A Classroom Investigation of the Effect of Population Size and Income on Success in the London 2012 Olympics

Nancy Carter
Nathan Felton
Neil Schwertman
California State University, Chico

Journal of Statistics Education Volume 22, Number 2 (2014),
www.amstat.org/publications/jse/v22n2/carter.pdf

Key Words: least squares; linear models; regression

Abstract

Engaging students in active learning can enhance their understanding and appreciation of a subject such as statistics. Classroom activities and projects help to engage students and further promote the learning process. In this paper, an activity investigating the influence of population size and wealth on the medal counts from the 2012 London Olympics is suggested, and the relevant data is provided.

1. Introduction

Group activities or projects engage students in active learning and can be useful in establishing enthusiasm and understanding of a subject such as statistics. Projects investigating timely, real world events of wide interest stimulate student appreciation of the subject and promote a deeper understanding. Clearly there are numerous data sets available, but it is sometimes difficult to find ones that are current and of interest to a substantial number of students. Fortunately, every two years, sports enthusiasts are treated to a special event, the Olympics. The summer and winter Olympics alternate biannually, with the summer Olympics in years divisible by four and the winter games in years divisible by two but not four. One interesting and timely data set is the medal counts from the recent 2012 London Olympics which can be found on the internet at, for example, http://espn.go.com/olympics/summer2012/medals and the list of nations at http://www.london2012.com/countries/

The outcomes of the Olympics are a source of great national pride and have been dominated by the larger and wealthier countries. This is not surprising since larger population nations have a larger pool from which to draw their athletes and richer nations can devote more resources to training and preparation. The project suggested here is for at least second year undergraduate statistics majors up to introductory linear model students, and the focus is the relationship between a country's Olympic success and its population size and wealth. In our case, this project was used to inspire a particularly motivated senior statistics major. While the proposed project is for more advanced students, the data could be used for beginning students to illustrate scatterplots or, for example, histograms of distribution of medal counts, per capita income or population of the participating nations.

Finding the population size and wealth of all 203 nations participating in the 2012 Olympics can be difficult and tedious. Since our primary purpose is to recommend this project to enhance student learning, we facilitate this by providing access to not only the medal counts for all 203 nations but also their population size and wealth at http://www.csuchico.edu/math/theses-projects.shtml (under Felton, Nathan). Since the population size and wealth of each nation is dynamic and continuously changing, Wikipedia was used as the source of the population size and wealth for consistency and because it included all 203 nations. While not perfect, this should provide at least some measure of the relative population sizes and wealth. Using these factors, students can creatively investigate different models, transformations and the various criteria for measuring the adequacy of the models. The discussion and comparison of the various models should provide a more thorough appreciation of the flexibility and power of least squares regression.

Olympic success is usually measured in one of three ways: first by the total number of Gold medals, GoldMedals=Y_1; second, by the total number of medals, TotalMedals=Y_2; and third, by the total number of points using the Borda method of assigning 3 points for first place (gold), 2 points for second place (silver) and 1 point for third place (bronze), BordaPoints=Y_3. As expected, these dependent variables are highly correlated: .963, .982, .996 for Y_1 and Y_2, Y_1 and Y_3, and Y_2 and Y_3 respectively.

We suggest very simple models as starting points in quantifying the influence on population size and a nation's wealth on success in the 2012 Olympics. These models are not intended as definitive models but rather to illustrate that simple statistical methodology can accommodate these factors in predicting a nation's Olympic success. Note that all statistical analyses in this paper were done using the Minitab statistical program (version 16). The data from this paper can be found at www.amstat.org/publicatons/jse/v22n2/carter/Olympics.csv, and the documentation file can be found at http://www.amstat.org/publciations/jse/v22n2/carter/Olympics_documentation.docx.

> *Helpful Hint: Ask the students if they believe that a nation's wealth and population size are important factors in determining Olympic performance. Why? How do they explain, for example, the small, relatively poor Jamaica winning 12 medals and coming in 20th among the 203 participants? Is their success inconsistent with the majority of the data? What are such observations called in statistics?*

2. The Effect of Population Size

It is intuitive that the population of a country should have a substantial impact on Olympic success. Large populations provide a larger pool of potential Olympic athletes. The impact of population size may be greater on nations with small populations where the number of Olympic quality participants is already quite limited.

3. The Effect of Wealth

The precise amount of total resources each nation devotes to their Olympic program is nearly impossible to determine. Furthermore, the countries may distribute the resources to the Olympics in a variety of ways, such as free or subsidized housing or expanded use of government services. Consequently, we considered both the nominal and the PPP (purchasing power parity) income per capita as a surrogate of a general measure of a nation's wealth in these analyses. To determine which measure of wealth to use in the models, the correlations with the response variables GoldMedals(Y_1), TotalMedals(Y_2) and BordaPoints(Y_3) were calculated using both income variables. For Y_1 the correlations were .13 and .13, for Y_2 they were .17 and .16 and for Y_3 they were .16 and .15, using nominal and PPP respectively. Since the correlations were identical or nearly so with a slight advantage to the nominal per capita income, we suggest, at least initially, that the students use the nominal value, Income(X_1), in their models.

> *Helpful Hint: Have the students discuss how they could determine the influence of wealth and population size on Olympic performance. What graphical and numerical methods might be useful?*

4. Models

> *Helpful Hint: Ask students the following questions: What is the purpose of the model? What criterion should be used to determine the adequacy of the model and to compare models?*

The process of creating models requires an intuitive assessment of factors that may affect the dependent variable. There clearly are numerous variables which can influence a nation's Olympic success. Besides population size and wealth, other factors such as experience in world class sporting competitions, a tradition of excellence in particular events (for example, China in women's diving) and the environment can be influencing factors on success in sports. In the Winter Olympics, countries such as Switzerland, Austria and other countries with prime winter sports venues, have a clear environmental advantage over tropical countries. Similarly, countries with access to high elevations for training of runners, especially long distance runners, may have an advantage. It should be noted that some of these variables could be confounded in predicting Olympic success. For example, the actual distribution of wealth in a country could be confounded with GDP. Also, monetary incentives given to winning athletes in some countries can be affected by the GDP or the distribution of wealth in the country. Furthermore, environmental factors such as elevation and population size could be confounded since population size tends to be concentrated at lower elevations. Some of these variables are exceedingly difficult to quantify but for those that are quantifiable, multiple regression could be used to separate the confounding effects. Because of the difficulty in quantifying many of these factors, we investigated only the two most obvious influential variables: population size and per capita wealth. To better understand the patterns of the independent variables population and wealth, we suggest as a first step creating scatterplots of each independent variable and the success variables. These plots may suggest a linear or higher order relationship or an appropriate transformation of the independent variable to help explain the response variable Y. In the model building process, the simplest linear model is usually a good place to start. The basic independent variables are Income=X_1 (the nominal income per capita in \$10,000) and PopnSize=X_2 (size of population in 1,000,000,000). The product of a country's income per capita and population size is called the Gross Domestic Product (GDP) and is considered a potential variable in the model denoted by X_3. The first model considered, however, used just the two fundamental variables X_1 and X_2 and was $E(Y) = \beta_0 + \beta_1X_1 + \beta_2X_2$ where Y is Olympic success. For this model with Y_1, Y_2 and Y_3, the R-squares are: .2615, .2625 and .2682 respectively. The next model included X_3, that is, $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3X_3$ and when used with Y_1, Y_2 and Y_3 and had R-squares of .7148, .7069 and .7230 respectively. Clearly this was a vast improvement in explaining total variability.

