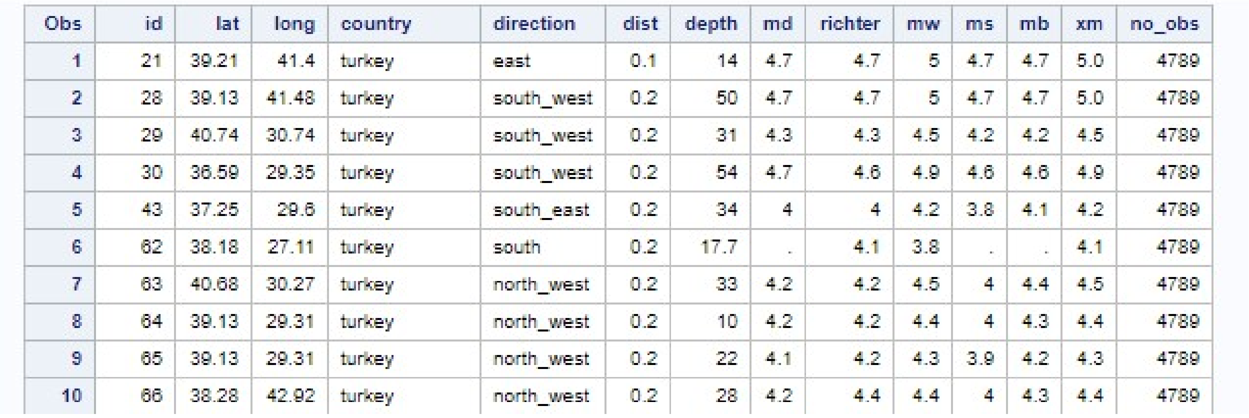




Part B: *Is there a difference in the moment magnitude scale value of an earthquake (Mw) between countries in which the earthquakes occurred, on average?*

We should be noticed that **mw** variable contains only 4789 valid entries (see **no_obs** column in below table). We should consider whether using only 4789 observations or fill null values with suitable values. Let us examine the table with only valid values of **mw** variable, we can see that **mw** is usually the highest value among other magnitudes, and the differences between measurements are from 0.0 to 0.3.

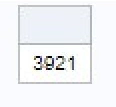
```
proc sql;
	create table valid_mw_only as
	select * from work.earthquakes where mw is not missing;
run;

proc print data=work.valid_mw_only;
run;
```



Obs	id	lat	long	country	direction	dist	depth	md	richter	mw	ms	mb	xm	no_obs
1	21	39.21	41.4	turkey	east	0.1	14	4.7	4.7	5	4.7	4.7	5.0	4789
2	28	39.13	41.48	turkey	south_west	0.2	50	4.7	4.7	5	4.7	4.7	5.0	4789
3	29	40.74	30.74	turkey	south_west	0.2	31	4.3	4.3	4.5	4.2	4.2	4.5	4789
4	30	36.59	29.35	turkey	south_west	0.2	54	4.7	4.6	4.9	4.6	4.6	4.9	4789
5	43	37.25	29.6	turkey	south_east	0.2	34	4	4	4.2	3.8	4.1	4.2	4789
6	62	38.18	27.11	turkey	south	0.2	17.7	.	4.1	3.8	.	.	4.1	4789
7	63	40.68	30.27	turkey	north_west	0.2	33	4.2	4.2	4.5	4	4.4	4.5	4789
8	64	39.13	29.31	turkey	north_west	0.2	10	4.2	4.2	4.4	4	4.3	4.4	4789
9	65	39.13	29.31	turkey	north_west	0.2	22	4.1	4.2	4.3	3.9	4.2	4.3	4789
10	66	38.28	42.92	turkey	north_west	0.2	28	4.2	4.4	4.4	4	4.3	4.4	4789

Actually, in the table **valid_mw_only**, there are 3921 over 4789 observations of **xm** equal to mw. Therefore, it is fine to fill null entries in **mw** by **xm**.

```
proc sql;
	select count(*) from work.valid_mw_only where mw=xm;
quit;
```



3921

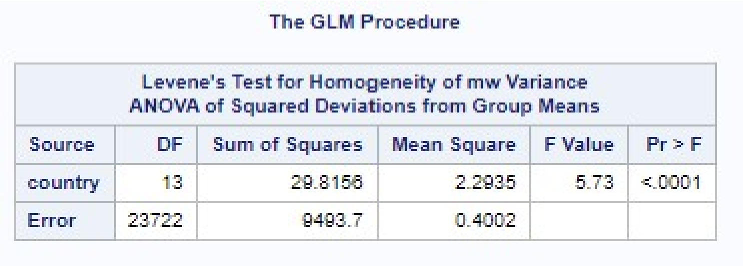
Below code fill missing **mw** by valid **xm** and create a new **dataset work.earthquakes_partb**

```
data work.earthquakes_partb;
	set work.earthquakes;
	if mw=. then mw=xm;
run;
```

In order to use One-Way ANOVA test to confirm whether means of **mw** are different between countries, we need to verify assumptions by conducting a levene test. It is noted that SAS removes countries with fewer threes observations (Israel, Albania, Egypt) for the test, so we do not worry about the effects of these countries.

