Title: Predicting Olympic Medal Counts
Author: Curtis Gross
Date: 1/8/2013

For my project I've decided to look at the economic factors leading to Olympic success. The 2012 Olympics from this past Summer came with a lot of claims for predicting medal counts. Many articles on the subject mention using regression, but don't show any of their steps. This is my attempt at explaining the factors for Olympic success.

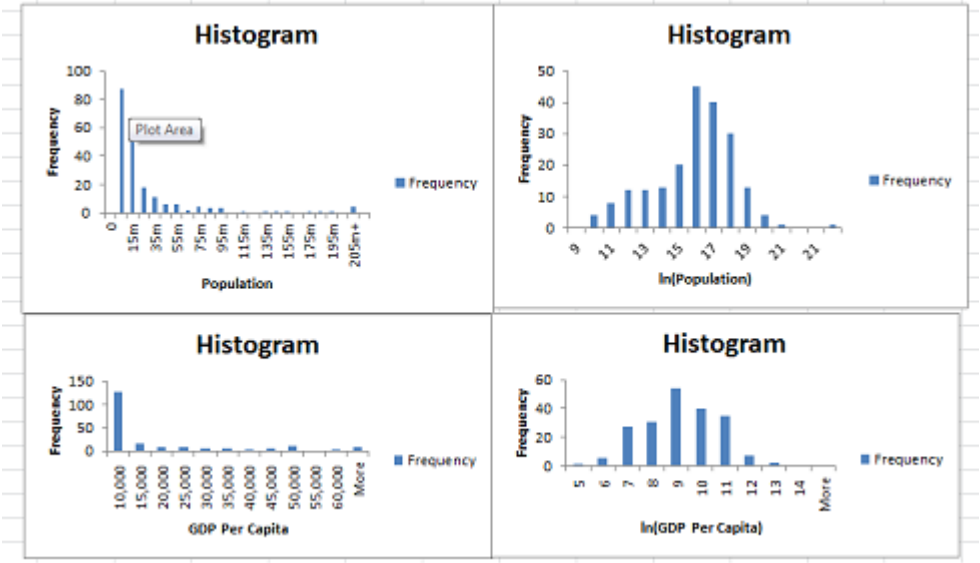
Success is defined as the total number of medals won(gold, silver, and bronze). The factors that I decided to work with are: Gross Domestic Product (GDP), Population, monetary incentives for winning medals, and whether a country is or has been communist. GDP and population are clear explanatory variables for consideration. The other two explanatory variables (although the Communist variable will be treated as a dummy variable) were hot button topics during the Olympic games, and I was curious about their effect on success.

This assignment will use data from the 2012 olympics, broken out by each participating country. Most of the data was collected from Wikipedia, which has lots of relevant lists. The medal incentives were difficult to find, and I had to pull from a number of sources to build a decent, yet probably not exhaustive list. All of my data is found in the attached excel file.

The first concern I had with my explanatory variables was the possibility of collinearity between GDP and population, because a larger workforce implies more production. The correlation between population and GDP is correl() = .507, which is more than I would like. For this reason I chose to use GDP per Capita = $\frac{\text{GDP}}{\text{Population}}$. The correlation between population and GDP per Capita is -.057; much more acceptable.

Before finding a linear model, let's begin by adjusting our data to meet the basic assumptions for normalized linear regression in Fox's text. My concerns lie with population and GDP per Capita. There is a positive (right) skew in both variables due to countries like China, India, U.S., and Luxembourg. To create more symmetry I will employ the natural log to "tighten up" their

distributions. Below is the effect before and after.



variables.png

These transformed variables are much closer to normal than before.

Now I would like to see the effect that the medal incentives have on a simplified version of the model. For this portion I will be using the Regression tool in Excel.

Regression Statistics	
Multiple R	0.541548
R Square	0.293275
Adjusted R Square	0.282621
Standard Error	11.40466
Observations	203

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regression	3	10740.9	3580.31	27.5268	6.2E-15
Residual	199	25883.2	130.066		
Total	202	36624.1			

	Coefficients	*ndard Er*	*t Stat*	*P-value*	*Lower 95%U*
Intercept	(67.886)	8.124	(8.356)	1.1E-14	-83.9064
ln(Population)	2.847	0.359	7.922	1.6E-13	2.138348
ln(gdp per cap)	3.360	0.539	6.239	2.6E-09	2.298281
Gold medal Incentiv	-8.3E-07	5.9E-06	(0.141)	0.8879	-1.2E-05

There are a few issues with the medal incentives. The test statistic is -.141, which yielded a p-value of .888. Thus it would be better to just leave this variable out of the model. The coefficient is near 0, and on the negative side. How could increasing the incentive for winning a medal worsen your predicted medal count? The reason is because many countries that produce small numbers of medal winners find it easy to give high incentives, since no one from their country is likely to win. This is evident in Singapore, which rewards a gold medal winner with $800,000 (U.S.), but Singapore only yielded 2 bronze medals in these Olympic games. It's easy to promise a lot of money when your country never wins anything. A symmetric argument could be made about Great Britain, which doesn't give any monetary reward for a medal, but won 65 medals nonetheless.

The other reason that incentives aren't reliable is because I couldn't find incentives for smaller countries, resulting in about 75% of the incentives data to be 0. Also, some countries give out non-monetary prizes, like homes, cars, and even lifetime supplies of beer (Germany), or a milk cow (Zimbabwe).

My list also doesn't take into account sponsorships that result from winning, such as Michael Phelps on the Wheaties box.