A heuristic argument can be made for using a natural log transformation. It seems intuitive that an increase of one hundred thousand in the population of China, which has 1.3 billion people, would have only a very marginal influence on enhancing the quality of the pool from which China can draw their Olympic athletes. Conversely, the same increase would likely greatly enhance the quality of the pool of potential athletes for Grenada by nearly doubling its population of 110,000. It appears that any change in population size should be measured relative to the initial population X_2, that is, $\Delta X_2/ X_2$ where ΔX_2 is the change in population size X_2. If Y is the nation's Olympic success, then ΔY is the change in success and if this change is assumed to be proportional to the change in population size ΔX_2 relative to the initial population X_2, then $\Delta Y = k\, \Delta X_2/ X_2$ (k is the proportionality constant). The limit of this equation gives $dY = k\, dX_2/ X_2$. Solving this differential equation produces $Y = k \ln(X_2) + c$. This suggests that the $\ln(X_2)$, the natural log of the population size, may be an appropriate transformation of population size and furthermore, is consistent with the law of diminishing return.

A similar argument can be made for income per capita. An increase in per capita income of \$1000 for a wealthy nation such as the United States may have only a minor effect on success whereas the change for a poor African nation could have a dramatic impact on success.

A positive difference between the observed and predicted success gives a measure of the degree of a country's superior performance, while a negative difference indicates the degree of underperformance. The numerous variables and their higher order products and transformations provide the students with many independent variables. More advanced students may wish to investigate the appropriateness and the adequacy of the fit in finding the "best" model. The students can creatively use these variables to find which combinations provide the "best" linear model for describing Olympic success.

> *Helpful Hint: A discussion of what is meant by "best" and the criteria such as R-square, adjusted R-square, significance level of the estimated model coefficients, residual*

differences or mean square error (MSE = variance + squared bias) may stimulate discussion.

We note that the adjusted R-square accounts for the degrees of freedom in the model by introducing a penalty term. For prediction purposes, PRESS could be considered as appropriate criteria for determining the best model.

While all three dependent variables, Y_1, Y_2 and Y_3, are commonly used measures of Olympic success, the choice of which one to use is a matter of personal preference. For brevity, we illustrate the model building process for the most popular measure of success, TotalMedals(Y_2). Included are scatterplots (along with a Lowess smoother) for Income(X_1), PopnSize(X_2) and GDP(X_3) versus Y_2 (see Figures 1, 2, and 3). To measure the relative slope and linear relationship of these independent variables with Y_2, the simple correlations were found to be .167, .474 and .838 for X_1, X_2 and X_3, respectively. That the correlations are all positive shows that all three variables have a positive influence on the Olympic success as measured by total medals. The large correlation between Y_2 and X_3 indicates a strong positive slope and that GDP(X_3) may be a major component in the model.

Figure 1.

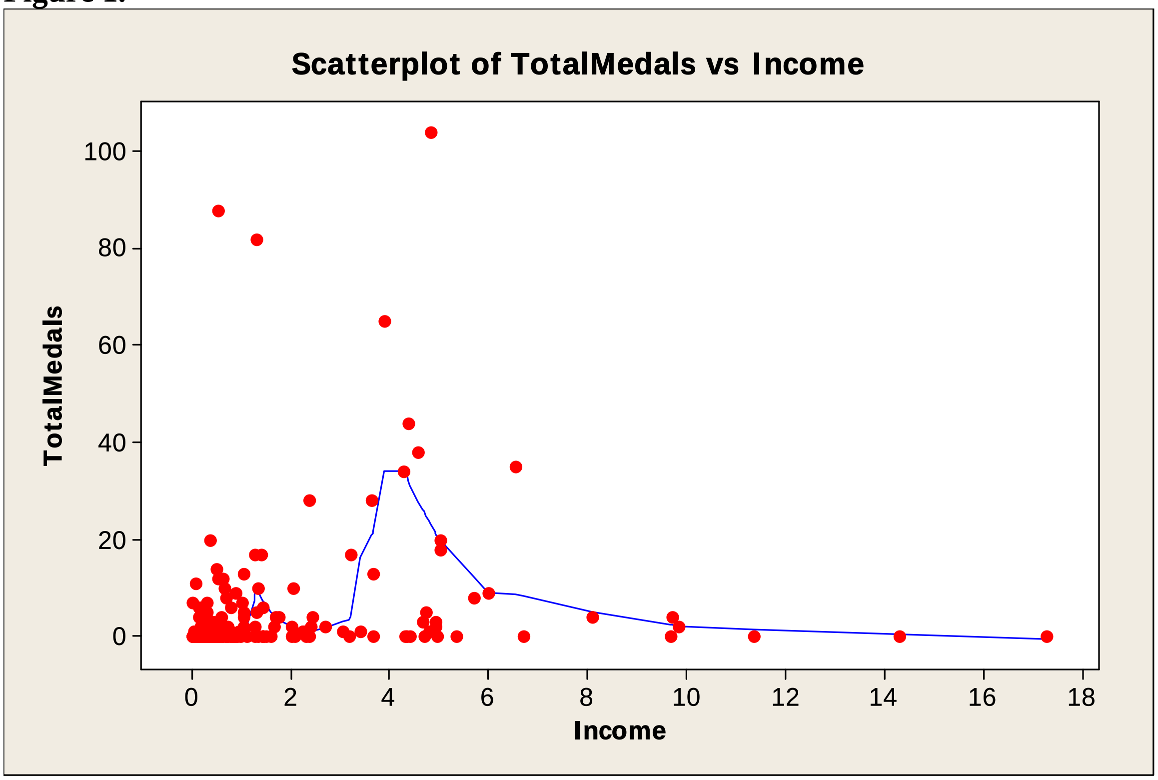


Figure 2.