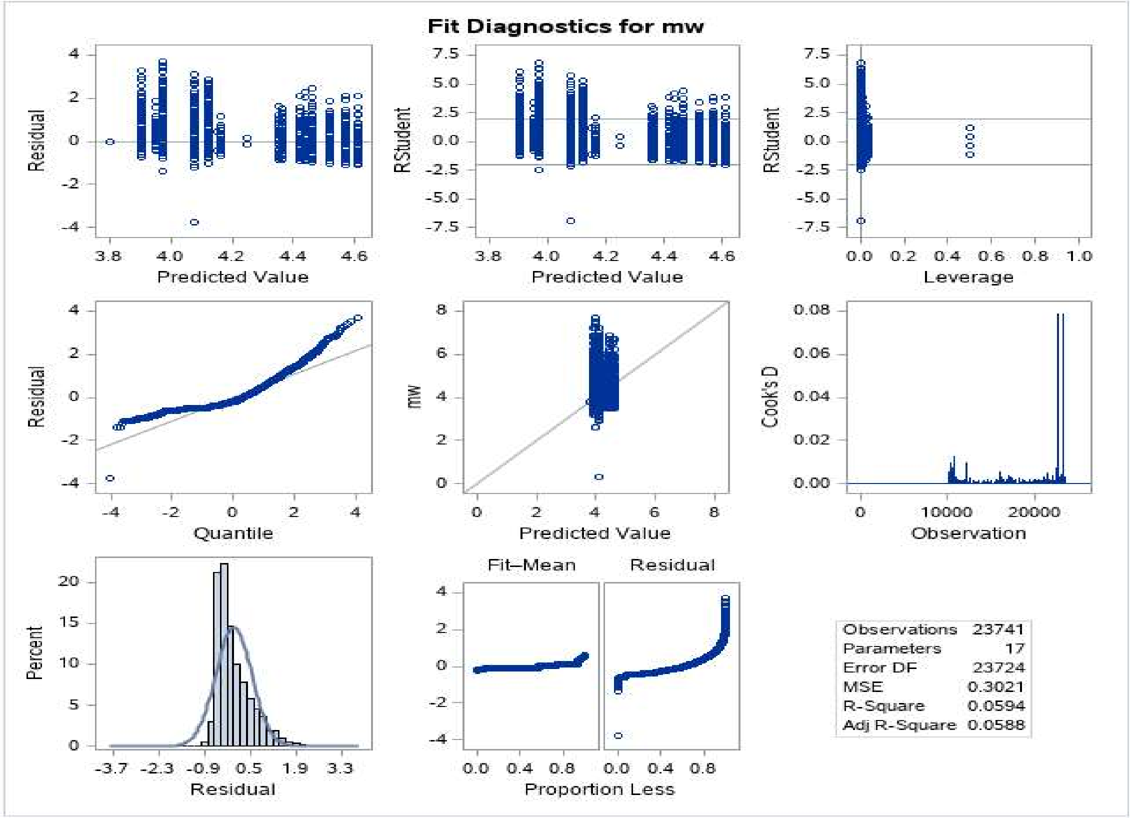
```
proc glm data=work.earthquakes_partb plots(maxpoints=24000)=diagnostics;
        class country;
        model mw=country;
        means country/hovtest=levene;
run;
quit;
```

The diagnostics plots do not support normal distribution assumption (residual plot has bell-shaped, and residuals follow diagonal line). The p-value <0.001, we reject the hypothesis of equal variances between countries. Since the normal distribution assumption is violated, we cannot confirm whether average of moment magnitudes is different between countries for this case.



The GLM Procedure

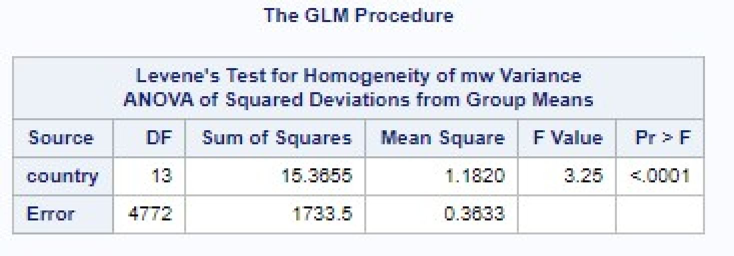
Levene's Test for Homogeneity of mw Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
country	13	29.8156	2.2935	5.73	<.0001
Error	23722	9493.7	0.4002		



Observations	23741
Parameters	17
Error DF	23724
MSE	0.3021
R-Square	0.0594
Adj R-Square	0.0588

