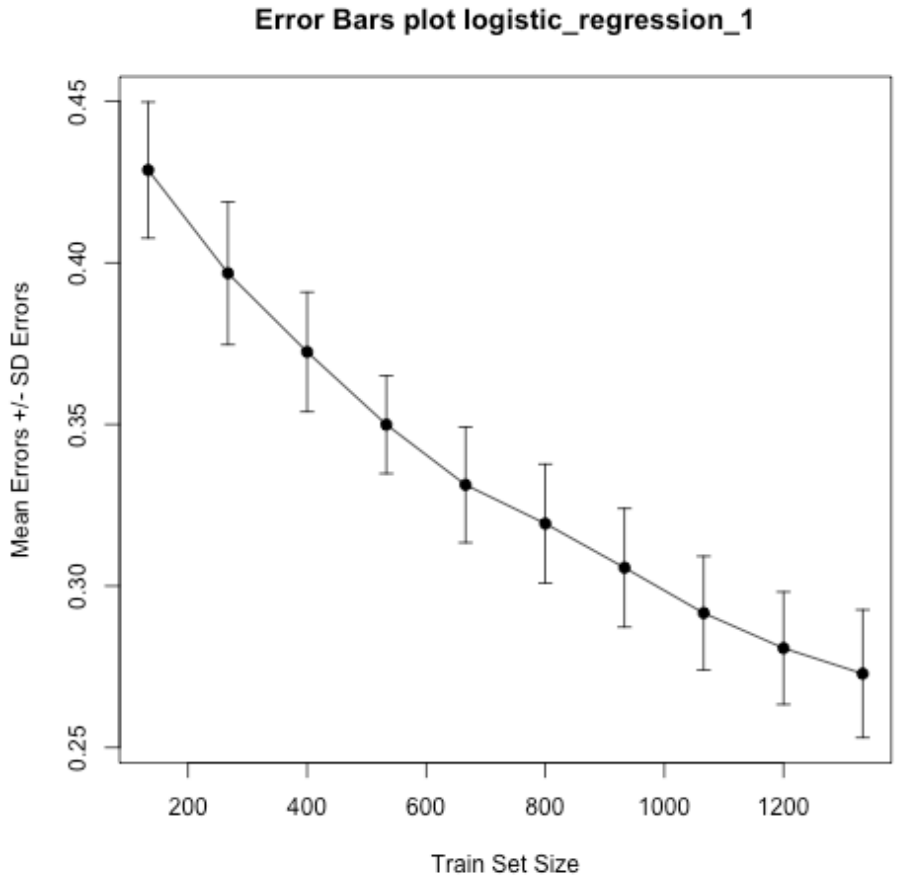
MACHINE LEARNING PROGRAMMING PROJECT 3:

This Project was completed in RStudio.

We were given 4 datasets for logistic Regression, poisson regression and ordinal regression.