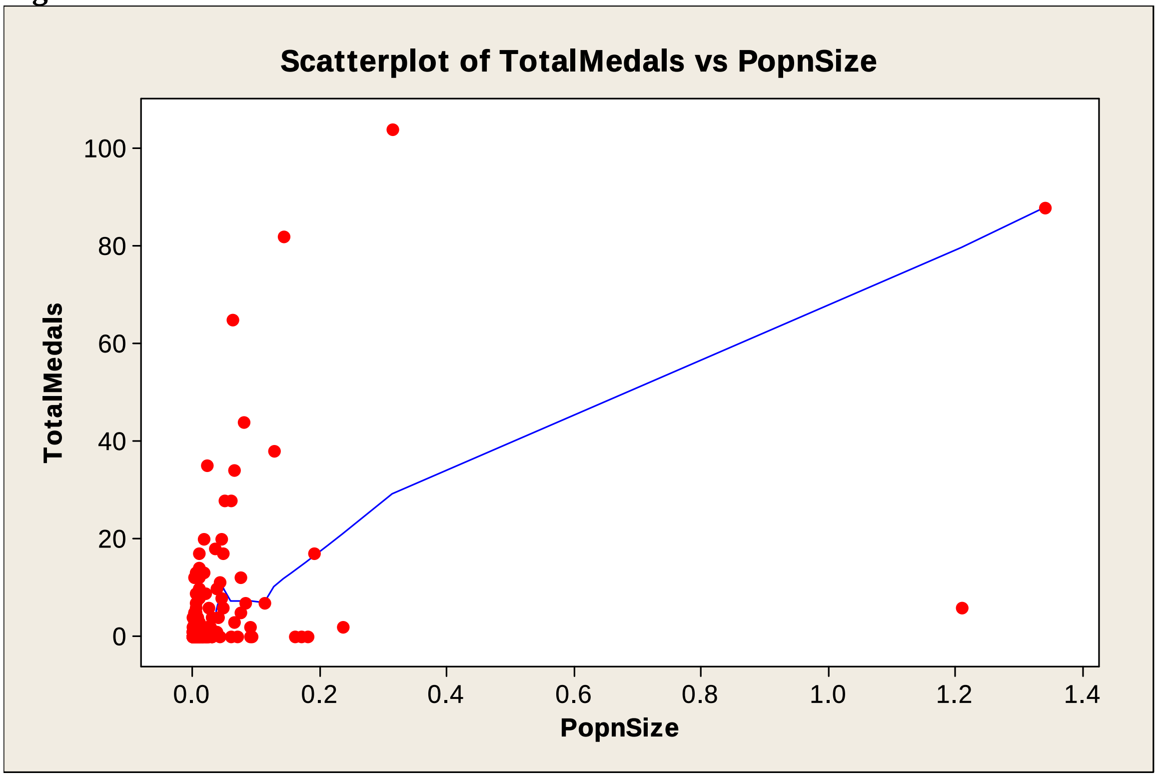
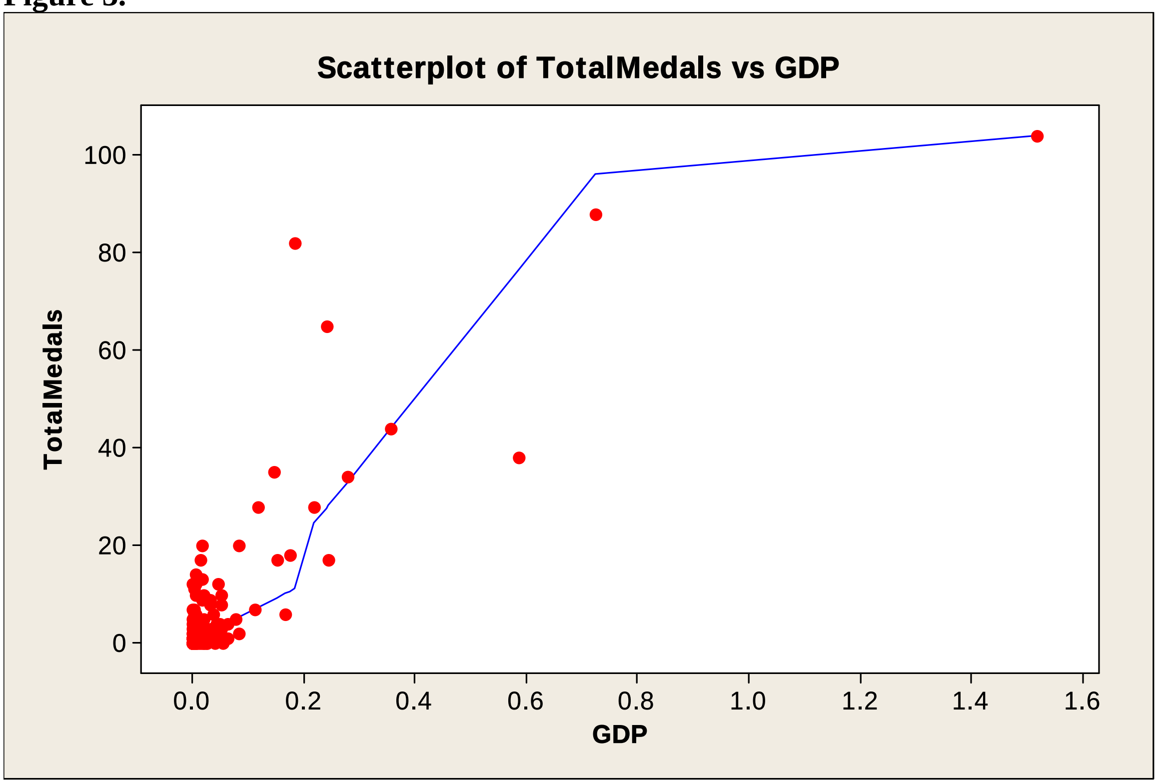


Figure 3.



We investigated several simple models using X_1, X_2 and X_3, along with their interactions, higher order terms and natural logs. Descriptive statistics for the principal variables are included in Table 1. All principle variables are extremely skewed to the right as indicated by the skewness parameter in the table.

Table 1. *Descriptive Statistics for Principle Variables*

Variable	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
GoldMedals	1.488	5.225	0.000	0.000	0.000	1.000	46.000
TotalMedals	4.739	13.457	0.000	0.000	0.000	3.000	104.000
BordaPoints	9.21	27.77	0.00	0.00	0.00	5.00	225.00
Income	1.573	2.493	0.022	0.150	0.564	1.764	17.268
PopnSize	0.03402	0.13110	0.00001	0.00123	0.00650	0.02270	1.34009
GDP	0.03460	0.13249	0.00000	0.00046	0.00215	0.01791	1.51947

Variable	Range	Skewness
GoldMedals	46.000	6.07
TotalMedals	104.000	4.96
BordaPoints	225.00	5.30
Income	17.246	3.08
PopnSize	1.34008	8.65
GDP	1.51946	8.

Regression summary results for several of the models investigated are included in Appendix A. The model $E(Y_2) = \beta_0 + \beta_1 X_3 + \beta_2 X_3^2$ had R-square = .7655, highly significant estimated coefficients, MSE = 42.9 and PRESS = 11557.0. We selected this as our "best" model because of its simplicity and because it had the best PRESS and MSE values and the R-square was only slightly below the highest R-square (.7672) of all models considered. This particular model is especially appealing in its simplicity and the interpretability of its terms. We include the residual plot (along with smoother) for this model versus GDP (Figure 4) along with two diagnostic plots of fit vs. residuals (Figure 5, including smoother) and normal probability plot (Figure 6) which may be useful in better understanding the model. It should be noted that all principle variables contained numerous outliers. This is particularly important in the independent variables. Using fences to identify outliers, per capita income (X_1) had 25 mild and 7 extreme outliers. Population size (X_2) had 2 extreme outliers and GDP (X_3) had 3 mild and two extreme outliers. These outliers have large standard deviations and leverage and will have a powerful effect on the models as can be seen in Figures 4, 5 and 6. We also examined the residual plot using the natural log transformed variables. The natural log transformations of the independent variables were ineffective in improving the models by all criteria used to find the "best" model.