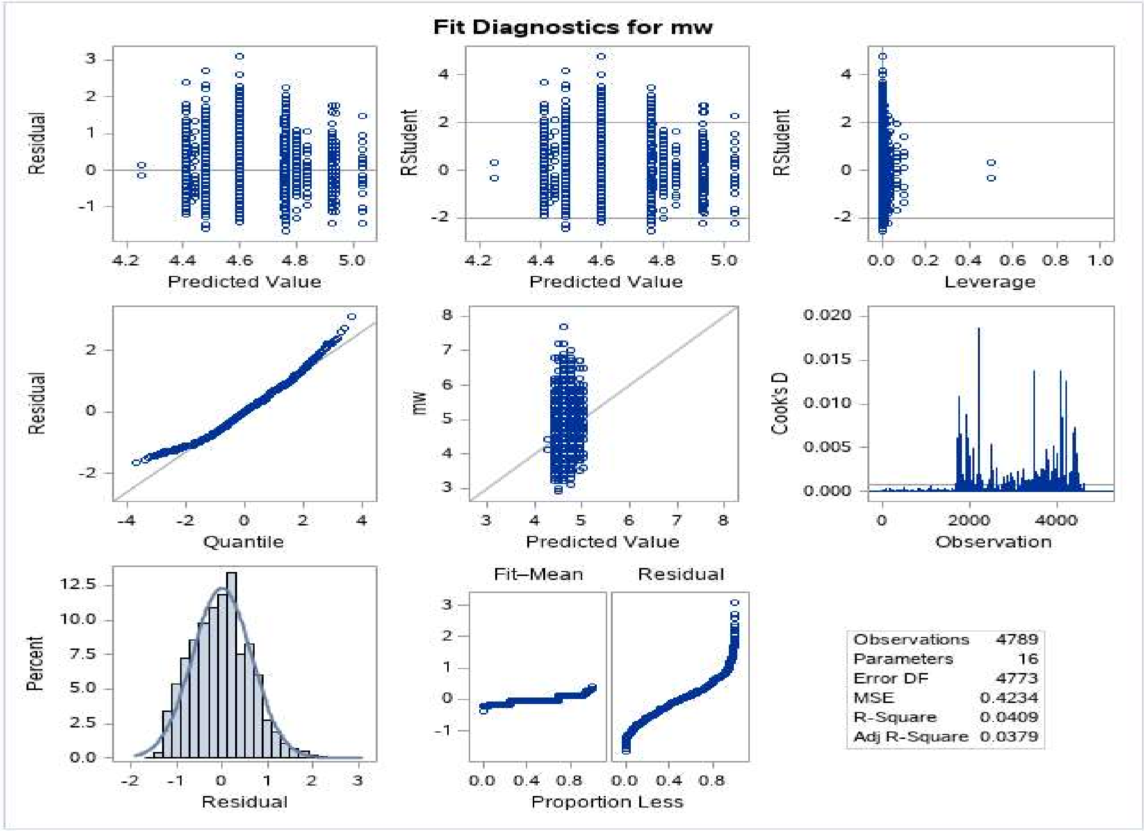
Note that if we decide to work with only valid moment magnitudes (4789 obs) as shown below. The QQplot and residual plot look better and may be considered to satisfy normal distribution (not strong) but the variances are different. In that case we could use Welch's variance-weighted one-way ANOVA, and the results provide evidence to reject hypothesis of equal moment magnitude (average) for all countries

```
proc glm data=work.valid_mw_only plots(maxpoints=24000)=diagnostics;
        class country;
        model mw=country;
        means country/hovtest=levene;
run;
```



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	15.3655	1.1820	3.25	<.0001
Error	4772	1733.5	0.3633		



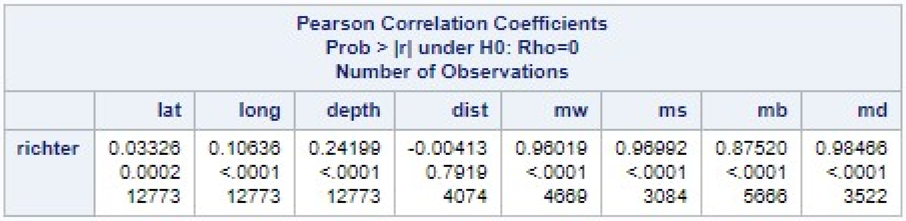
Observations	4789
Parameters	16
Error DF	4773
MSE	0.4234
R-Square	0.0409
Adj R-Square	0.0379

For this case (use only 4789 valid values of mw), if we transform **mw** to **log(mw),** we have the same result (but this time is the average of **log(mw)**, not average(mw)), however the normal distribution assumption complies better

Part C: *Fit a regression with "richter" as the response and consider the other variables in the dataset as potential explanatory variables, but do not use the variable "id" or the variable "xm"*

Since null values dominate in variables **mw**, **ms**, **md**, **mb**, **richter** then if we replace missing values by value of **xm**, the replaced values are the same for **mw**, **ms**, **md**, **mb**, **richter** for a lot of cases and that impacts the analysis results. Thus, we use only records that are available for all five variables. Following that direction, we use **proc corr** procedure to calculate correlations between **richter** and other numerical variables. From the correlation table, it is highly likely that **mw**, **ms**, **mb**, **md**, and may be **depth** variables are the predictors of **richter**.

```
%let var_numerical=lat long depth dist mw ms mb md;
proc corr data=work.earthquakes;
        var &var_numerical;
        with richter;
run;
```

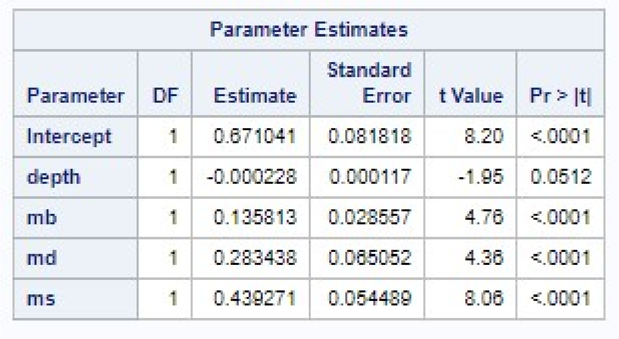


Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	lat	long	depth	dist	mw	ms	mb	md
richter	0.03326 0.0002 12773	0.10636 <.0001 12773	0.24199 <.0001 12773	-0.00413 0.7919 4074	0.96019 <.0001 4669	0.96992 <.0001 3084	0.87520 <.0001 5666	0.98466 <.0001 3522

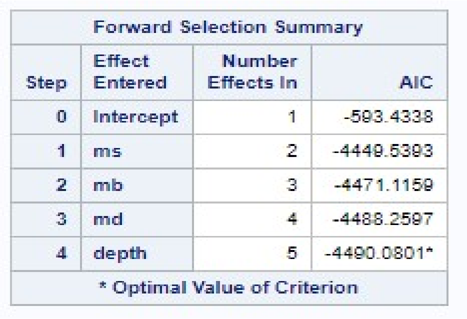
However, **ms**, **mw**, **mb**, **md** may face problems of collinearity, so we should go further by using procedure **glmselect**. That give the optimal model which includes **depth**, **ms**, **mb**, **md** (so **mw** is removed, it may be due to collinearity matter). From the "Fit Criteria" Chart, we can see that AIC, AICC, Adj-R-sq also support to include **ms**, **mb**, **md**, while **SBC**, p-values(p-value of ms is at 0.0512) support to exclude **depth** predictor. Therefore, we go with model which includes predictors **ms**, **mb** and **md** as it is simpler and the performance is not much different.

```
%let var_all=lat long depth dist mw ms mb md country;
proc glmselect data=work.earthquakes plots=all;
        class country;
        model richter=&var_all/selection=forward select=AIC showpvalues;
run;
```



