UNIVERSITY OF LEEDS

School of Mathematics

Declaration of Academic Integrity
for Individual Pieces of Work

I am aware that the University defines plagiarism as presenting someone else's work as your own. Work means any intellectual output, and typically includes text, data, images, sound or performance.

I promise that in the attached submission I have not presented anyone else's work as my own and I have not colluded with others in the preparation of this work. Where I have taken advantage of the work of others, I have given full acknowledgement. I have read and understood the University's published rules on plagiarism and also any more detailed rules specified at School or module level. I know that if I commit plagiarism I can be expelled from the University and that it is my responsibility to be aware of the University's regulations on plagiarism and their importance.

I re-confirm my consent to the University copying and distributing any or all of my work in any form and using third parties (who may be based outside the EU/EEA) to monitor breaches of regulations, to verify whether my work contains plagiarised material, and for quality assurance purposes.

I confirm that I have declared all mitigating circumstances that may be relevant to the assessment of this piece of work and that I wish to have taken into account. I am aware of the School's policy on mitigation and procedures for the submission of statements and evidence of mitigation. I am aware of the penalties imposed for the late submission of coursework.

Student Signature ______________________________ Date __**16 March 2020**__

Student Name _______**Nico Septianus**_______ Student Number _____**201380903**_____

Please note. When you become a registered student of the University at first and any subsequent registration you sign the following authorisation and declaration: "I confirm that the information I have given on this form is correct. I agree to observe the provisions of the University's Charter, Statutes, Ordinances, Regulations and Codes of Practice for the time being in force. I know that it is my responsibility to be aware of their contents and that I can read them on the University web site. I acknowledge my obligation under the Payment of Fees Section in the Handbook to pay all charges to the University on demand. I agree to the University processing my personal data (including sensitive data) in accordance with its Code of Practice on Data Protection http://www.leeds.ac.uk/dpa . I consent to the University making available to third parties (who may be based outside the European Economic Area) any of my work in any form for standards and monitoring purposes including verifying the absence of plagiarised material. I agree that third parties may retain copies of my work for these purposes on the understanding that the third party will not disclose my identity."'

PART 1: Regression

1.) Using the GLM function in R on medal_data, we obtained both 2008 and 2012 prediction such as,

Medal Count in 2008

```
Call:
glm(formula = Medal2008 ~ GDP + Population, data = medal_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-27.154   -4.856   -1.702    0.842   51.037

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.613e+00  1.506e+00   3.728 0.000395 ***
GDP         7.613e-03  7.353e-04  10.354 1.29e-15 ***
Population  8.435e-09  7.220e-09   1.168 0.246750
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 133.1455)

    Null deviance: 29595.1  on 70  degrees of freedom
Residual deviance:  9053.9  on 68  degrees of freedom
AIC: 553.72

Number of Fisher Scoring iterations: 2
```

Medal Count in 2012

```
Call:
glm(formula = Medal2012 ~ GDP + Population, data = medal_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-20.568   -5.961   -2.462    3.932   60.121

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.076e+00  1.500e+00   4.051 0.000133 ***
GDP         7.564e-03  7.325e-04  10.326 1.45e-15 ***
Population  5.247e-09  7.193e-09   0.729 0.468225
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 132.1562)

    Null deviance: 28402.8  on 70  degrees of freedom
Residual deviance:  8986.6  on 68  degrees of freedom
AIC: 553.19

Number of Fisher Scoring iterations: 2
```

2.) Next to check the consistency for each variable (Population and GDP) corresponding to the prediction. We perform 95% confidence interval test for each variable in each year.

2008 Data:

- GDP

```
"GDP 2008: 0.00614591191989105 0.00908035801420896"
```

The interval does not contain regression parameter 0 in it, which can be conclude that GDP is consistent towards the prediction in 2008.

- Population

```
"Population 2008: -5.97148075440265e-09 2.28411294839574e-08"
```

The interval contains regression parameter 0 in it, which means population inconsistent towards the prediction in 2008. This could be due to error or inefficiency in data points.

2012 Data:

- GDP

```
"GDP 2012: 0.00610231906043588 0.00902584306021206"
```

Similar to 2008, GDP is consistent towards the prediction in 2012. This can be seen from the interval which does not contain regression parameter 0.

- Population

```
"Population 2012: -9.10593444565584e-09 1.95994344112297e-08"
```

Similar to 2008, the variable population is inconsistent towards the prediction in 2012. We can see from the interval contains regression parameter 0.

From comparison in 2008 and 2012, can be concluded that GDP plays big part on predicting medal obtained for every country. However, population in a country does not always refer to medal obtained for the countries.