1. Logistic Regression:

a) Dataset: A.csv



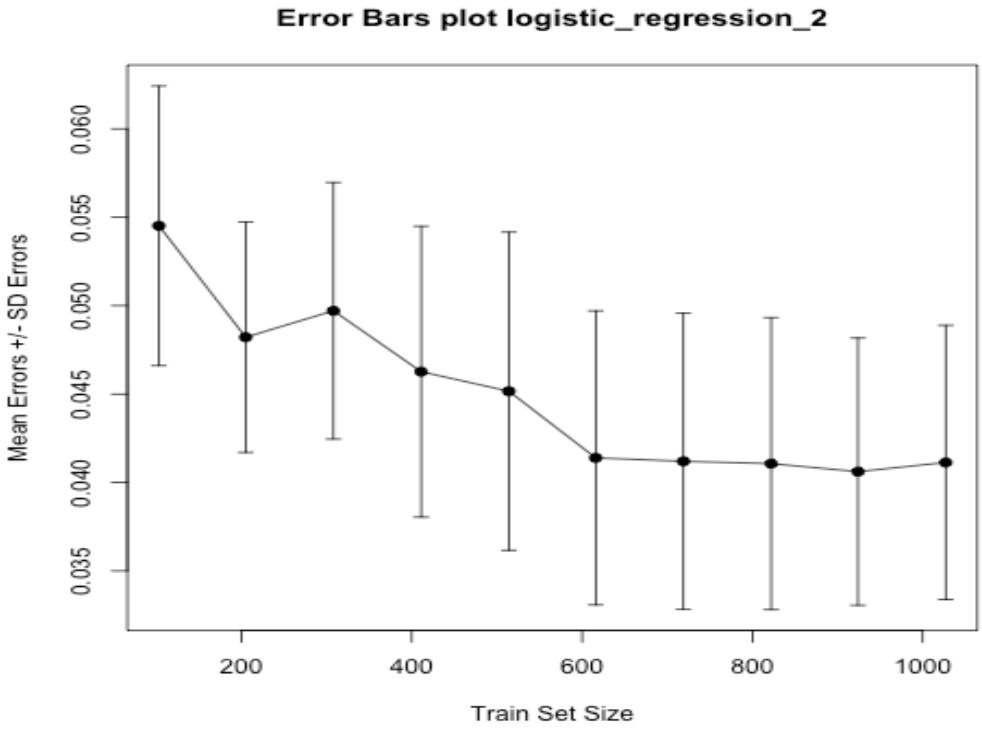
	trainsetsize	IterationsSample_mean	Run_time_mean	error_mean
1	133	3	0.01099687	0.4237881
2	267	3	0.02450873	0.3935532
3	400	3	0.04793542	0.3662169
4	533	3	0.08787071	0.3486757
5	666	3	0.12718617	0.3327836
6	800	3	0.18218427	0.3228386
7	933	3	0.23501536	0.3135932
8	1066	3	0.30481384	0.2957521
9	1200	3	0.44709725	0.2861569
10	1333	3	0.71663645	0.2795602

This is a dataframe consisting of all the statistics for the Mean number of Iterations, Run time mean and error mean when this process was replicated 30 times for random train data and test data.

For this dataset the mean error rate starts from approx. **0.42** of the test set for the smallest size of train set i.e 133 and **reduces** to as low as approx. **0.28** of the test set. This follows the general trend that as the training set size increases, the error rate on the test set decreases.

The mean number of iterations are 3 for each epoch.
The mean run time of the 30 iterations for each training set size starts from 0.01 seconds and ends at 0.71 seconds