> *Helpful Hint: Some students may observe that using the natural log transformations in the linear model produces essentially a multiplicative type model.*

Both the residual plots (residuals verses Total medals and fitted values) show there is a substantial increase in variability as the total medals increase. This is to be expected since most nations had no or very few medals and hence little or no variation. This is typical since the coefficient of variation is usually relatively constant and suggests that the standard deviation or variance generally increase with the mean.

As expected from such skewed variables, the normal probability plots show the residuals deviated from normality with outliers at both extremes. Using the natural log transformation did reduce the non-normality of the plots. The purpose of this activity, however, was to determine how much wealth and population size predicted the 2012 Olympic success rather than statistical inference, and hence the non-normality was a very minor concern. The natural log transformation greatly reduced the prediction quality of the model as measured by PRESS (see Appendix A).

Similarly, the log transformation of the outcome variables (log of 1 plus the outcome, that is, log(Y+1)) would vastly reduce the increasing variability of these counts with variance linked to the mean. Again, the purpose of this paper was predictive. However, if this exercise is to be extended to statistical inference, it could be useful in making the variances more consistent as required. While there are numerous models for counts, that is beyond the scope of this paper.

> *Helpful Hint: Ask the students to discuss what would be the ramifications of reducing the data to using only the results of nations that received medals.*

Most countries received no medals, thus creating an extremely skewed distribution. To overcome this one might consider including in the analysis only those countries which had some success in the 2012 Olympics. This will of course bias the overall data but perhaps give a more interpretative analysis of the data not influenced the preponderance of zeroes. To examine the effect of the abundance of zeroes, a comparison of analyses with and without the zero responses may be very informative for illustrating the effect of highly skewed data. The analysis of the success for nations that had received at least one medal produced the same "best" model as the full data set. With nations that earned no medals removed, R-square decreased to.7371from .7655 and MSE increased to 97.0 from 42.9. PRESS is not valid for comparison because of the vastly different number of observations. For gold medals at least, it appears that removal of the zeroes did not improve the analysis of the data.

Figure 4.

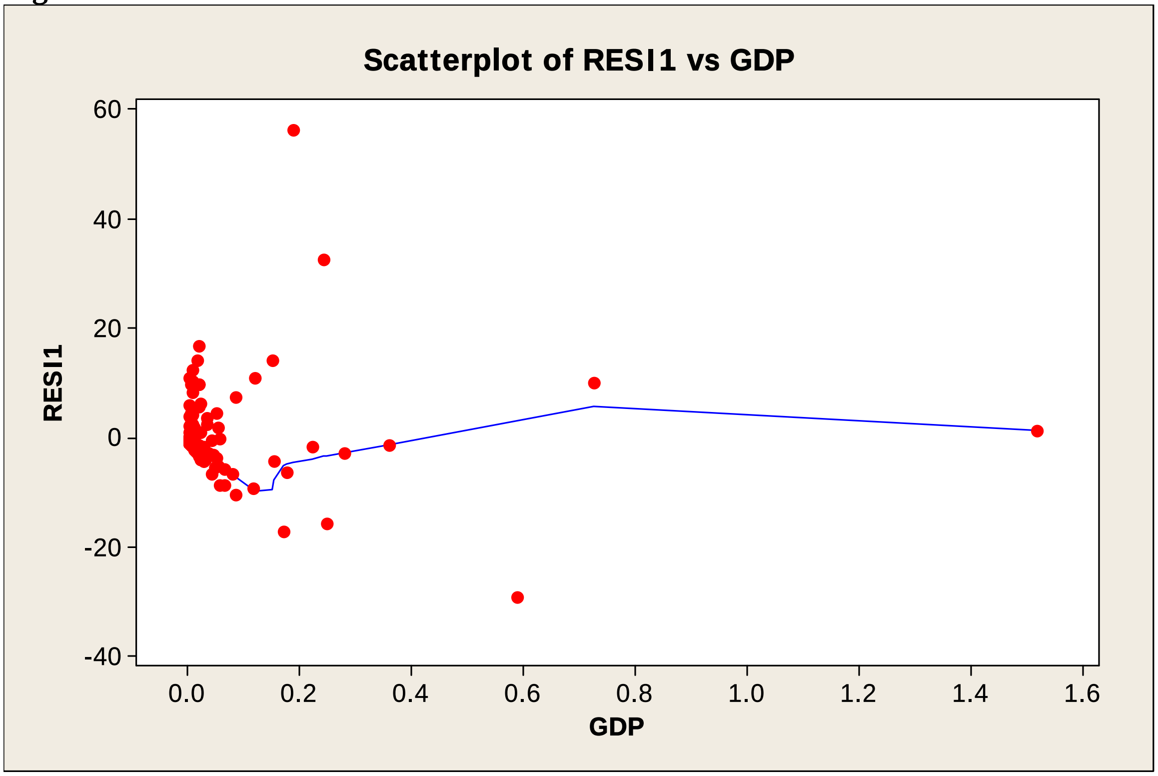


Figure 5.