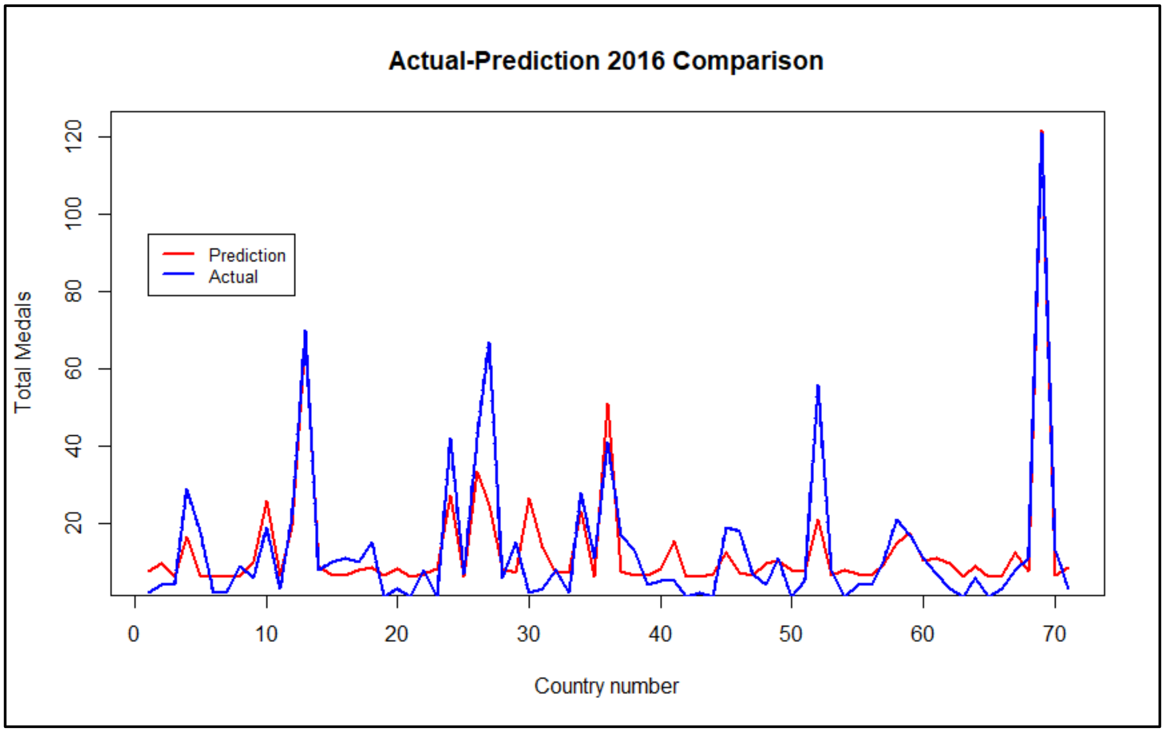
3.) For prediction 2016 medal from 2012, we use the predict function on the linear regression of 2012 medal count. Below there are countries number, countries name, prediction results for 2016 and the actual data from medal_data in 2016 respectively.

	Country	Prediction	Actual
1	Algeria	7.697937	2
2	Argentina	9.660081	4
3	Armenia	6.170773	4
4	Australia	16.572245	29
5	Azerbaijan	6.603459	18
6	Bahamas	6.136872	2
7	Bahrain	6.248223	2
8	Belarus	6.542817	9
9	Belgium	10.002805	6
10	Brazil	25.819025	19
11	Bulgaria	6.519486	3
12	Canada	19.390152	22
13	China	68.348721	70
14	Colombia	8.828562	8
15	Croatia	6.581570	10
16	Cuba	6.595043	11
17	Czech Republic	7.759147	10
18	Denmark	8.621790	15
19	Dominican Republic	6.545939	1
20	Egypt	8.242126	3
21	Estonia	6.250779	1
22	Ethiopia	6.758360	8
23	Finland	8.117037	1
24	France	27.394391	42
25	Georgia	6.208237	7
26	Germany	33.513444	42
27	Great Britain	24.795509	67
28	Greece	8.392310	6
29	Hungary	7.187558	15
30	India	26.568161	2
31	Indonesia	13.728427	3
32	Iran	7.130986	8
33	Ireland	7.743689	2
34	Italy	22.996238	28
35	Jamaica	6.204280	11
36	Japan	51.125438	41
37	Kazakhstan	7.572239	17
38	Kenya	6.532974	13
39	Lithuania	6.416057	4
40	Malaysia	8.332637	5
41	Mexico	15.404428	5
42	Moldova	6.147716	1
43	Mongolia	6.155200	2
44	Morocco	7.004824	1
45	Netherlands	12.489418	19
46	New Zealand	7.087823	18
47	North Korea	6.368698	7
48	Norway	9.776986	4
49	Poland	10.169817	11
50	Portugal	7.928127	1
51	Romania	7.535952	5
52	Russian Federation	20.878996	56
53	Serbia	6.454139	8
54	Singapore	7.916400	1
55	Slovakia	6.830738	4
56	Slovenia	6.461612	4
57	South Africa	9.429469	10
58	South Korea	14.774385	21
59	Spain	17.595080	17
60	Sweden	10.196346	11
61	Switzerland	10.925493	7
62	Taiwan	9.722857	3
63	Tajikistan	6.165369	1
64	Thailand	9.034171	6
65	Trinidad and Tobago	6.253046	1
66	Tunisia	6.478984	3
67	Turkey	12.315867	8
68	Ukraine	7.565541	11
69	United States	121.892569	121
70	Uzbekistan	6.572002	13
71	Venezuela	8.612422	3