Parameter Estimates

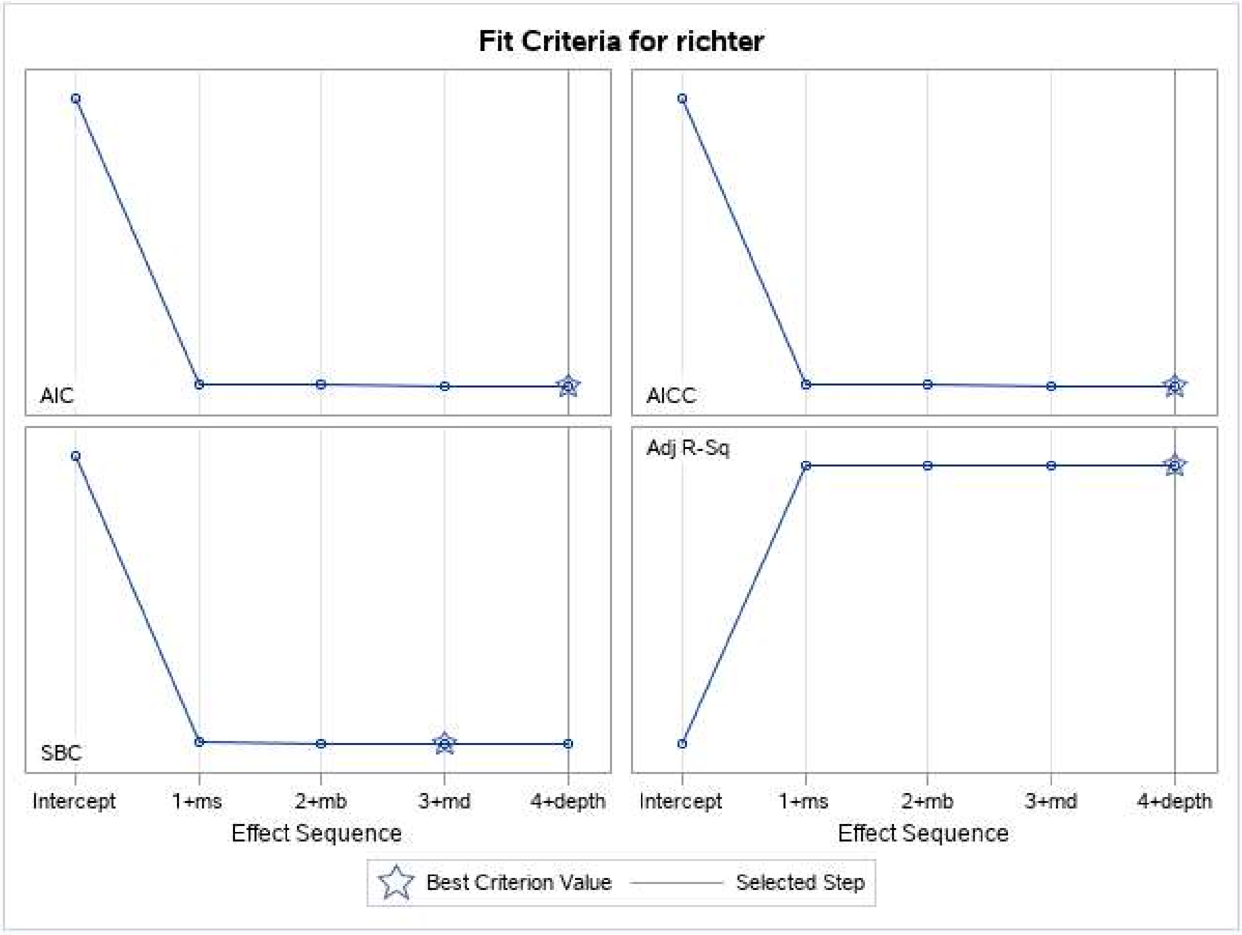
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.671041	0.081818	8.20	<.0001
depth	1	-0.000228	0.000117	-1.95	0.0512
mb	1	0.135813	0.028557	4.76	<.0001
md	1	0.283438	0.065052	4.36	<.0001
ms	1	0.439271	0.054489	8.06	<.0001



Forward Selection Summary

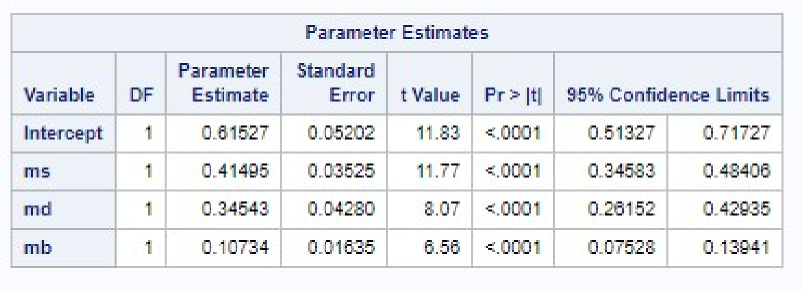
Step	Effect Entered	Number Effects In	AIC
0	Intercept	1	-593.4338
1	ms	2	-4449.5393
2	mb	3	-4471.1159
3	md	4	-4488.2597
4	depth	5	-4490.0801*