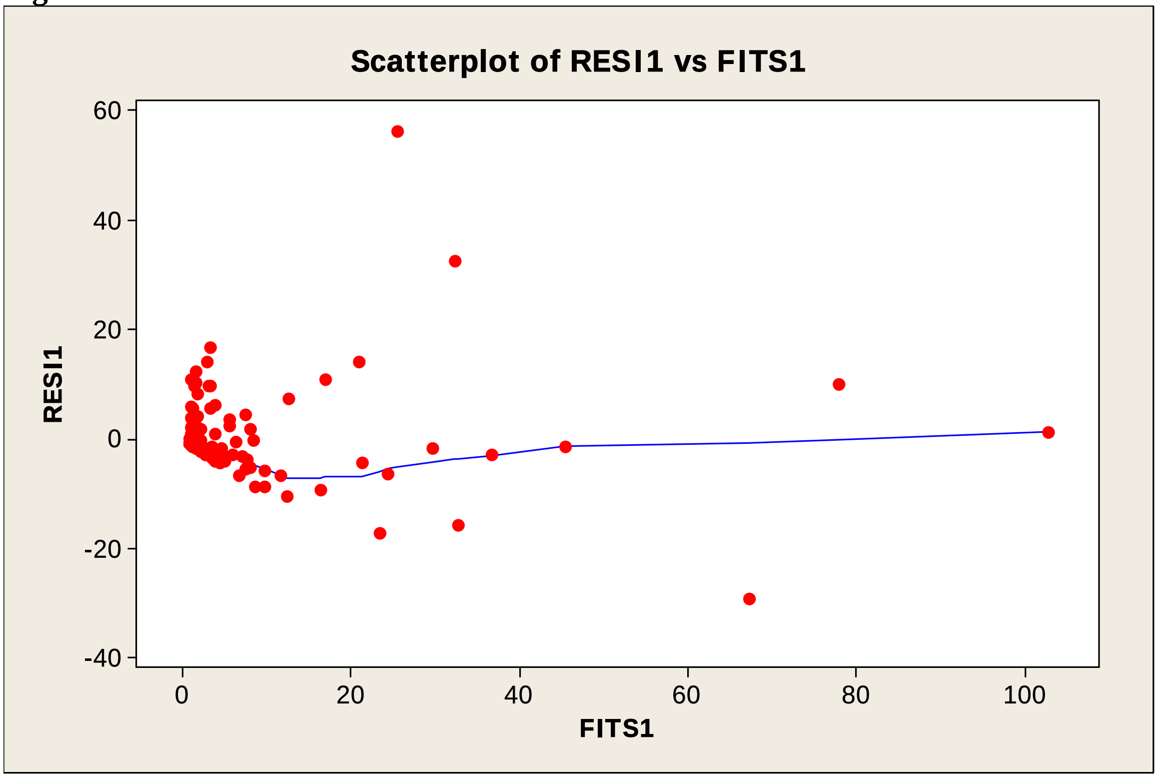
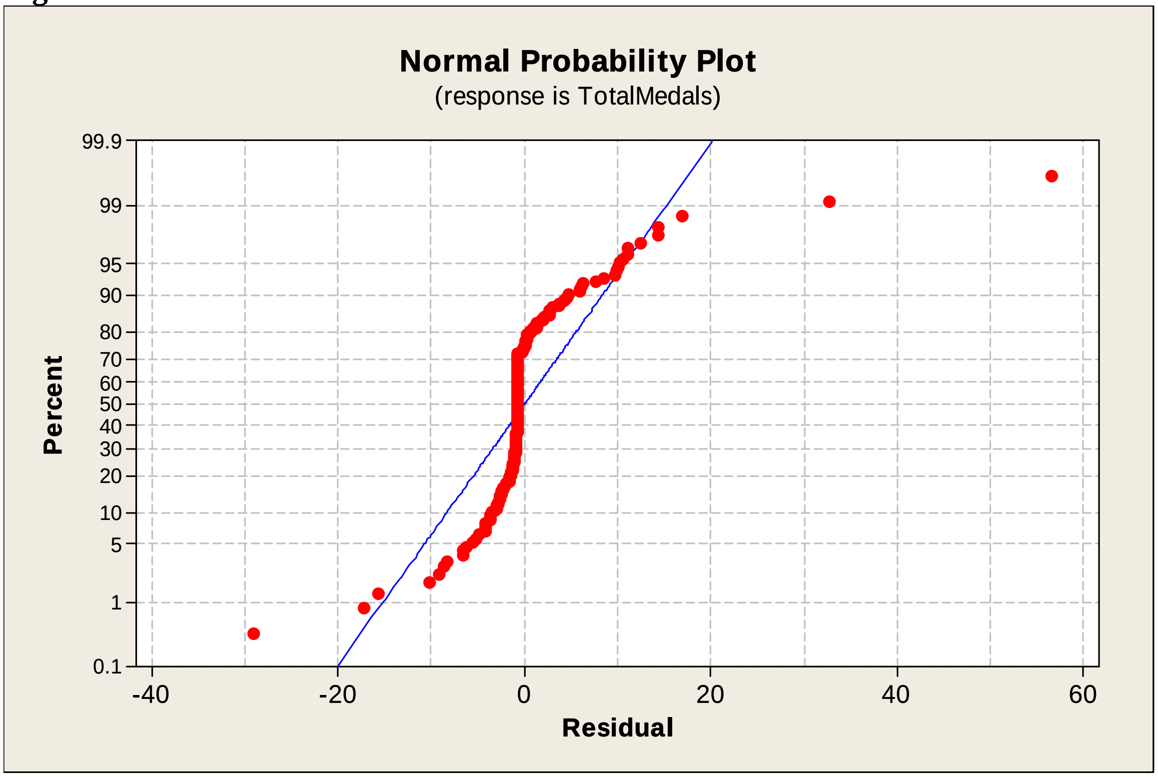


Figure 6.



5. Conclusions

While the focus of this paper is to enhance student learning through an activity or project, it is also important to interpret what information the model does provide about success in the 2012 Olympics. While the analysis clearly shows that population size and wealth were major components in success in the 2012 Olympics, it is their product, i.e. the GDP, which is the most influential component in the model. The GDP and its square explain over 75% of the total variation in the response variable.

Perhaps a country's performance in the Olympics should not be judged strictly on how many gold medals, total medals or Borda points it achieves, but rather on its performance compared to its predicted outcomes. This would allow a more balanced "level playing field" for comparing the success of even the smallest and poorest countries with the larger and wealthier ones. Tables 2 and 3 are quite extensive for all 203 nations and therefore, for spatial reasons, only the first 20 entries are included in this paper. The complete tables are available at http://www.csuchico.edu/math/theses-projects.shtml (under Felton, Nathan). Also included on the website are the Excel spreadsheet of the data, the Minitab regression results, and the scatterplots for GoldMedals(Y_2) with Income(X_1), PopnSize(X_2), and GDP(X_3). As can be seen on the complete output at http://www.csuchico.edu/math/theses-projects.shtml , the countries with large standardized residuals and large leverage (most influential) were the larger richer nations and those with the most medals. Since a large majority of the countries did not win any medals or had limited success, it is to be expected that the most influential observations are countries with considerable success. Table 2 displays the number of gold, silver and bronze medals, the total number of medals and Borda points for the first 20 entries. Table 3 displays the first 20 entries for the predicted outcomes for the total number of medals, Y_2. The ranks were achieved by ranking the residuals, that is, the difference of the actual outcome minus the model predicted outcome.

Table 2. *Original Data Set*

Country	Y1 Gold Medals Won	Silver Medals Won	Bronze Medals Won	Y2 Total Medals Won	Y3 Borda Points
United States	46	29	29	104	225
China	38	27	23	88	191
Russia	24	26	32	82	156
Great Britain	29	17	19	65	140
Germany	11	19	14	44	85
Japan	7	14	17	38	66
Australia	7	16	12	35	65
France	11	11	12	34	67
South Korea	13	8	7	28	62
Italy	8	9	11	28	53
Netherlands	6	6	8	20	38
Ukraine	6	5	9	20	37
Canada	1	5	12	18	25
Hungary	8	4	5	17	37
Spain	3	10	4	17	33
Brazil	3	5	9	17	28
Cuba	5	3	6	14	27
Kazakhstan	7	1	5	13	28
New Zealand	6	2	5	13	27
Iran	4	5	3	12	25