The country number will be used to track for finding the trend and outliers. Therefore, for question 4 need to refer to these tables.

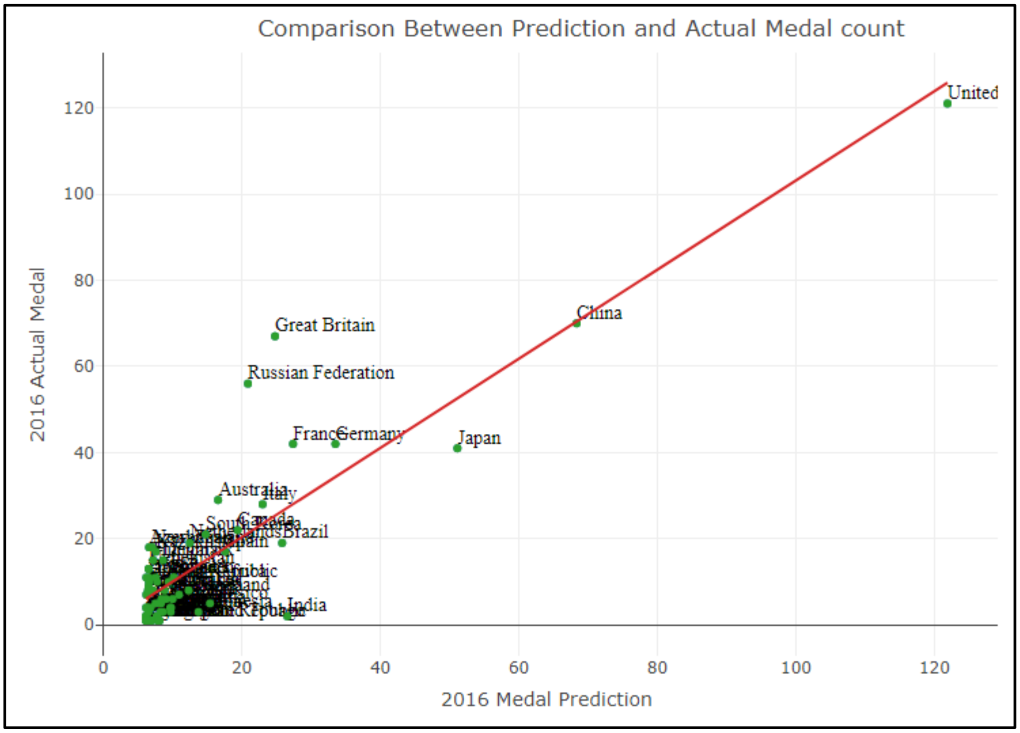
4.) To compare between both predicted and actual medal data in 2016. We plot 3 different graphs. There will be line graph comparison, scatter plot with transformation axes and boxplot to find the outliers.

