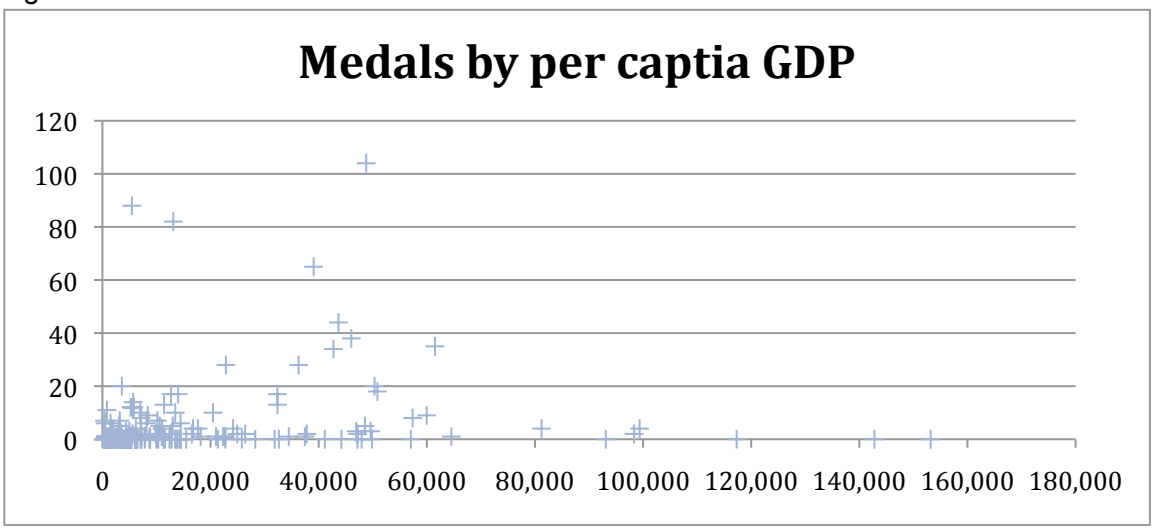
The count of Olympic medals won by country is observational data, not experimental. Explanatory variables are certainly observed and we can never know if all relevant factors have been identified. In this study, I look at 2011 GDP, 2012 population and 2012 Summer Olympic team size as explanatory variables for the number of medals won by country. Of course, team size may be partly explained by population and GDP so it may not be possible to consider that an independent explanatory variable.

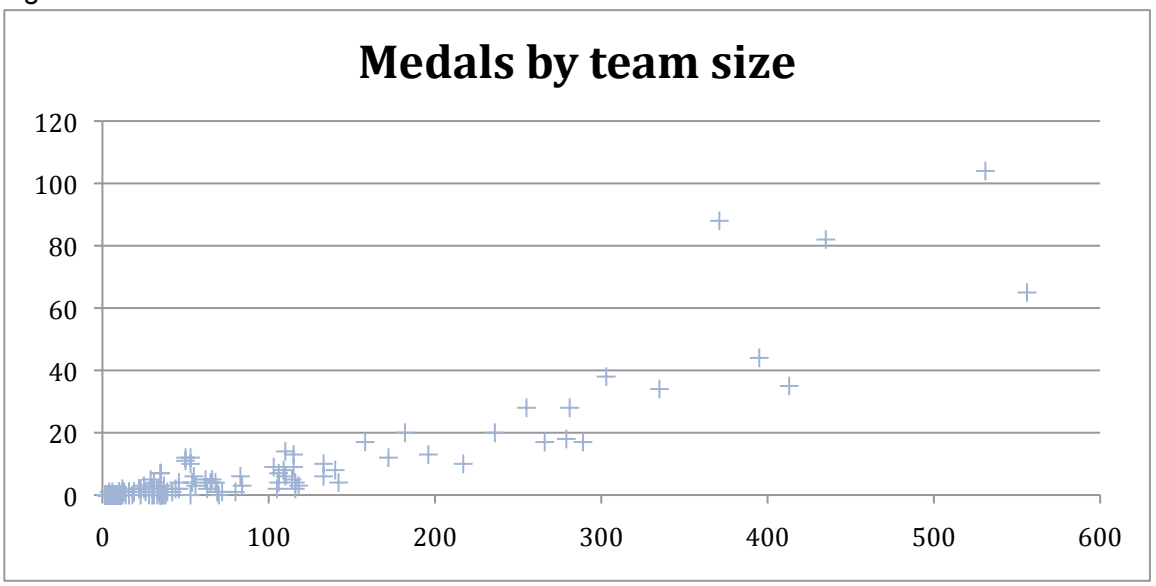
Based on a preliminary scatter plot of medals by per capita GDP, the data appears positively skewed, and at higher per capita GDP has a much larger spread than at lower values. There do not appear to be multiple modes or a strong nonlinearity but it's hard to tell visually if there are heavy tails.