Table 3. *Total Medals Residual Ranking*

Country	Total Medals Won	Residual	Residual Ranking
United States	104	1.2478	35
China	88	10.1093	10
Russia	82	56.531	1
Great Britain	65	32.7217	2
Germany	44	-1.2999	155
Japan	38	-29.1737	203
Australia	35	14.2141	5
France	34	-2.6322	175
South Korea	28	11.064	7
Italy	28	-1.6893	165
Ukraine	20	16.9087	3
Netherlands	20	7.5484	15
Canada	18	-6.2387	194
Hungary	17	14.2666	4
Spain	17	-4.3082	189
Brazil	17	-15.732	201
Cuba	14	12.4449	6
New Zealand	13	9.9628	11
Kazakhstan	13	9.7496	12
Jamaica	12	11.0366	8

While the United States did very well in total medals (actually in all three criteria for success) using the raw data, the predicted success from the model using GDP was less spectacular. The observed success for the United States was better than the prediction, but clearly Russia and Great Britain far outperformed their predicted success. Obviously, there are many other factors besides GDP that influence a nation's Olympic success. There may be a "home field advantage" for the host country. This may be due, in part, to the increased hype of hosting this premier athletic event. Furthermore, cultural differences could have a profound influence on Olympic outcomes since some nations give athletics and sports a higher priority than other nations. In some sports, such as swimming and track and field, top athletes can compete in multiple events which can further influence Olympic medal totals. There are many factors that are exceedingly difficult to evaluate and contribute to the approximately 24 percent of the total variation (R-square) unaccounted for by our model. Hence this exercise focused on the influence of the nonsubjective GDP and its square.

The modeling in this paper of Olympic success shows that statistical methods can be used to develop a fairer way to compare a nation's success relative to the other nations in the 2012 Olympics, in essence, "leveling the playing field." Students, by using their creativity, may be able to devise innovative ways of using statistical methods to refine these models to account for the disparity in highly influential variables. The model proposed in this paper is just a starting point for a more in depth exploration by students. Instructors may wish to use the other measures of success, Y_1 (the total number of gold medals) and Y_3 (total Borda points), in similar analyses to allow their students to innovatively create similar prediction models and to investigate various regression criteria for adequacy of the model.

Appendix A: Regression Output for Several Models Investigated

General Regression Analysis: TotalMedals versus Income, PopnSize

Regression Equation

TotalMedals = 1.38849 + 1.05281 Income + 49.8057 PopnSize

Coefficients

Term	Coef	SE Coef	T	P	95% CI
Constant	1.3885	0.99453	1.39612	0.164	(-0.5726, 3.3496)
Income	1.0528	0.32834	3.20644	0.002	(0.4054, 1.7003)
PopnSize	49.8057	6.24393	7.97666	0.000	(37.4933, 62.1181)

Summary of Model

S = 11.6146 R-Sq = 26.25% R-Sq(adj) = 25.51%
PRESS = 34945.8 R-Sq(pred) = 4.47%

Analysis of Variance

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	2	9601.3	9601.3	4800.66	35.5870	0.0000000
Income	1	1018.1	1386.9	1386.93	10.2813	0.0015647
PopnSize	1	8583.2	8583.2	8583.25	63.6271	0.0000000
Error	200	26979.8	26979.8	134.90		
Total	202	36581.2				

Fits and Diagnostics for Unusual Observations

Obs	TotalMedals	Fit	SE Fit	Residual	St Resid		
1	104	22.1231	2.25591	81.8769	7.18633	R	
2	88	68.7024	8.18299	19.2976	2.34125	R	X
3	82	9.8832	1.06266	72.1168	6.23530	R	
4	65	8.5844	1.13564	56.4156	4.88068	R	
5	44	10.0686	1.27732	33.9314	2.93927	R	
6	38	12.5883	1.43446	25.4117	2.20479	R	
7	35	9.4127	1.82328	25.5873	2.23068	R	
8	34	9.1477	1.22991	24.8523	2.15184	R	
37	6	61.8065	7.37645	-55.8065	-6.22045	R	X
43	4	11.8768	2.79401	-7.8768	-0.69870		X
66	2	11.8330	2.82795	-9.8330	-0.87288		X
97	0	11.6039	2.78651	-11.6039	-1.02914		X
143	0	16.4614	4.25569	-16.4614	-1.52323		X
144	0	13.3460	3.30804	-13.3460	-1.19872		X
155	0	19.5698	5.20947	-19.5698	-1.88519		X

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.

General Regression Analysis: TotalMedals versus Income, PopnSize, GDP

Regression Equation

TotalMedals = 1.42638 + 0.155803 Income + 8.3513 PopnSize + 80.4512 GDP

Coefficients

Term	Coef	SE Coef	T	P	95% CI
Constant	1.4264	0.62848	2.2696	0.024	(0.1870, 2.6657)
Income	0.1558	0.21382	0.7287	0.467	(-0.2658, 0.5774)
PopnSize	8.3513	4.61113	1.8111	0.072	(-0.7416, 17.4442)
GDP	80.4512	4.63077	17.3732	0.000	(71.3195, 89.5828)

Summary of Model

S = 7.33966 R-Sq = 70.69% R-Sq(adj) = 70.25%
PRESS = 18906.4 R-Sq(pred) = 48.32%

Analysis of Variance

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	3	25860.9	25860.9	8620.3	160.019	0.000000
Income	1	1018.1	28.6	28.6	0.531	0.467059
PopnSize	1	8583.2	176.7	176.7	3.280	0.071630
GDP	1	16259.6	16259.6	16259.6	301.827	0.000000
Error	199	10720.2	10720.2	53.9		
Total	202	36581.2				

Fits and Diagnostics for Unusual Observations

Obs	TotalMedals	Fit	SE Fit	Residual	St Resid		
1	104	127.046	6.20532	-23.0457	-5.87935	R	X
2	88	71.061	5.17289	16.9394	3.25327	R	X
3	82	17.781	0.81094	64.2187	8.80345	R	
4	65	22.036	1.05570	42.9643	5.91522	R	
6	38	50.424	2.35893	-12.4236	-1.78750		X
7	35	14.595	1.19018	20.4047	2.81735	R	
9	28	11.772	0.62816	16.2282	2.21917	R	
12	20	3.203	0.57594	16.7967	2.29556	R	
37	6	25.077	5.11845	-19.0771	-3.62654	R	X
43	4	6.900	1.78872	-2.8996	-0.40735		X
66	2	4.441	1.83703	-2.4407	-0.34347		X
97	0	2.988	1.82939	-2.9884	-0.42042		X
143	0	3.699	2.78784	-3.6988	-0.54478		X
144	0	3.282	2.16924	-3.2819	-0.46806		X
155	0	4.168	3.40933	-4.1676	-0.64119		X

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.