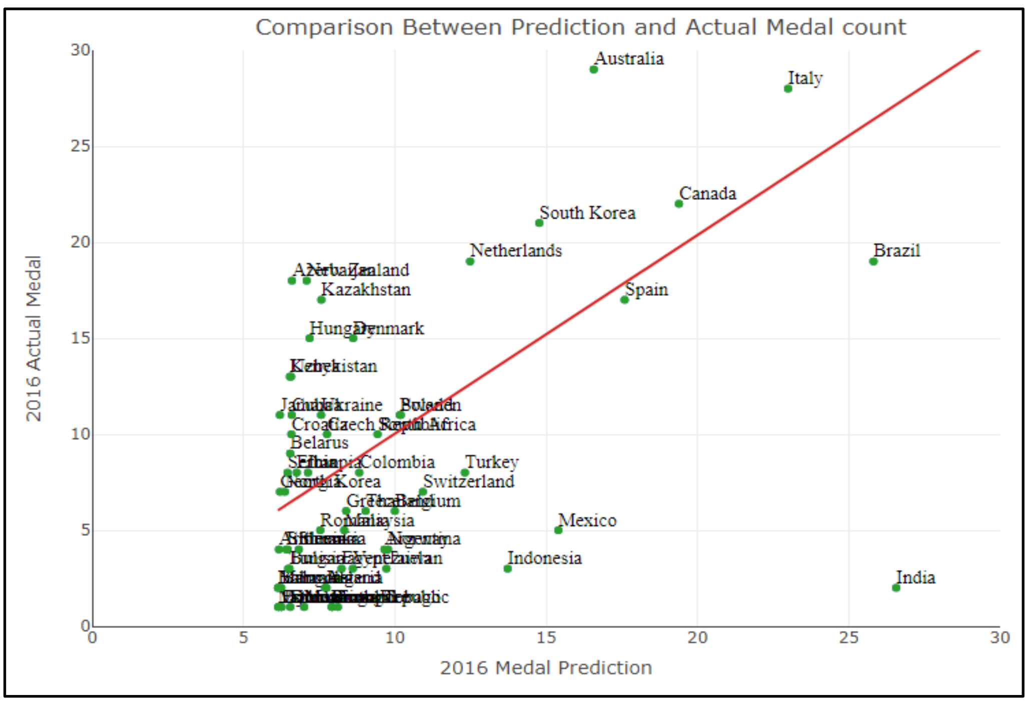
- Line graph



From the graphs we can see the comparison for predicted and actual medal count. The x-axis refers to country number and y-axis is total medal obtained. We can see more or less the predicted graph (red) follows the actual data (blue). For instance, the country number 69 (USA) reached the peak at 121 medals is predicted same with the actual data. Another good prediction come from country number 12 (China) which receive roughly 68 medals. Some other low receiving medals countries also show some good prediction such as country number 14 (Columbia), 57 (South Africa), 59 (Spain) and more.

- Scatter Plot





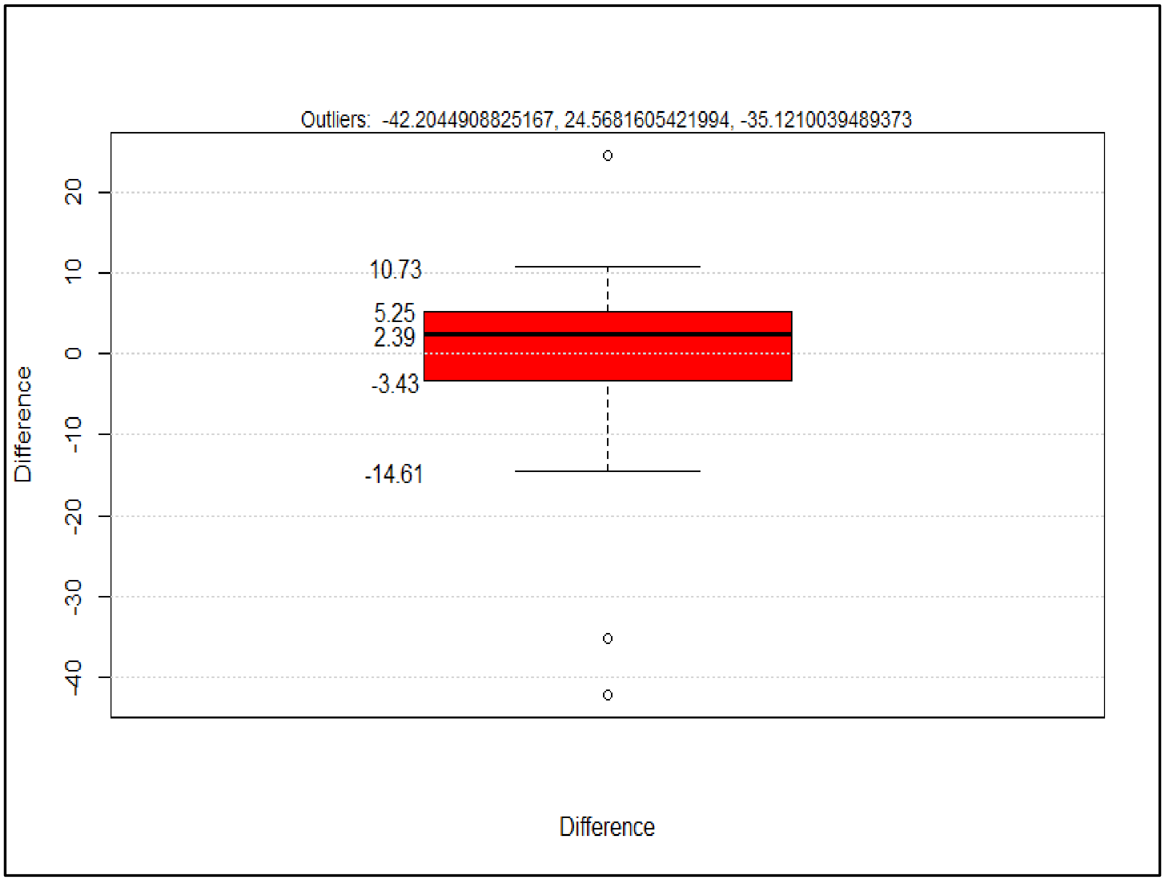
The red line is the linear regression in 2016 where x = y where x is predicted medal data and y is actual data. As we can see the linear regression go in same way with the previous line chart which China and USA have the majority medals and they fit nicely with the regression.

