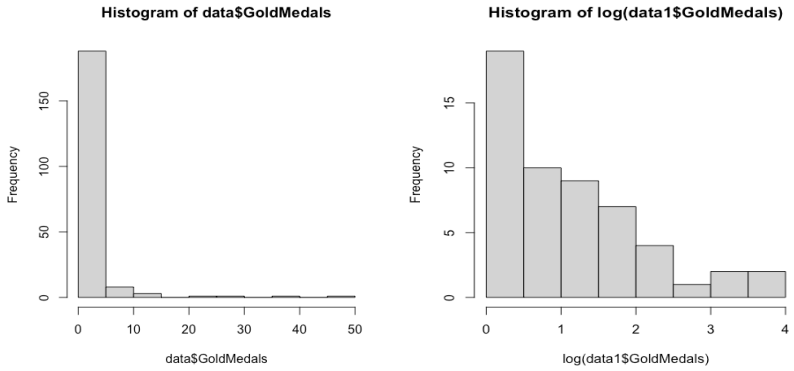
Multiple and Logistic Regression

Author By: Arivarasan Ramasamy

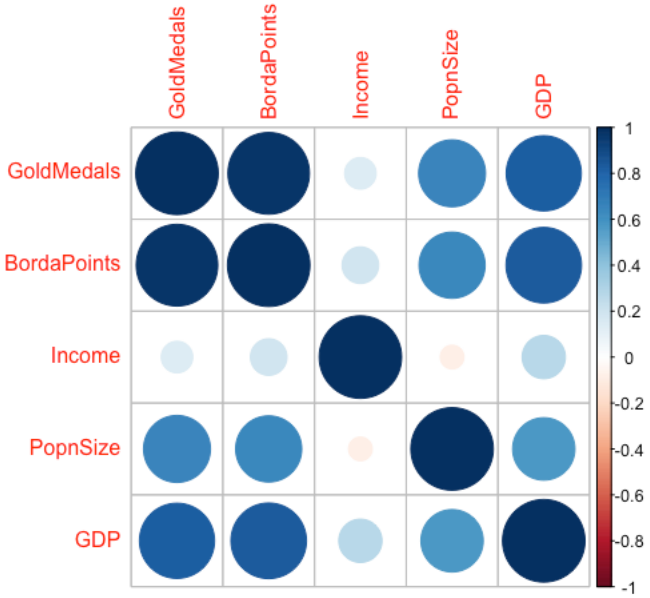
Data investigation and analysis :

On investigating the model on number of gold medals, there were many numbers of countries does not own any medals. Histogram view of total gold medals column shows a right skewed graph as shown below



First histogram shows that there is a outlier in the model data. To make the comparison more reliable, we are transforming the data set by taking a subset of the data frame with only counties those own gold medals. And taking the log of the total gold medals column to make it more symmetrical. Adding it as a new column to the data frame table for compression process. Now plotting histogram for the log gold medals column we get better hist plot which is represented in the right side.

On comparing the correlations between several important factors with the total gold medals of the country shows the below correlation plot



The correlation plot above shows that there is a strong correlation between total gold medals own by a country is more dependent on Borda points, GDP, Population size and other variables. The size and the color of the plot show the correlation percent of the variables.

This show that GDP of a country plays a important factor in athletes winning more number of gold medals in Olympics. At the same time Population size also influence but not as strong as GDP. This shows that countries with more population and less GDP wins less medals than countries with high GDP and less population with more infrastructure.

Investigate the interactions between the G20 variable and other significant predictors

By fitting liner models with the G20 variables with several other independent variables like Income, GDP and population size and the result is displayed below