* Optimal Value of Criterion



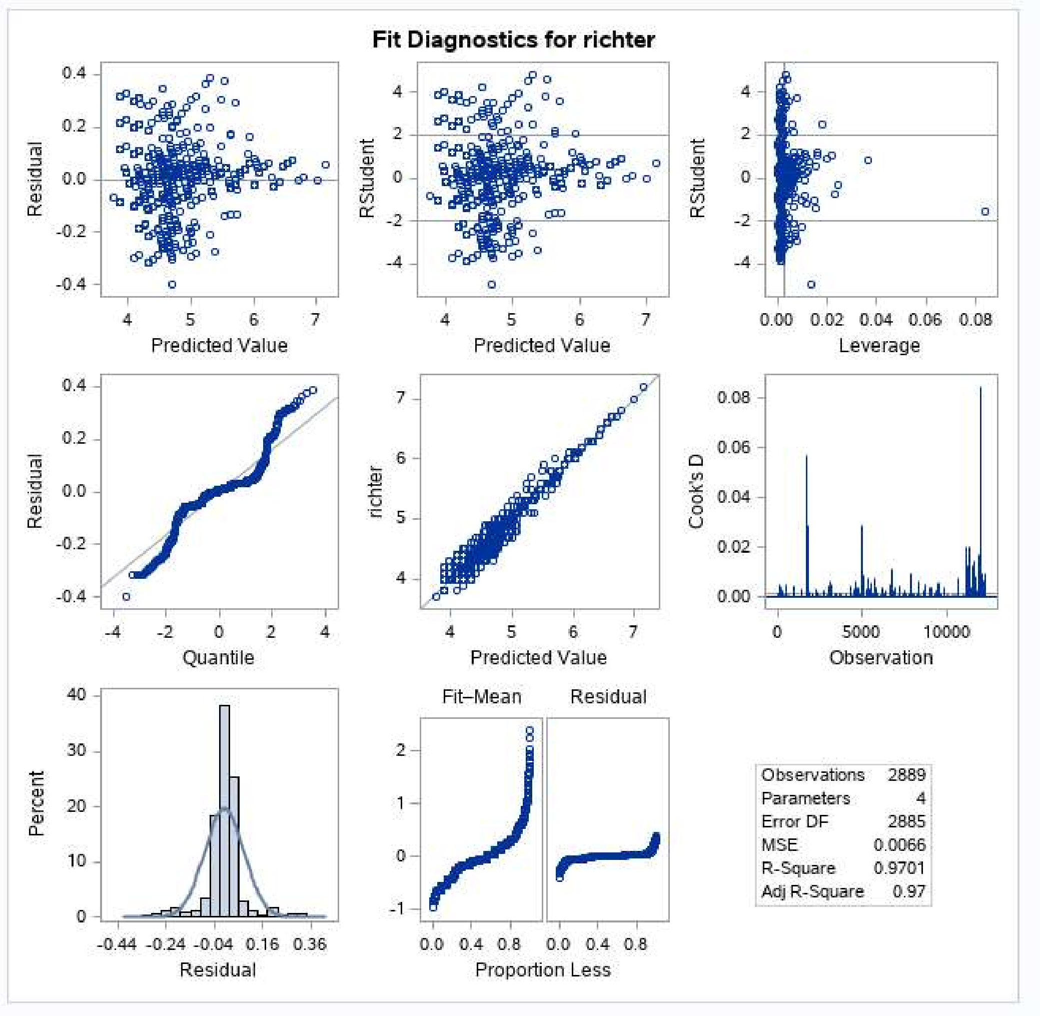
Below are diagnostic plots of final model which indicate that all assumptions seem hold, residuals distribute randomly around horizontal line 0, the plot richter against predicted values follow a diagonal line. QQplot shows that points follow diagonal line (not perfectly).

```
proc reg data=work.earthquakes;
        model richter=ms md mb/ clb;
run;
quit;
```



Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	0.61527	0.05202	11.83	<.0001	0.51327	0.71727
ms	1	0.41495	0.03525	11.77	<.0001	0.34583	0.48406
md	1	0.34543	0.04280	8.07	<.0001	0.26152	0.42935
mb	1	0.10734	0.01635	6.56	<.0001	0.07528	0.13941



Observations	2889
Parameters	4
Error DF	2885
MSE	0.0066
R-Square	0.9701
Adj R-Square	0.97

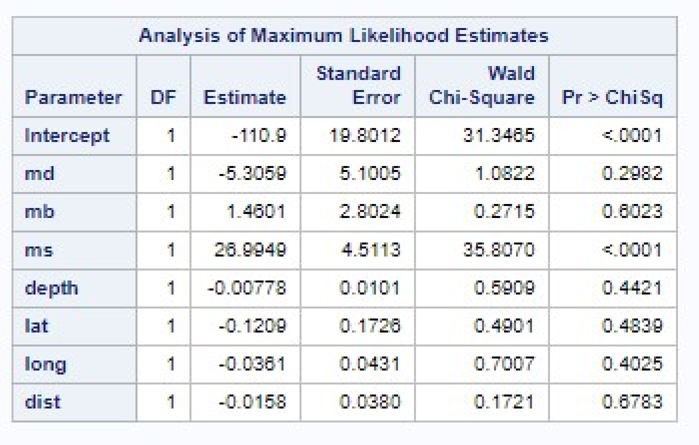
Part D: *A magnitude of 5 and above on the Richter scale is considered to be a moderate or stronger earthquake, causing damage and loss of life. Consider a new variable "serious", where the value of "serious" is 1 if the corresponding Richter scale value is 5 or more and 0 if the corresponding Richter value is below 5. Fit a regression with "serious" as the response and consider the other variables in the dataset as potential explanatory variables, but do not use the variables "id", "mw", "richter" or "xm".*

We first set up the variable **serious** as following:

```
data work.earthquakes;
        set work.earthquakes;
        if richter >=5 then serious=1;
                else if richter <5 and richter>0 then serious=0;
                else if serious=.;/*null value*/
run;
```