- Boxplot

For boxplot we use the difference between the predicted data and actual data or we can say the error in the regression line from each predicted value. This in order to find the outlier from trend (huge difference). First, we make data frame of the difference (predicted – actual) for each country,

	Country	Difference
1	Algeria	5.6979373
2	Argentina	5.6600810
3	Armenia	2.1707729
4	Australia	-12.4277552
5	Azerbaijan	-11.3965415
6	Bahamas	4.1368719
7	Bahrain	4.2482230
8	Belarus	-2.4571829
9	Belgium	4.0028050
10	Brazil	6.8190248
11	Bulgaria	3.5194861
12	Canada	-2.6098481
13	China	-1.6512793
14	Colombia	0.8285623
15	Croatia	-3.4184296
16	Cuba	-4.4049566
17	Czech Republic	-2.2408534
18	Denmark	-6.3782098
19	Dominican Republic	5.5459390
20	Egypt	5.2421261
21	Estonia	5.2507787
22	Ethiopia	-1.2416397
23	Finland	7.1170365
24	France	-14.6056091
25	Georgia	-0.7917633
26	Germany	-8.4865558
27	Great Britain	-42.2044909
28	Greece	2.3923104
29	Hungary	-7.8124415
30	India	24.5681605
31	Indonesia	10.7284275
32	Iran	-0.8690137
33	Ireland	5.7436890
34	Italy	-5.0037617
35	Jamaica	-4.7957204
36	Japan	10.1254379
37	Kazakhstan	-9.4277608
38	Kenya	-6.4670260
39	Lithuania	2.4160571
40	Malaysia	3.3326367
41	Mexico	10.4044280
42	Moldova	5.1477165
43	Mongolia	4.1551999
44	Morocco	6.0048243
45	Netherlands	-6.5105821
46	New Zealand	-10.9121769
47	North Korea	-0.6313021
48	Norway	5.7769863
49	Poland	-0.8301831
50	Portugal	6.9281268
51	Romania	2.5359518
52	Russian Federation	-35.1210039
53	Serbia	-1.5458613
54	Singapore	6.9163999
55	Slovakia	2.8307385
56	Slovenia	2.4616121
57	South Africa	-0.5705314
58	South Korea	-6.2256153
59	Spain	0.5950801
60	Sweden	-0.8036537
61	Switzerland	3.9254927
62	Taiwan	6.7228569
63	Tajikistan	5.1653692
64	Thailand	3.0341710
65	Trinidad and Tobago	5.2530464
66	Tunisia	3.4789836
67	Turkey	4.3158671
68	Ukraine	-3.4344586
69	United States	0.8925686
70	Uzbekistan	-6.4279980
71	Venezuela	5.6124222

Next, we plot the boxplot from these data to find the outliers from trend and we can refer back to the line chart and see which country number that have the huge difference or outliers.



We obtained there are 3 countries that contains outliers, such as country number

- 27 (Great Britain = -42,2),
- 30 (India = 24,56),
- 52 (Russian Federation = -35,12)

This shows overall the model between actual and predicted data are quite close with only 3 outliers in it.

PART 2: Model Selection

1.) First, we name each model for question 2 and 3.

- Model 1 (Population only)

```
Call:
glm(formula = Medal2012 ~ Population, data = medal_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-54.311   -8.421   -5.507    1.786   81.059

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.032e+01  2.295e+00   4.498  2.7e-05 ***
Population  4.026e-08  1.009e-08   3.990 0.000162 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 334.4558)

    Null deviance: 28403  on 70  degrees of freedom
Residual deviance: 23077  on 69  degrees of freedom
AIC: 618.15

Number of Fisher Scoring iterations: 2
```

- Model 2 (GDP only)