General Regression Analysis: TotalMedals versus GDP, GDPSQ

```
Regression Equation

TotalMedals  =  0.741716 + 142.183 GDP - 49.3902 GDPSQ


Coefficients

Term          Coef  SE Coef        T      P        95% CI
Constant     0.742  0.49623   1.4947  0.137  ( -0.237,   1.720)
GDP        142.183  8.49562  16.7360  0.000  (125.430, 158.935)
GDPSQ      -49.390  6.70554  -7.3656  0.000  (-62.613, -36.168)


Summary of Model

S = 6.54954      R-Sq = 76.55%        R-Sq(adj) = 76.31%
PRESS = 11557.0  R-Sq(pred) = 68.41%


Analysis of Variance

Source       DF   Seq SS   Adj SS   Adj MS        F          P
Regression    2  28001.9  28001.9  14000.9  326.389  0.0000000
  GDP         1  25674.7  12015.0  12015.0  280.093  0.0000000
  GDPSQ       1   2327.2   2327.2   2327.2   54.252  0.0000000
Error       200   8579.3   8579.3     42.9
Total       202  36581.2


Fits and Diagnostics for Unusual Observations

Obs  TotalMedals      Fit   SE Fit  Residual  St Resid
  1          104  102.752  6.45642    1.2478   1.13380     X
  2           88   77.891  3.12905   10.1093   1.75699     X
  3           82   25.469  1.27486   56.5310   8.79960  R
  4           65   32.278  1.59073   32.7217   5.15024  R  X
  5           44   45.300  2.14847   -1.2999  -0.21010     X
  6           38   67.174  2.88023  -29.1737  -4.95962  R  X
  7           35   20.786  1.05358   14.2141   2.19887  R
  8           34   36.632  1.78521   -2.6322  -0.41771     X
 10           28   29.689  1.47203   -1.6893  -0.26471     X
 12           20    3.091  0.46420   16.9087   2.58818  R
 14           17    2.733  0.46653   14.2666   2.18381  R
 16           17   32.732  1.61131  -15.7320  -2.47817  R  X
 37            6   23.245  1.16985  -17.2455  -2.67611  R

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.
```

General Regression Analysis: TotalMedals versus Income, PopnSize, GDP, ...

```
Regression Equation

TotalMedals  =  0.421726 - 0.261726 Income + 32.3141 PopnSize + 235.479 GDP -
                0.0702992 IncomeSQ - 436.605 PopnSQ + 0.00522077 IncomeCubed +
                325.39 PopnCubed - 367.388 GDPSQ + 176.442 GDPCubed

Coefficients

Term              Coef  SE Coef         T      P         95% CI
Constant         0.422   0.7783   0.54183  0.589  (   -1.113,    1.957)
Income          -0.262   0.8921  -0.29338  0.770  (   -2.021,    1.498)
PopnSize        32.314  23.7735   1.35925  0.176  ( -14.575,   79.203)
GDP            235.479  23.7266   9.92469  0.000  ( 188.682,  282.276)
IncomeSQ        -0.070   0.1663  -0.42264  0.673  (   -0.398,    0.258)
PopnSQ        -436.605  86.6421  -5.03918  0.000  (-607.491, -265.718)
IncomeCubed      0.005   0.0075   0.69333  0.489  (   -0.010,    0.020)
PopnCubed      325.390  58.6508   5.54791  0.000  ( 209.711,  441.069)
GDPSQ         -367.388  62.8817  -5.84253  0.000  (-491.412, -243.365)
GDPCubed       176.442  33.6018   5.25097  0.000  ( 110.168,  242.716)


Summary of Model

S = 6.02744      R-Sq = 80.83%           R-Sq(adj) = 79.94%
PRESS = 1525792  R-Sq(pred) = -4070.98%


Analysis of Variance

Source          DF   Seq SS   Adj SS   Adj MS        F         P
Regression       9  29569.5  29569.5  3285.50  90.4349  0.000000
  Income         1   1018.1      3.1     3.13   0.0861  0.769546
  PopnSize       1   8583.2     67.1    67.12   1.8476  0.175653
  GDP            1  16259.6   3578.5  3578.48  98.4994  0.000000
  IncomeSQ       1    189.0      6.5     6.49   0.1786  0.673026
  PopnSQ         1      0.1    922.5   922.54  25.3933  0.000001
  IncomeCubed    1    151.9     17.5    17.46   0.4807  0.488938
  PopnCubed      1   1550.2   1118.2  1118.21  30.7793  0.000000
  GDPSQ          1    815.7   1240.1  1240.13  34.1352  0.000000
  GDPCubed       1   1001.7   1001.7  1001.72  27.5727  0.000000
Error          193   7011.7   7011.7    36.33
Total          202  36581.2


Fits and Diagnostics for Unusual Observations

Obs  TotalMedals      Fit   SE Fit  Residual  St Resid
  1          104  103.833  6.02703    0.1668   2.37727  R  X
  2           88   87.413  5.95434    0.5865   0.62672     X
  3           82   28.827  1.80391   53.1731   9.24561  R
  4           65   37.032  1.93642   27.9681   4.89988  R
  5           44   43.574  2.38349    0.4256   0.07687     X
  6           38   43.246  5.40144   -5.2457  -1.96115     X
 12           20    4.726  0.71571   15.2740   2.55214  R
 13           18   29.622  1.76379  -11.6218  -2.01641  R
 14           17    3.452  0.60137   13.5476   2.25892  R
 16           17   30.681  2.16832  -13.6806  -2.43258  R
 17           14    1.917  0.49001   12.0830   2.01132  R
 33            7   21.011  1.50994  -14.0107  -2.40105  R
 37            6    6.794  5.88443   -0.7940  -0.60837     X
 60            2    4.908  2.53393   -2.9080  -0.53173     X
```

```
 82            1   13.758  0.99738  -12.7576  -2.14617  R
143            0   -2.292  3.03838    2.2924   0.44036     X
144            0   -3.717  2.64107    3.7171   0.68606     X
155            0    1.970  5.59977   -1.9703  -0.88358     X

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.
```

General Regression Analysis: TotalMedals versus GDP, GDPSQ, GDPCubed

```
Regression Equation

TotalMedals  =  0.522881 + 162.695 GDP - 113.297 GDPSQ + 33.6892 GDPCubed


Coefficients

Term           Coef  SE Coef         T      P         95% CI
Constant      0.523   0.5170   1.01144  0.313  (  -0.497,   1.542)
GDP         162.695  16.3897   9.92663  0.000  ( 130.375, 195.015)
GDPSQ      -113.297  44.2212  -2.56206  0.011  (-200.500, -26.095)
GDPCubed     33.689  23.0436   1.46198  0.145  ( -11.752,  79.130)


Summary of Model

S = 6.53100     R-Sq = 76.80%          R-Sq(adj) = 76.45%
PRESS = 324484  R-Sq(pred) = -787.03%


Analysis of Variance

Source       DF   Seq SS   Adj SS   Adj MS        F         P
Regression    3  28093.0  28093.0  9364.34  219.542  0.000000
  GDP         1  25674.7   4203.0  4203.03   98.538  0.000000
  GDPSQ       1   2327.2    280.0   279.99    6.564  0.011145
  GDPCubed    1     91.2     91.2    91.17    2.137  0.145325
Error       199   8488.1   8488.1    42.65
Total       202  36581.2


Fits and Diagnostics for Unusual Observations

Obs  TotalMedals      Fit   SE Fit  Residual  St Resid
  1          104  104.339  6.52901   -0.3393  -2.10896  R  X
  2           88   71.783  5.21431   16.2170   4.12379  R  X
  3           82   27.071  1.67847   54.9287   8.70277  R  X
  4           65   33.757  1.88142   31.2426   4.99551  R  X
  5           44   45.781  2.16752   -1.7812  -0.28911     X
  6           38   63.792  3.68781  -25.7917  -4.78496  R  X
  8           34   37.889  1.97690   -3.8892  -0.62480     X
 10           28   31.248  1.81399   -3.2475  -0.51761     X
 12           20    3.196  0.46838   16.8041   2.57963  R
 13           18   25.834  1.63192   -7.8341  -1.23883     X
 14           17    2.791  0.46686   14.2093   2.18124  R
 16           17   34.193  1.89226  -17.1933  -2.75054  R  X
 37            6   24.829  1.59181  -18.8289  -2.97265  R  X

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.
```