Since serious is a binary variable, so we will use logistic regression to fit for **serious** response. We start with full model, which include all numerical variables.

```
proc logistic data=work.earthquakes plots(only)=(effect oddsratio);
        model serious(event='1')=md mb ms depth lat long dist/clodds=pl;
run;
```

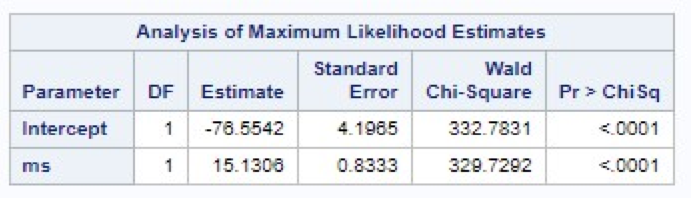


Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-110.9	19.8012	31.3465	<.0001
md	1	-5.3059	5.1005	1.0822	0.2982
mb	1	1.4601	2.8024	0.2715	0.6023
ms	1	26.9949	4.5113	35.8070	<.0001
depth	1	-0.00778	0.0101	0.5909	0.4421
lat	1	-0.1209	0.1726	0.4901	0.4839
long	1	-0.0361	0.0431	0.7007	0.4025
dist	1	-0.0158	0.0380	0.1721	0.6783

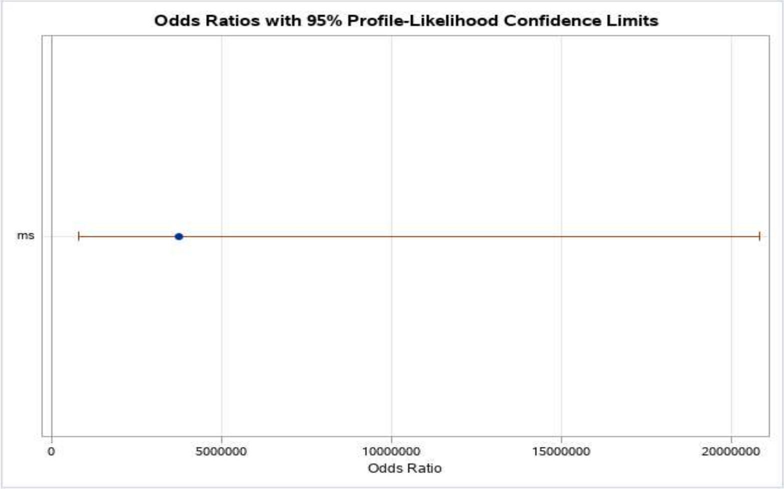
Using p-value, we see that only coefficient of **ms** is significant, so we just keep ms as the predictor for our model. This is a simple method to select predictor for logistic regression, otherwise it is too complicated and may be beyond the content of this course. Let us examine the results of model with only predictor **ms**. The odds ratio is extremely high and wide range due for interpretation and visualization (as shown below plots). We may need some transformations for the predictor. An exponential transformation will help to reduce coefficient of ms and then reduce odds ratio in this case (see next page).

```
proc logistic data=work.earthquakes plots(only)=(effect oddsratio);
        model serious(event='1')=ms/clodds=pl;
run;
```



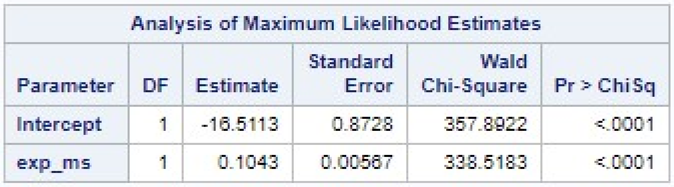
Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-76.5542	4.1965	332.7831	<.0001
ms	1	15.1306	0.8333	329.7292	<.0001



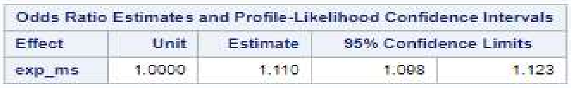
Here is the code to perform transformation and fit the model. Now, we have better confidence interval for odds ratio.

```
data work.earthquakes;
        set work.earthquakes;
        exp_ms=exp(ms);
run;

proc logistic data=work.earthquakes plots(only)=(effect oddsratio);
        model serious(event='1')=exp_ms /clodds=pl;
run;
```



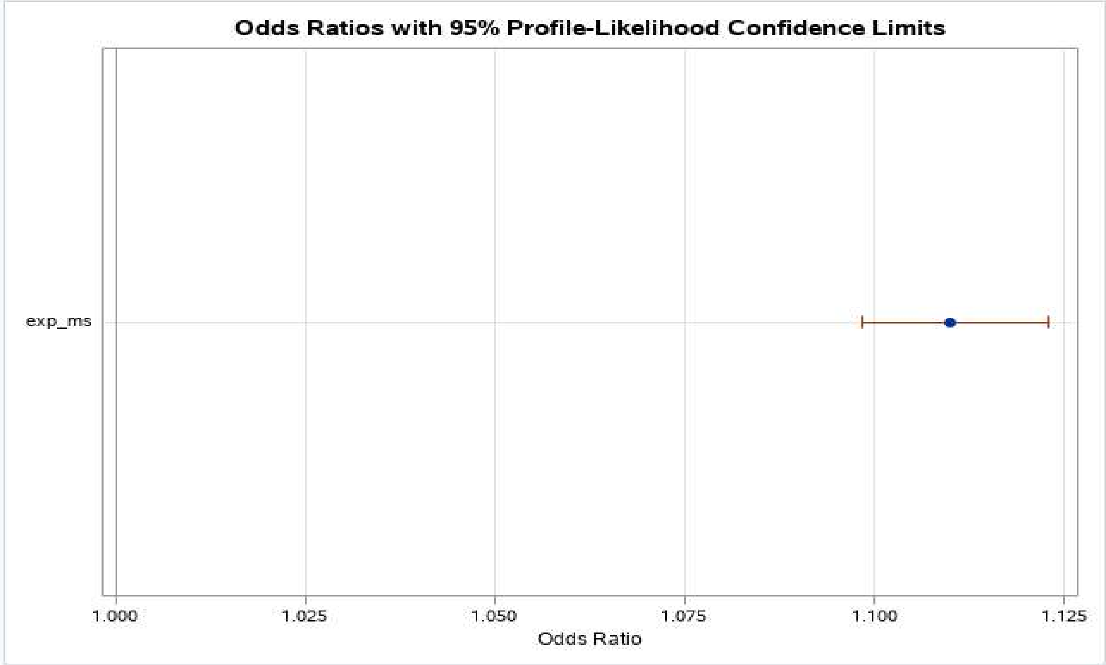
Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-16.5113	0.8728	357.8922	<.0001
exp_ms	1	0.1043	0.00567	338.5183	<.0001