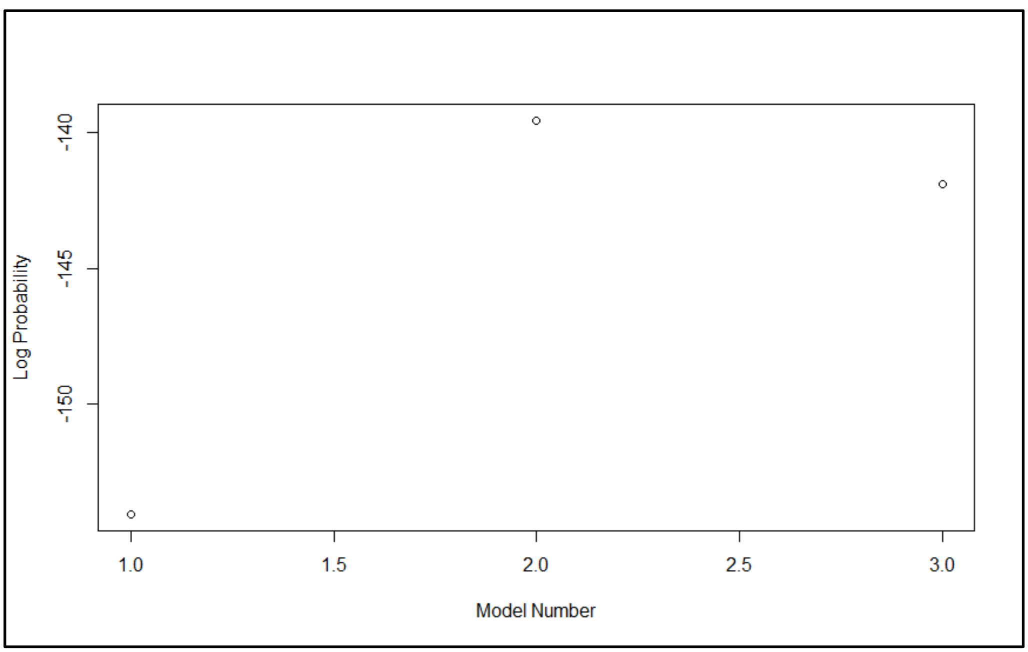
```
Call:
glm(formula = Medal2012 ~ GDP, data = medal_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-20.211   -6.025   -2.590    3.885   60.244

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.2359593  1.4788174   4.217 7.39e-05 ***
GDP         0.0078160  0.0006438  12.140  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 131.2601)

    Null deviance: 28402.8  on 70  degrees of freedom
Residual deviance:  9056.9  on 69  degrees of freedom
AIC: 551.74

Number of Fisher Scoring iterations: 2
```

- Model 3 (Population + GDP)

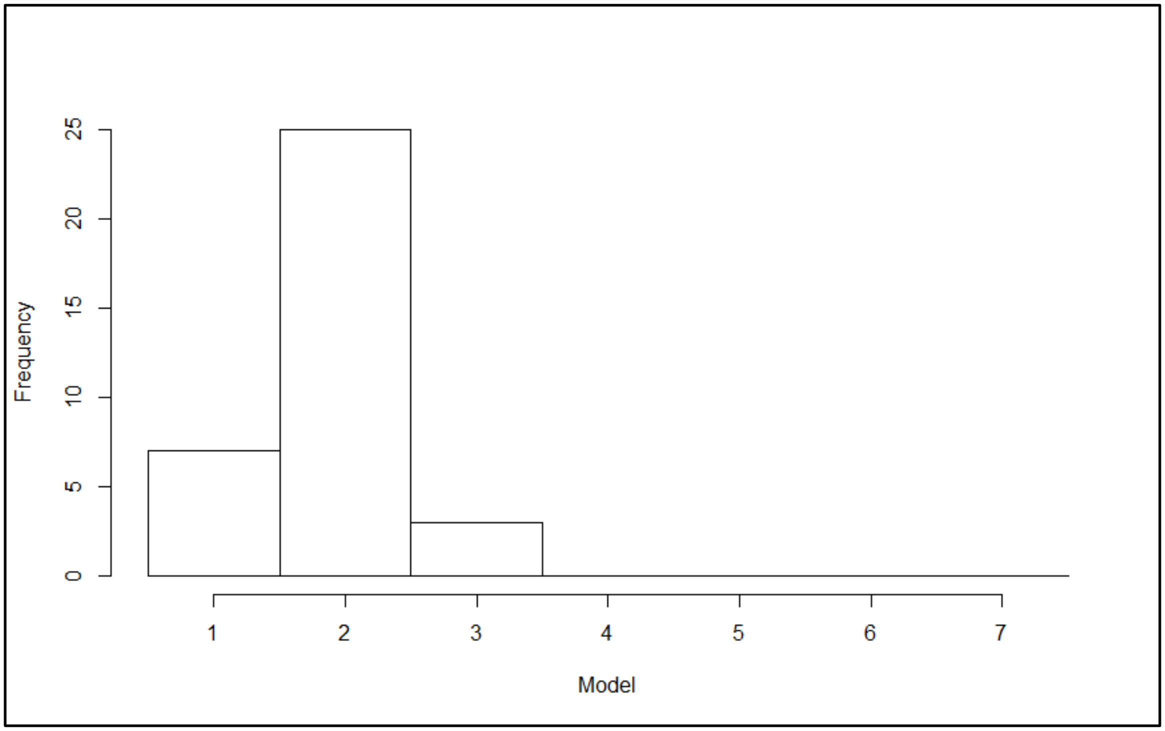
```
Call:
glm(formula = Medal2012 ~ GDP + Population, data = medal_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-20.568   -5.961   -2.462    3.932   60.121

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.076e+00  1.500e+00   4.051 0.000133 ***
GDP         7.564e-03  7.325e-04  10.326 1.45e-15 ***
Population  5.247e-09  7.193e-09   0.729 0.468225
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 132.1562)

    Null deviance: 28402.8  on 70  degrees of freedom
Residual deviance:  8986.6  on 68  degrees of freedom
AIC: 553.19

Number of Fisher Scoring iterations: 2
```