b) Dataset: USPS.csv

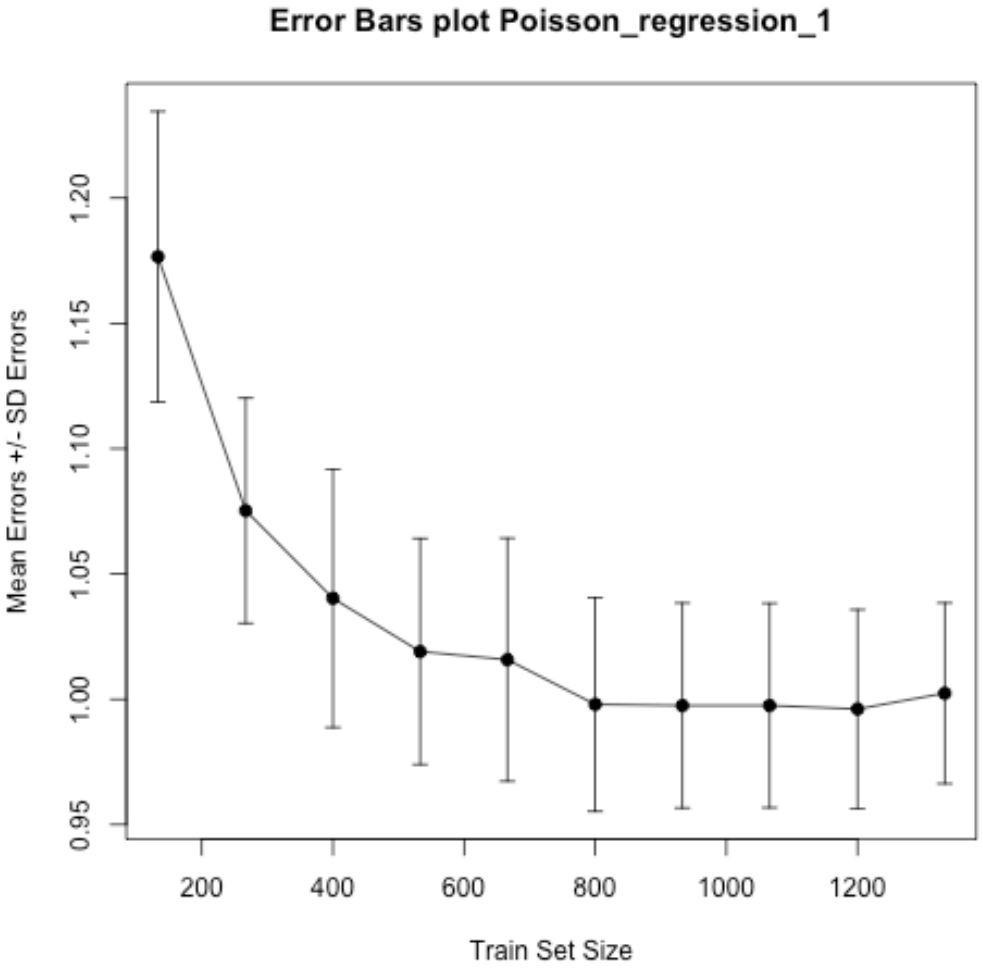


	trainsetsize	IterationsSample_mean	Run_time_mean	error_mean
1	103	5.000000	0.3002251	0.05451592
2	205	6.000000	0.4628332	0.04821313
3	308	6.000000	0.6095601	0.04970760
4	411	6.000000	0.7869688	0.04626381
5	514	6.000000	1.0619109	0.04515919
6	616	6.200000	1.4015498	0.04139051
7	719	6.566667	1.8806573	0.04119558
8	822	7.000000	2.4242388	0.04106563
9	924	7.000000	2.9604102	0.04061079
10	1027	7.000000	3.5273525	0.04113060

For this dataset the mean error rate starts from approx. **0.05** of the test set for the smallest size of train set i.e 103 and **reduces** to as low as approx. **0.04** of the test set. This follows the general trend that as the training set size increases, the error rate on the test set decreases.

The mean number of iterations are starts from 5 and go to 7 for the largest training size. So, as the training set increases, the iterations required to converge also increases
The mean run time of the 30 iterations for each training set size starts from 0.05 seconds and goes upto 3.52 seconds.