For the last effort to make a slightly better model I want to analyze the effect that communism has on success. Much of my research has indicated that national pride is highest in communist countries, and much more funding is allocated towards Olympic training. Within my data I created a field called "Communist?" and labeled its entries with a 1 if the country is, or was communist, and a 0 otherwise. The reason for adding the "or was communist" portion is because there are currently only 5 countries that are considered communist, and that makes for a poor dummy variable.

I will separate the data into communist and non-communist countries and create the linear regression for each, then deduce the dummy variable regression from there.

SUMMARY OUTPUT	Non-Communist				
Regression Statistics					
Multiple R	0.534256969				
R Square	0.285430509				
Adjusted R Square	0.277401638				
Standard Error	10.801582				
Observations	181				
ANOVA					
	df	*SS*	*MS*	*F*	*ignificance F*
Regression	2	8295.654546	4147.83	35.5505	1E-13
Residual	178	20768.00291	116.674		
Total	180	29063.65746			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%U*
Intercept	-62.14325471	7.937113746	-7.82945	4.2E-13	-77.8062
ln(GDP Per Capita)	3.272663376	0.536838411	6.09618	6.6E-09	2.21328
ln(population)	2.495294393	0.350542982	7.11837	2.6E-11	1.80354

SUMMARY OUTPUT	Communist					
Regression Statistics						
Multiple R	0.713574917					
R Square	0.509189163					
Adjusted R Square	0.457524864					
Standard Error	13.76218719					
Observations	22					
ANOVA						
	df	*SS*	*MS*	*F*	*ignificance F*	
Regression	2	3733.305506	1866.65	9.85573	0.00116	
Residual	19	3598.55813	189.398			
Total	21	7331.863636				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Up*
Intercept	-152.6201551	36.38816094	-4.19423	0.00049	-228.781	-
ln(GDP per capita)	5.206094228	2.194358896	2.37249	0.02838	0.61325	
ln(Pop)	7.289664702	1.777485163	4.10111	0.00061	3.56935	

It is clear from the Regression output that both models are significant, due to their large F-values. The non-communist and communist fitted regressions are respectively:

$$\hat{Y}_{NC} = -62.14 + 3.273X_1 + 2.500X_2$$
$$\hat{Y}_{C} = -152.62 + 5.206X_1 + 7.29X_2$$

where X_1 is the natural log of GDP per capita and X_2 is the natural log of Population. Therefore the dummy regression model is

$$\hat{Y} = \hat{Y}_{NC} + D_1[-90.48 + 1.933X_1 + 4.79X_2].$$

where D_1 is 1 when the country is communist, and 0 otherwise.

The Regression analysis of this model yields:

Regression Statistics						
Multiple R	0.546387					
R Square	0.298539					
Adjusted R Square	0.287964					
Standard Error	11.3621					
Observations	203					
ANOVA						
	df	*SS*	*MS*	*F*	*gnificance F*	
Regression	3	10933.7	3644.57	28.2312	2.97E-15	
Residual	199	25690.4	129.097			
Total	202	36624.1				
	Coefficients	*ndard Er*	*t Stat*	*P-value*	*Lower 95%*	*Up*
Intercept	-68.0409	8.02176	-8.482	5E-15	-83.8594	-
ln(Population)	2.816565	0.35515	7.93061	1.5E-13	2.116222	3
ln(gdp per cap)	3.388778	0.53323	6.35513	1.4E-09	2.337262	4
Communist?	3.169389	2.57622	1.23025	0.22006	-1.91081	8

model.png

The p-value for the dummy variable is too high to be of any significance, which is depressing. It must be that communist countries do not significantly differ from non communist countries in Olympic success. My final model will not include this dummy variable.

Now we're down to 2 reliable explanatory variables, so the final model's analysis is:

Regression Statistics	
Multiple R	0.541483
R Square	0.293204
Adjusted R Square	0.286136
Standard Error	11.37668
Observations	203

ANOVA					
	df	*SS*	*MS*	*F*	*gnificance F*
Regression	2	10738.3	5369.17	41.4835	8.5E-16
Residual	200	25885.8	129.429		
Total	202	36624.1			

	Coefficients	*ndard Er*	*t Stat*	*P-value*	*Lower 95%*	*U*
Intercept	-67.7293	8.02805	-8.4366	6.5E-15	-83.5598	
ln(Population)	2.840053	0.35509	7.99806	1E-13	2.139847	
ln(gdp per cap)	3.350819	0.53302	6.28643	2E-09	2.29975	

model.png

And hence the fitted regression is:

$$\hat{Y} = -67.73 + 3.351X_1 + 2.84X_2.$$

This model is far more significant than the .0001 level, which is nice to know.

Unfortunately the model only has an R^2 of .3, meaning that even though the coefficients are likely good fits, the model is still a poor estimator overall. I will conclude by mentioning some of the steps that one could take to improve upon my 2 variable model.

For the interest of time and simplicity I only began by looking at 4 explanatory variables. Other studies used many variables ranging from number of internet users to host country and to women's equality with men. No doubt the model would benefit significantly from more explanatory variables.

The other large issue is that the variance of the data does not remain constant. To check this I grouped the data into 4 equal sized groups, ranking from the lowest GDP to the highest and analyzed the means and variances

of the medal counts.

Group 1: mean is 0.04 variance is 0.0392
Group 2: mean is 0.92 variance is 5.0955
Group 3: mean is 2.56 variance is 18.864
Group 4: mean is 11.08 variance is 275.014.

There is a clear relationship between the mean growth and the variance growth, and it's not constant. To conclude, a Poisson or Exponential model probably would have been a better fit for the data.