From 3 different models we can see if model 2 (GDP only) has the lowest AIC with 551,74. Followed by model 3 (AIC = 553,19) and lastly model 1 remain with the highest AIC (618,15). Thus, we choose model 2.

2.) To ensure the AIC result we use cross-validation method to check if model 2 is really feasible model. Here we are going to plot in 2 different way which is just the training set in scatter plot and histogram which contain the iteration of 35 times which is the real result we want to obtain.



Above we have reliable result through the log-likelihood probability. We got model 2 with the highest log probability. Nevertheless, another test of 35 iteration is still needed. To make it different we put the iteration result in histogram,



Similarly, to the AIC and the scatter graph, we have model 2 as the highest frequency. This is good sign since we can conclude that the model 2 is the preferable model compare to model 1 and 3. This could be because model 3 (GDP + population) might overfit with more variable and model 1 (population only) is weak to explain the medal prediction.

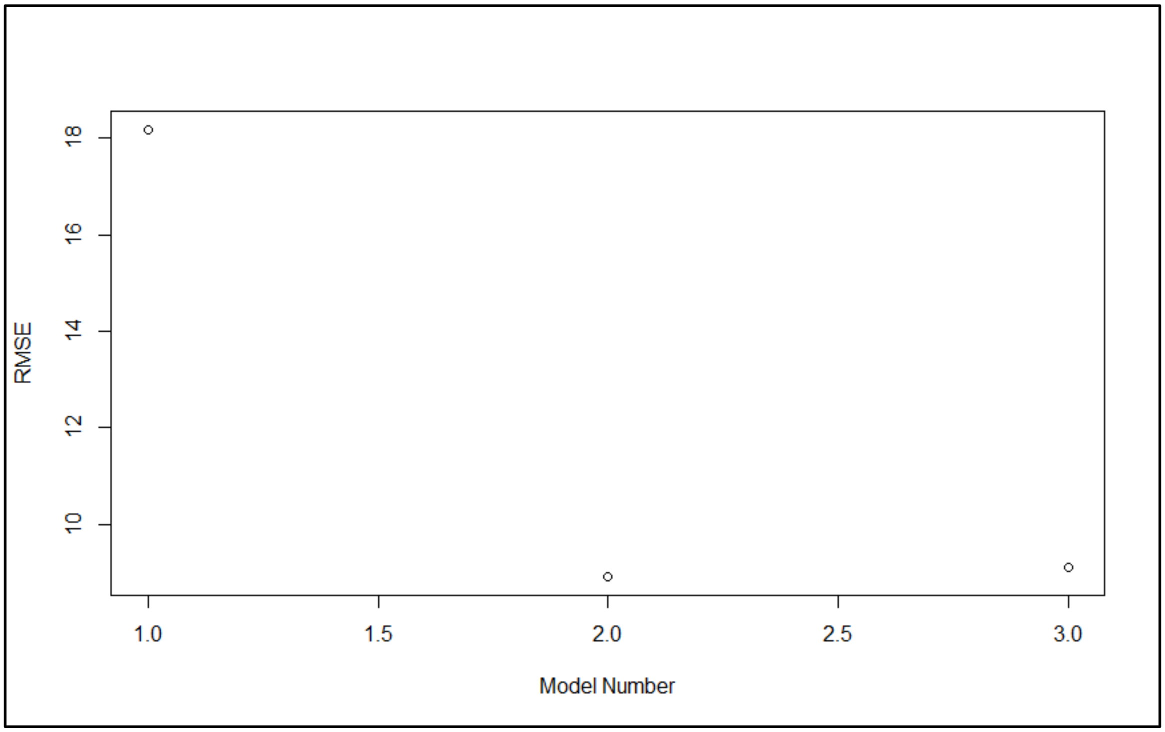
3.) Similar to part 1, we use the predict function to predict each model corresponding to each variable.