	Dependent variable:					
	G20					
	(1)	(2)	(3)	(4)	(5)	(6)
GDP	0.758***	1.065***		1.117***		
	(0.276)	(0.271)		(0.221)		
Income	0.070***		0.096***		0.089***	
	(0.025)		(0.024)		(0.026)	
PopnSize	0.416	0.118	1.006***			0.911***
	(0.347)	(0.353)	(0.289)			(0.328)
Constant	0.052	0.191***	0.038	0.193***	0.118	0.256***
	(0.074)	(0.059)	(0.078)	(0.058)	(0.082)	(0.064)
Observations	54	54	54	54	54	54
R^2	0.425	0.332	0.338	0.330	0.182	0.129
Adjusted R^2	0.390	0.306	0.312	0.317	0.166	0.113
Residual Std. Error	0.366 (df = 50)	0.391 (df = 51)	0.389 (df = 51)	0.387 (df = 52)	0.428 (df = 52)	0.442 (df = 52)
F Statistic	12.311*** (df = 3; 50)	12.658*** (df = 2; 51)	13.031*** (df = 2; 51)	25.643*** (df = 1; 52)	11.534*** (df = 1; 52)	7.723*** (df = 1; 52)
Note:						*p<0.1; **p<0.05; ***p<0.01

From the above html file the adjusted R square value is showing a significant range for the combined Income, GDP and Population size but the value is less for independent variables. This show that all factors like GDP and population and income combinedly improves the Position of the country in the top 20 Gold medal list.

The interactions between the significant predictors can be calculated from the linear regression and combining them in a stargazer function and the results are interpreted below

	Dependent variable:		
	G20		
	(1)	(2)	(3)
GDP	5.580***	9.953***	2.070***
	(1.861)	(1.732)	(0.264)
Income	0.059**	0.101***	0.066***
	(0.023)	(0.020)	(0.017)
PopnSize	1.547	-4.081***	6.022***
	(2.209)	(0.885)	(0.807)
GDP:Income	-0.543	-1.767***	
	(0.387)	(0.330)	
GDP:PopnSize	-4.484*		-8.723***
	(2.396)		(1.197)
Income:PopnSize			
GDP:Income:PopnSize	-0.901		
	(0.668)		
Constant	-0.109	-0.048	-0.166***
	(0.068)	(0.062)	(0.060)
Observations	54	54	54
R^2	0.749	0.637	0.724
Adjusted R^2	0.717	0.607	0.701
Residual Std. Error	0.249 (df = 47)	0.294 (df = 49)	0.256 (df = 49)
F Statistic	23.388*** (df = 6; 47)	21.503*** (df = 4; 49)	32.136*** (df = 4; 49)
Note:			*p<0.1; **p<0.05; ***p<0.01

From the above html output, we can see that the R square and adjusted R square value got increased this shows a positive correlation between the countries having high GDP, Income and population has more probability of getting into the top 20 gold medal list. The negative symbol in the values shows that there is a negative correlation i.e it affects the success of the winning a gold is negative impact.

Building a model to predict the probability of being at the top10 by total medal

Calculated the top 10 countries in the total medal list and added them with a dummy variable as a new column in the data set. Using the logistic regression, we predicted the probability of being in top 10 total medal list. The html file generated from the logistic regression considering the interactions between significant variables

	Dependent variable:					
	top10medals					
	(1)	(2)	(3)	(4)	(5)	(6)
GDP	41.735***	34.283***		33.717***		
	(12.310)	(9.278)		(8.468)		
Income	0.065	0.247	0.225***		0.189**	
	(0.323)	(0.209)	(0.083)		(0.080)	
PopnSize	-14.832*		4.001***			3.576**
	(8.953)		(1.441)			(1.402)
Constant	-6.501***	-7.151***	-3.793***	-6.275***	-3.394***	-3.219***
	(1.731)	(1.888)	(0.481)	(1.359)	(0.415)	(0.362)
Observations	203	203	203	203	203	203
Log Likelihood	-7.814	-8.588	-32.859	-9.144	-37.590	-35.847
Akaike Inf. Crit.	23.627	23.177	71.717	22.288	79.181	75.694
Note:						*p<0.1; **p<0.05; ***p<0.01

html

From the above logistic regression output we can see that GDP of Country contributes to a max of 41% and it's the most important factor for being at the top 10 total medals list.
Whereas the population shows a negative coefficient. This shows that countries with more population are not in the top 10 list may be due to less infrastructure facilities.
Income of a country also has a positive impact, but this does not contribute more to the top 10 list medals. These shows that some countries have more income but the interest towards the athlete training and winning a medal is less. They might have invested in some other development projects.
Some countries have more population and income but less GDP due to the income is distributed to the huge population and these countries are not in the list of top 10 medal list.