General Regression Analysis: TotalMedals versus Ln(Income), Ln(PopnSize), ...

```
Regression Equation

TotalMedals  =  19.3377 + 3.3851e+014 Ln(Income) + 3.3851e+014 Ln(PopnSize) -
                3.3851e+014 Ln(GDP)


Coefficients

Term                   Coef      SE Coef         T      P
Constant       1.93377E+01  2.28667E+00   8.45669  0.000
Ln(Income)     3.38510E+14  1.20156E+14   2.81726  0.005
Ln(PopnSize)   3.38510E+14  1.20156E+14   2.81726  0.005
Ln(GDP)       -3.38510E+14  1.20156E+14  -2.81726  0.005

Term                     95% CI
Constant      ( 1.48284E+01,  2.38469E+01)
Ln(Income)    ( 1.01568E+14,  5.75451E+14)
Ln(PopnSize)  ( 1.01568E+14,  5.75451E+14)
Ln(GDP)       (-5.75451E+14, -1.01568E+14)


Summary of Model

S = 11.2299      R-Sq = 31.40%        R-Sq(adj) = 30.36%
PRESS = 28371.0  R-Sq(pred) = 22.44%


Analysis of Variance

Source            DF   Seq SS   Adj SS   Adj MS        F          P
Regression         3  11485.1  11485.1  3828.37  30.3572  0.0000000
  Ln(Income)       1   2603.7    974.6   974.59   7.7281  0.0059584
  Ln(PopnSize)     1   7880.4    973.5   973.53   7.7196  0.0059853
  Ln(GDP)          1   1000.9    970.5   970.46   7.6953  0.0060635
Error            199  25096.0  25096.0   126.11
Total            202  36581.2


Fits and Diagnostics for Unusual Observations

Obs  TotalMedals      Fit   SE Fit  Residual  St Resid
  1          104  22.3188  2.22695   81.6812   7.42092  R
  2           88  17.5744  2.34294   70.4256   6.41237  R
  3           82  16.0363  1.62063   65.9637   5.93607  R
  4           65  18.3268  1.70406   46.6732   4.20484  R
  5           44  19.2273  1.81269   24.7727   2.23527  R
143            0   7.1224  2.88369   -7.1224  -0.65624     X
155            0   8.0333  2.99644   -8.0333  -0.74226     X
159            0  -9.8832  2.94415    9.8832   0.91197     X
195            0  -9.1765  2.74811    9.1765   0.84277     X

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.
```

General Regression Analysis: TotalMedals versus Ln(GDP), Ln(GDP)SQ

Regression Equation

TotalMedals = 47.0882 + 12.9589 Ln(GDP) + 0.838832 Ln(GDP)SQ

Coefficients

Term	Coef	SE Coef	T	P	95% CI
Constant	47.0882	2.88544	16.3192	0.000	(41.3984, 52.7779)
Ln(GDP)	12.9589	0.97937	13.2319	0.000	(11.0277, 14.8901)
Ln(GDP)SQ	0.8388	0.07854	10.6806	0.000	(0.6840, 0.9937)

Summary of Model

```
S = 9.13103      R-Sq = 54.42%        R-Sq(adj) = 53.96%
PRESS = 18686.6  R-Sq(pred) = 48.92%
```

Analysis of Variance

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	2	19906.0	19906.0	9953.0	119.375	0
Ln(GDP)	1	10394.9	14597.7	14597.7	175.083	0
Ln(GDP)SQ	1	9511.1	9511.1	9511.1	114.075	0
Error	200	16675.1	16675.1	83.4		
Total	202	36581.2				

Fits and Diagnostics for Unusual Observations

Obs	TotalMedals	Fit	SE Fit	Residual	St Resid		
1	104	52.6564	3.28046	51.3436	6.02525	R	X
2	88	43.0142	2.60303	44.9858	5.13998	R	X
3	82	27.6600	1.61688	54.3400	6.04668	R	
4	65	30.3982	1.78029	34.6018	3.86362	R	
5	44	34.6578	2.04712	9.3422	1.04985		X
6	38	40.4200	2.42669	-2.4200	-0.27492		X
37	6	26.6466	1.55834	-20.6466	-2.29481	R	
115	0	5.6888	1.99801	-5.6888	-0.63849		X
136	0	6.1157	2.05395	-6.1157	-0.68739		X
151	0	8.2091	2.32581	-8.2091	-0.92970		X
159	0	20.0642	3.80442	-20.0642	-2.41716	R	X
166	0	5.9516	2.03247	-5.9516	-0.66857		X
195	0	16.4843	3.36695	-16.4843	-1.94216		X

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.

References

Official London 2012 website. (2012, August 12). Retrieved from http://www.london2012.com/medals/medal-count/

Population data obtained from *International Data Base* (IDB), United States Census Bureau. Accessed on March 30, 2012.

Wikipedia. (2011, August). Retrieved from http://en.wikipedia.org/wiki/List_of_countries_by_GDP_(PPP)_per_capita

World Development Indicators database, World Bank. Accessed on 12 July 2012. Note: The GDP (PPP) per person employed was calculated by dividing the GDP (PPP) by the employed population. The employed population was obtained by subtracting the unemployed population from the labor force. Multiplying the unemployment rate by the labor force and dividing the result by 100 calculated the unemployed population. The GDP (PPP) per capita was calculated by dividing the GDP (PPP) by the total population.

The World Factbook, United States Central Intelligence Agency (GDP (PPP), labor force and unemployment rate). Accessed on March 30, 2012. Note: The GDP (PPP) per person employed was calculated by dividing the GDP (PPP) by the employed population. The employed population was obtained by subtracting the unemployed population from the labor force. Multiplying the unemployment rate by the labor force and dividing the result by 100 calculated the unemployed population. The GDP (PPP) per capita was calculated by dividing the GDP (PPP) by the total population.

Nancy Carter, Professor of Statistics
Department of Mathematics & Statistics
California State University, Chico
Chico, CA 95929-0525
ncarter@csuchico.edu

Neil Schwertman, Professor of Statistics
Department of Mathematics & Statistics
California State University, Chico 95929-0525
nschwertman@csuchico.edu

Nathan Felton, student
Department of Mathematics & Statistics
California State University, Chico 95929-0525