	Country	Population_model	GDP_model	PopGDP_model	Actual
1	Algeria	11.81632	7.710687	7.697937	2
2	Argentina	11.93780	9.721830	9.660081	4
3	Armenia	10.45411	6.316074	6.170773	4
4	Australia	11.24378	16.957674	16.572245	29
5	Azerbaijan	10.68936	6.731495	6.603459	18
6	Bahamas	10.33675	6.296846	6.136872	2
7	Bahrain	10.37222	6.407130	6.248223	2
8	Belarus	10.70346	6.666935	6.542817	9
9	Belgium	10.76345	10.234092	10.002805	6
10	Brazil	18.06845	25.593526	25.819025	19
11	Bulgaria	10.61904	6.654195	6.519486	3
12	Canada	11.72256	19.804975	19.390152	22
13	China	64.57287	63.278115	68.348721	70
14	Colombia	12.19380	8.828145	8.828562	8
15	Croatia	10.49527	6.735013	6.581570	10
16	Cuba	10.77513	6.711252	6.595043	11
17	Czech Republic	10.74545	7.918125	7.759147	10
18	Denmark	10.54720	8.836196	8.621790	15
19	Dominican Republic	10.70014	6.670609	6.545939	1
20	Egypt	13.62126	8.029972	8.242126	3
21	Estonia	10.37557	6.409319	6.250779	1
22	Ethiopia	13.71765	6.483806	6.758360	8
23	Finland	10.54022	8.315570	8.117037	1
24	France	12.95379	27.910040	27.394391	42
25	Georgia	10.50246	6.348276	6.208237	7
26	Germany	13.61739	34.143557	33.513444	42
27	Great Britain	12.82945	25.241335	24.795509	67
28	Greece	10.75687	8.570841	8.392310	6
29	Hungary	10.72362	7.330438	7.187558	15
30	India	60.31055	20.679823	26.568161	2
31	Indonesia	19.89101	12.854806	13.728427	3
32	Iran	13.39505	6.912280	7.130986	8
33	Ireland	10.50725	7.934226	7.743689	2
34	Italy	12.76964	23.390187	22.996238	28
35	Jamaica	10.43145	6.353747	6.204280	11
36	Japan	15.46227	52.093769	51.125438	41
37	Kazakhstan	10.99565	7.691304	7.572239	17
38	Kenya	11.87712	6.498734	6.532974	13
39	Lithuania	10.45106	6.569938	6.416057	4
40	Malaysia	11.46337	8.414052	8.332637	5
41	Mexico	14.84568	15.265973	15.404428	5
42	Moldova	10.46583	6.290671	6.147716	1
43	Mongolia	10.43270	6.302864	6.155200	2
44	Morocco	11.63218	7.019282	7.004824	1
45	Netherlands	10.99620	12.772191	12.489418	19
46	New Zealand	10.50098	7.257358	7.087823	18
47	North Korea	11.29096	6.407912	6.368698	7
48	Norway	10.52406	10.032986	9.776986	4
49	Poland	11.87273	10.257306	10.169817	11
50	Portugal	10.74776	8.092422	7.928127	1
51	Romania	11.08926	7.641203	7.535952	5
52	Russian Federation	16.08260	20.756342	20.878996	56
53	Serbia	10.60922	6.587993	6.454139	8
54	Singapore	10.53123	8.109461	7.916400	1
55	Slovakia	10.54176	6.986220	6.830738	4
56	Slovenia	10.40535	6.623165	6.461612	4
57	South Africa	12.35936	9.426775	9.429469	10
58	South Korea	12.27856	14.960601	14.774385	21
59	Spain	12.18258	17.888172	17.595080	17
60	Sweden	10.70464	10.441999	10.196346	11
61	Switzerland	10.63939	11.204218	10.925493	7
62	Taiwan	11.25801	9.878228	9.722857	3
63	Tajikistan	10.62916	6.286920	6.165369	1
64	Thailand	12.95900	8.937569	9.034171	6
65	Trinidad and Tobago	10.37556	6.411664	6.253046	1
66	Tunisia	10.75228	6.594402	6.478984	3
67	Turkey	13.33124	12.278453	12.315867	8
68	Ukraine	12.16036	7.527558	7.565541	11
69	United States	22.94067	124.211089	121.892569	121
70	Uzbekistan	11.49515	6.590494	6.572002	13
71	Venezuela	11.41569	8.709576	8.612422	3

To check which model predicts the best we use root mean squared error method (RMSE), by using rmse() function we obtained such as,

	model1_Population	model2_GDP	model3_GDPPopulation
1	18.18077	8.908145	9.112593

Finally, we plot it on the scatter graph which give us,



The x-axis refers to which model we use and y-axis is the RMSE value. The smallest RMSE value the better fit model with the real data. This means again model 2 (8,9) has the best prediction followed by model 3 (9,11) and model 1 (18,18) respectively.