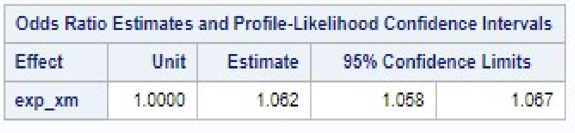
Odds Ratio Estimates and Profile-Likelihood Confidence Intervals

Effect	Unit	Estimate	95% Confidence Limits	
exp_ms	1.0000	1.110	1.098	1.123



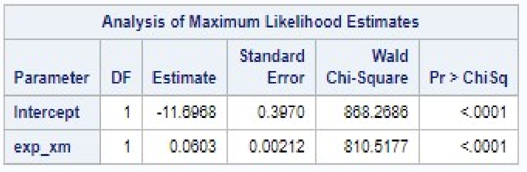
Part E: *Fit a regression with "serious" as the response and "xm" as the only explanatory variable. How does this model compare to your model from part d) in terms of out-of-sample predictive performance (i.e. the model's ability to predict data on which it has not been built)?*
It is similar to part D, we need exponential transformation of **xm** in order to get good fit.

```
data work.earthquakes;
	set work.earthquakes;
	exp_xm=exp(xm);
run;

proc logistic data=work.earthquakes plots(only)=(effect oddsratio);
	model serious(event='1')=exp_xm /clodds=pl;
run;
```



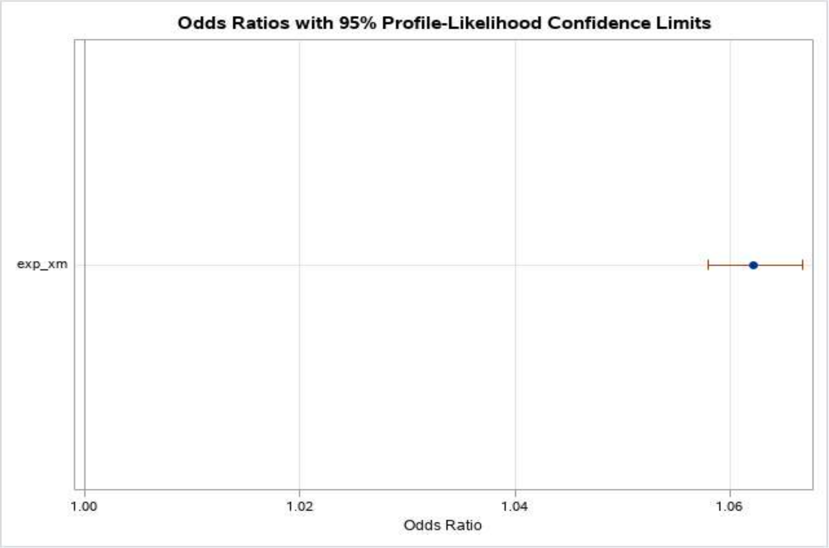
Odds Ratio Estimates and Profile-Likelihood Confidence Intervals

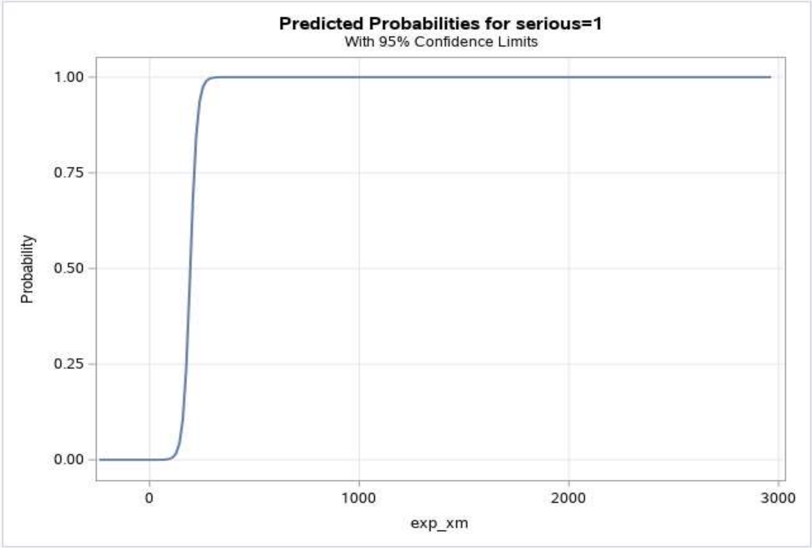
Effect	Unit	Estimate	95% Confidence Limits	
exp_xm	1.0000	1.062	1.058	1.067



Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-11.6968	0.3970	868.2686	<.0001
exp_xm	1	0.0603	0.00212	810.5177	<.0001





In order to evaluate the model in part D and E, we need to build a training sample and a test sample then train both models on training sample, after that we evaluate their performance on test sample by comparing ROC curves.