Figure 1



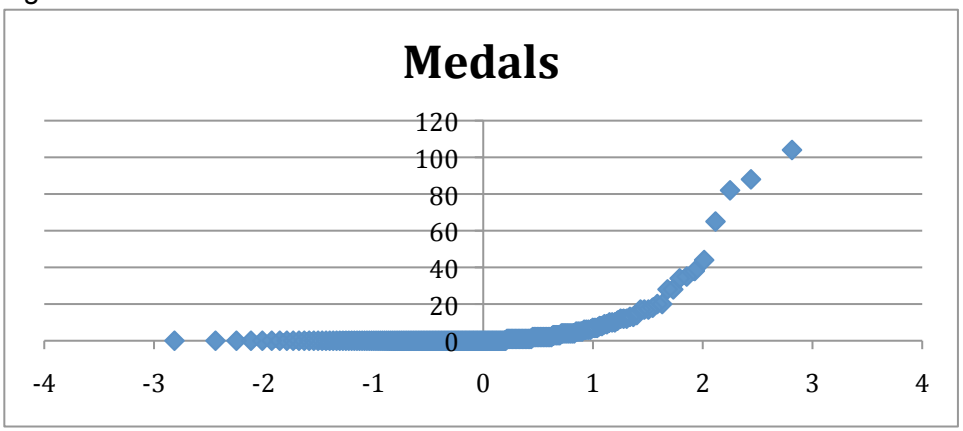
The distribution by team size has a much more obvious correlation but doesn't look completely linear. The spread also increases as team size increases.

Figure 2



The normal quantile-comparision plot confirms that the data is positively skewed and depicts that the majority of countries took home zero medals. A boxplot would be even less helpful since the minimum, first quartile and median are all equal to 0. The third quartile is 3 and about 20% of the data points are outliers with over half of those being “far outside”.

Figure 3



The scatter plot matrix shows data for medals, GDP and team size.

Figure 4

The data needs to be transformed in order to match the assumptions of classical statistical models. Since the data is positively skewed, the transformation should descend the ladder of powers and roots. Figures 5 through 10 show the medal count by team size under various transformations. As shown in figures 5, 6 and 7, taking the natural log of team size pulls in

the right tail. As shown in figures 8, 9, and 10, taking the natural log of medal count makes the distribution very nearly linear, if you exclude countries with no medals.

Figure 5: ((X^.5)-1)/.5

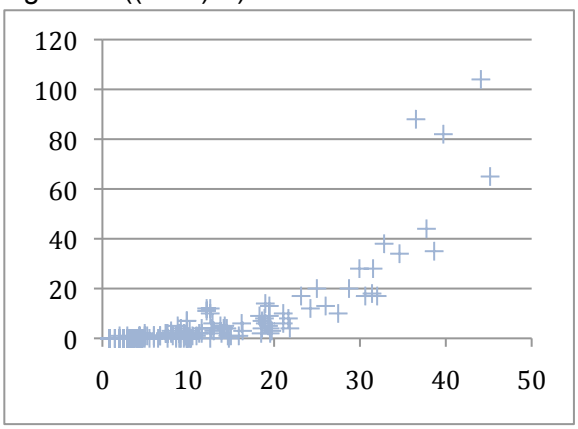


Figure 6: ln(X)