2. Poisson Regression: Dataset: AP.csv

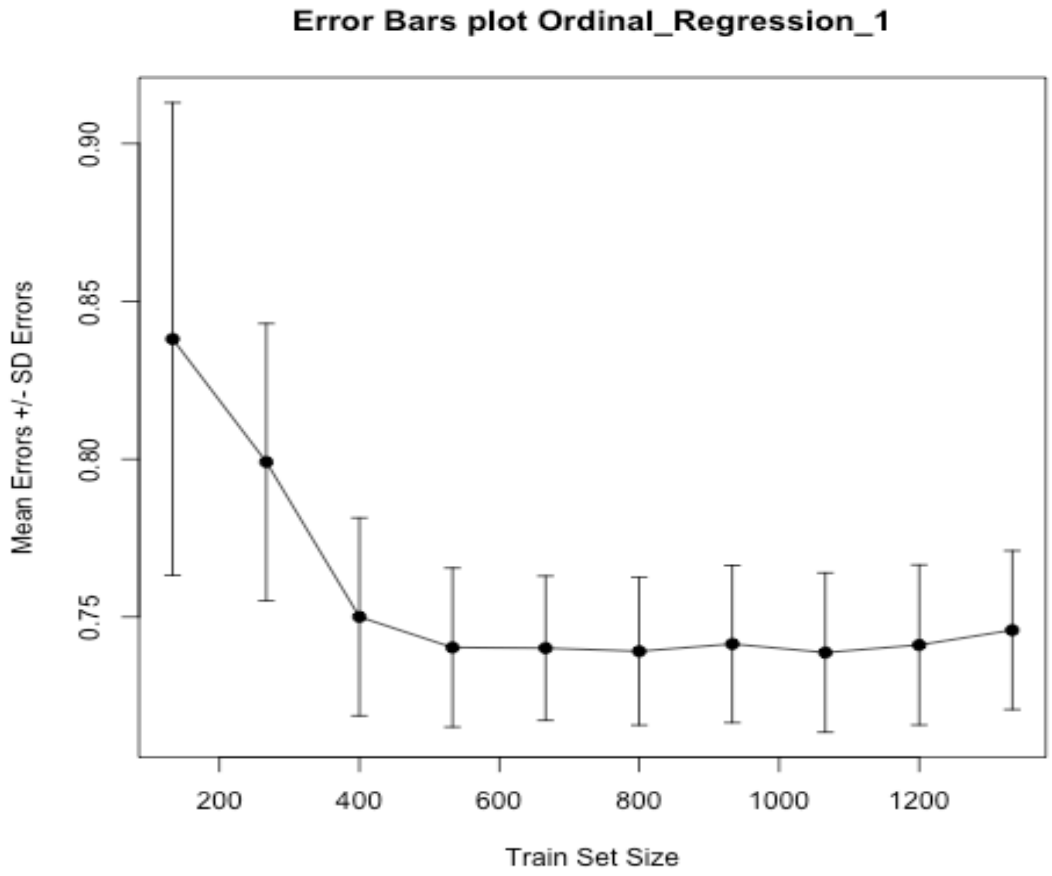


	trainsetsize	IterationsSample_mean	Run_time_mean	error_mean
1	133	7.066667	0.02397067	1.1766117
2	267	7.466667	0.06098317	1.0752624
3	400	7.266667	0.13339148	1.0402799
4	533	7.433333	0.21310761	1.0190405
5	666	7.333333	0.31043677	1.0157921
6	800	7.000000	0.44199229	0.9979510
7	933	7.000000	0.60330033	0.9975012
8	1066	7.300000	0.85982063	0.9975012
9	1200	7.266667	1.12485040	0.9961019
10	1333	7.333333	1.27183779	1.0023988

For this dataset the mean error rate starts from approx. **1.17** of the test set for the smallest size of train set i.e 133 and **reduces** to as low as approx. **0.99** of the test set. This follows the general trend that as the training set size increases, the error rate on the test set decreases.

The mean number of iterations are 7.06 for the smallest training set and go to 7 for the largest training size. So, as the training set increases, the iterations required to converge also increases
The mean run time of the 30 iterations for each training set size starts from 0.02 seconds and goes upto 1.27 seconds. So as the training set size increases the mean run time also increases.

3. Ordinal Regression



	trainsetsize	IterationsSample_mean	Run_time_mean	error_mean
1	133	3.133333	0.1661676	0.8380810
2	267	4.000000	0.4150538	0.7991004
3	400	4.000000	0.6681151	0.7500250
4	533	4.000000	0.8815354	0.7403298
5	666	4.000000	1.0098033	0.7401299
6	800	4.000000	1.5292627	0.7391304
7	933	4.000000	1.9559256	0.7414293
8	1066	4.000000	2.1187021	0.7387306
9	1200	4.033333	2.3160339	0.7411294
10	1333	4.133333	3.1498063	0.7458271

For this dataset the mean error rate starts from approx. **0.83** of the test set for the smallest size of train set i.e 133 and **reduces** to as low as approx. **0.74** of the test set. This follows the general trend that as the training set size increases, the error rate on the test set decreases.

The mean number of iterations are 3.13 for the smallest training set and go to 4.13 for the largest training size. So, as the training set increases, the iterations required to converge also increases

The mean run time of the 30 iterations for each training set size starts from 0.16 seconds and goes upto 3.14 seconds. So as the training set size increases the mean run time also increases.

Apart from individual functions for each of these distributions there is a GLM function which takes in argument the type of regression we want to perform.

4. For Model Selection part, I completed a linear search on the GLM Function with values of alpha between 0.1 to 50 and checked the values of mean error for each value of alpha. The lowest value of error rate was at the value of alpha= 0.01 and the test error rate goes down for the logistic regression dataset for the maximum training set size of 1333.