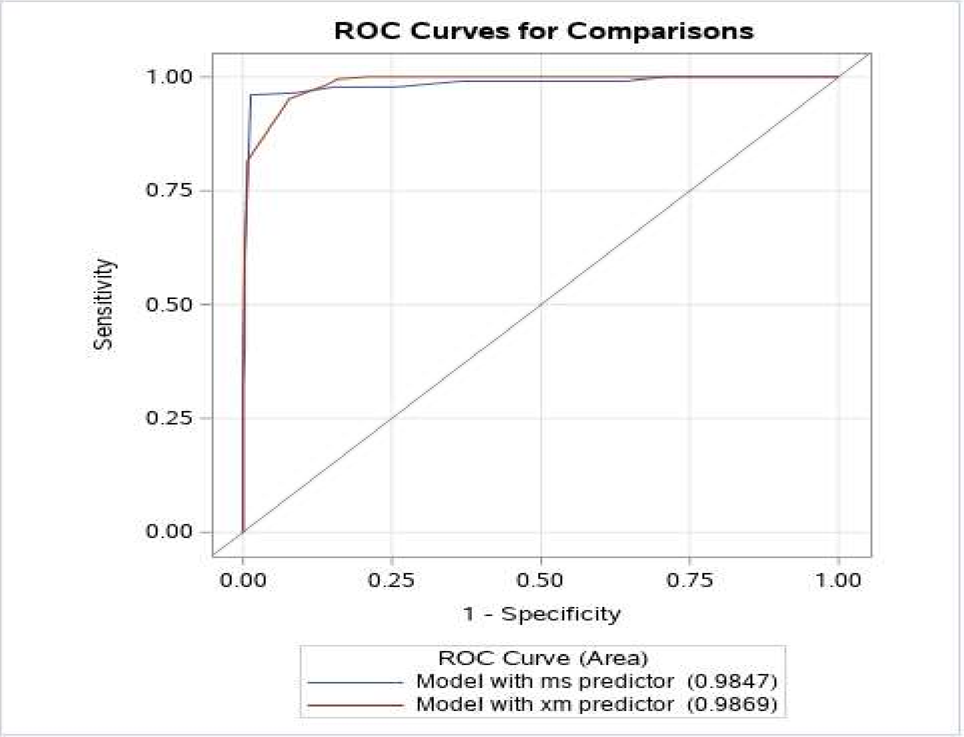
First we set up a full sample which contain all valid **ms** and **serious** entries and then create a train sample which equal to 70% of full sample, the remaining part is for test sample. The following code perform the task.

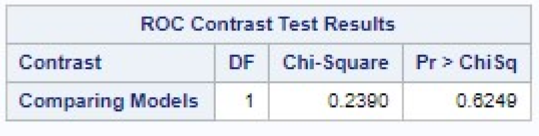
```
/*set up full sample*/
proc sql;
        create table full_sample as
        select * from work.earthquakes where ms is not missing and serious is not missing;
run;

/*rank before take sample*/
proc sort data=work.full_sample out=work.full_sample_sort;
        by serious;
run;
/*take train sample of 70% full sample*/
proc surveyselect noprint data=work.full_sample_sort samprate=0.7
        outall out=work.earthquakes_sampling;
        strata serious;
run;
/*separate the sampling into train sample and test sample*/
data work.train(keep=xm ms mb md exp_xm exp_ms serious) work.test(keep=xm ms mb md exp_xm
exp_ms serious);
        set work.earthquakes_sampling;
        if selected then output work.train;
        else output work.test;
run;
```

We now train and test two model with created train and test sample, after that we overlay two ROC curves of two models in order to compare their performance. Below code perform the tasks:

```
proc logistic data=work.train;
        model serious(event='1')=exp_ms;
        score data=work.test out=testAssess(rename=(p_1=p_ms)) outroc=work.roc;
run;
proc logistic data=work.train;
        model serious(event='1')=exp_xm;
        score data=work.testAssess out=testAssess(rename=(p_1=p_xm)) outroc=work.roc;
run;
proc logistic data=work.testAssess;
        model serious(event='1')=p_ms p_xm/nofit;
        roc "Model with ms predictor" p_ms;
        roc "Model with xm predictor" p_xm;
        roccontrast "Comparing Models";
run;
```

The ROC curves show no much differences between the two models. The ROC Contrast Test also do not provide evidence of a difference between them. So, we conclude there is no difference in the performance between two models. Both of them perform excellently with AUROC approximately at 0.98. It should be noted that **surveyselect** procedure generates different train/test sample each time we call it. Thus, the performances of models could change slightly according to.





ROC Contrast Test Results			
Contrast	DF	Chi-Square	Pr > ChiSq
Comparing Models	1	0.2390	0.6249

IV. Conclusion

Throughout the analysis process, we can see that thee data quality is not good which contains numerous zeros entries and null entries. This could lead to make confusion during analysis process. However, those zeros values are actually null values which already replaced by 0. So, by changing those zeros back to nulls values, it will help to avoid potential confusion.

This report also shows that earthquakes magnitudes on different measures are quite close together, the differences are usually from 0 to 0.3. And these magnitudes on different scales are highly correlated, so in case we do not have data on one scale, we could use the data available on other scales as an approximate value. The magnitudes have very little impacts by depth of earthquakes.

Some countries like Turkey, Greece and Mediterranean have very high number of earthquakes, while countries like Israel, Albania, Egypt have only one or two times. Average of earthquake-magnitudes in different countries **may** be different. In some cases, transformation of continuous variables will make the regression easier and more convenient for visualization.