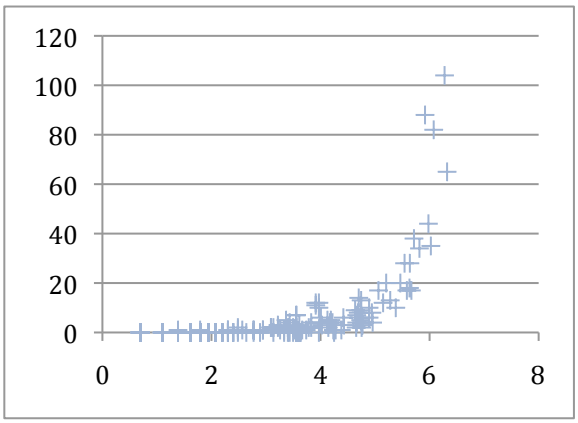


Figure 7: ((X^-.5)-1)/-.5

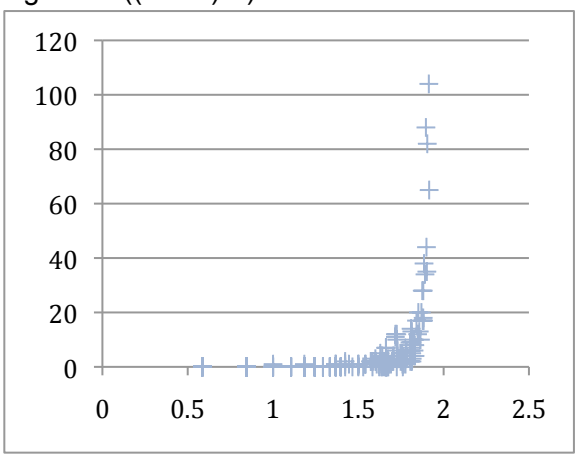


Figure 8: ln(X) and ((Y^.5)-1)/.5

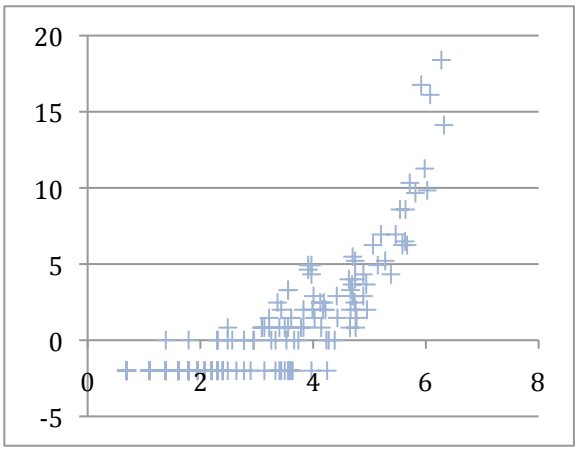


Figure 9: ln(X) and ln(Y)

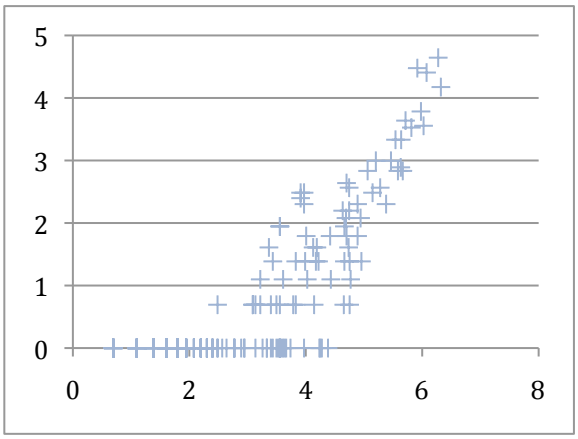
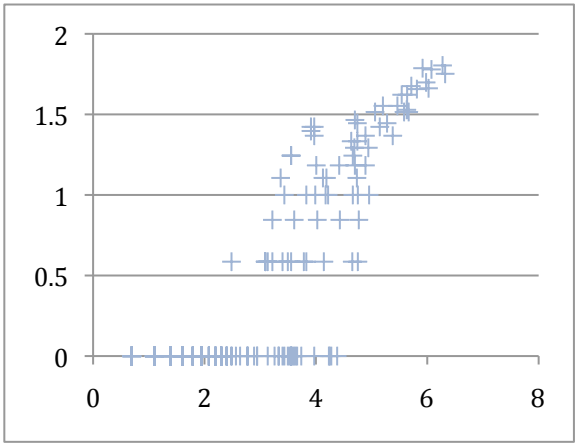


Figure 10



The unequal spread in the pre-transformed data (see figure 2) was mostly cleaned up by descending the ladder of powers. The transformed data (figure 9) has a slight negative association between level and spread but it's much closer to constant.

Further analysis of least squares regression will use X=ln(team size) and Y=ln(medal count). Setting countries who won zero medals to have Y=-1 produces ordinary least squares estimators B=0.8898 and A=-2.4528:

$\overline{X} = 2.8413$
$\overline{Y} = 0.0754$
$\Sigma (X_i - \overline{X})(Y_i - \overline{Y}) = 397.4040$
$\Sigma (X_i - \overline{X})^2 = 446.6173$
$B = \Sigma (X_i - \overline{X})(Y_i - \overline{Y}) / \Sigma (X_i - \overline{X})^2 = 397.4040 / 446.6173 = 0.8898$
$A = \overline{Y} - B\overline{X} = 446.6173 - 0.8898 * 2.8413 = -2.4528$

The calculations for the total sum of squares, residual sum of squares, regression sum of squares and the standard error of the regression are as follows:

TSS (using data points and average) = $\Sigma E_i'^2 = \Sigma (Y_i - \overline{Y})^2 = 462.3465$
RSS (using data points and regression line) = $\Sigma E_i^2 = \Sigma (Y_i - \hat{Y}_i)^2 = 108.7329$
RegSS (regression sum of squares) = $\Sigma (\hat{Y}_i - \overline{Y})^2 = TSS - RSS = 353.6135$
$S_E = sqrt(\Sigma E_i^2 / (n-2)) = sqrt(RSS / 202) = 0.7337$

The residual standard error of 0.7337 represents 2.08 medals, which is not that small considering the average medal count of 4.72. The square of the correlation coefficient r^2 = 353.6135 / 462.3465 = 0.7648 and the correlation coefficient r = 0.8745. The linear regression line is therefore a fairly decent estimate of the medal data.

Note that if countries with no medals are excluded from the data, a very similar least squares regression line is produced with A=-2.5064 and B=0.9693, but fit is better. TSS = 132.0840, RSS = 39.3780 and RegSS = 92.7061. This makes sense because there are a large number of countries that took home no medals and they have team sizes varying from 2 to 70. The residual standard error is 0.4415, however r = 0.8378 and therefore is not better.

For multiple correlation, X1=ln(team size), X2=ln(GDP) and Y=ln(medal count). Again, countries with zero medals are set to Y=-1.

n = 204
$\overline{X}1 = 2.8413$
$\overline{X}2 = 23.9461$
$\overline{Y} = 0.0754$
$\Sigma X1_i = 579.6151$
$\Sigma X2_i = 4,884.9981$
$\Sigma Y_i = 15.3717$
$\Sigma X1_i^2 = 2,093.4491$
$\Sigma X2_i^2 = 118,224.1359$
$\Sigma X1_iX2_i = 14,488.5346$
$\Sigma X1_iY_i = 441.0788$
$\Sigma X2_iY_i = 894.7153$

For regression coefficients A, B1 and B2:
$An + B1 \Sigma X1_i + B2 \Sigma X2_i = \Sigma Y_i$
$A \Sigma X1_i + B1 \Sigma X1_i^2 + B2 \Sigma X1_iX2_i = \Sigma X1_iY_i$
$A \Sigma X2_i + B1 \Sigma X1_iX2_i + B2 \Sigma X2_i^2 = \Sigma X2_iY_i$

A*204 + B1*579.6151 + B2*4,884.9981 = 15.3717
A*579.6151 + B1*2,093.4491 + B2*14,488.5346 = 441.0788
A*4,884.9981 + B1*14,488.5346 + B2*118,224.1359 = 894.7153

B1 = 0.9398
B2 = -0.0367
A = -1.7167

The partial coefficient for the effect of GDP on medal count is actually slightly negative, but close to zero. A change in $1 of GDP has very little effect on medal count, after team size is taken into account, but an increase seems to be correlated with a slight decrease in medals.

The standard error of regression is

S_E = sqrt(Σ E_i^2 / (n-k-1)) = sqrt(108.1719 / 201) = 0.7336

We have n-k-1 = 204 – 2 - 1 degrees of freedom because we have 2 explanatory variables. We also have the sums of squares and correlation coefficient as follows.

TSS = 462.3465
RegSS = 354.1746
RSS = 108.1719
r^2 = RegSS / TSS = 0.7660
r = 0.8752

As expected the multiple correlation of 0.8752 is greater than the single correlation coefficient of 0.8745, but not by much. Adding GDP as a second explanatory variable does not make a significant difference to the correlation of the regression line to the data.

The adjusted squared multiple correlation is
1 – (RSS / (n-k-1)) / (TSS / (n-1)) = 0.7649

Given this, the correlation between GDP and team size is, as expected, very high.
r12 = Σ $X1_iX2_i$ / sqrt(Σ $X1_i^2$ * Σ $X2_i^2$) = 0.9210

Medal count could have been regressed on a number of other explanatory variables as well. Because of the social science nature of the data though, perfect collinearity isn't really an issue. The multiple regression using team size and GDP shows a relationship between the two. Likely, per capita GDP would also have a strong relationship with team size. Countries where people have the wealth to dedicate to recreational activities understandably send more athletes to the Olympics. Likewise, athletes with the wealth to train with the best coaches and in the best facilities understandably win more medals. It would be impossible however, to redesign the study to decrease the collinearity since explanatory variables must be observational.

Going back to the single regression model with least squares estimators B=0.8898 and A=-2.4528, recall that the residual standard error was 0.7337. The assumptions of the simple regression model are that the errors are linear, with constant variance, normal, independent and that X is measured without error and not invariant. In this case, the variance of errors is dependent on X with a larger variance at lower X values and X is observed rather than fixed.

For analysis of the data, these characteristics are assumed to hold. Certainly the intercept and slope, A and B, are unbiased linear estimators. The variances of A and B may not be perfect but the calculations are shown here for demonstration.

$V(A) = (S_E\text{^}2 * \Sigma\ x_i\text{^}2) / (n * \Sigma\ (x_i - \bar{X})\text{^}2) = 0.7337\text{^}2 * 2093.4491 / 204 * 446.6173 = 0.0124$
$V(B) = S_E\text{^}2 / \Sigma\ (x_i - \bar{X})\text{^}2 = 0.7337\text{^}2 / 446.6173 = 0.0012$

The t statistic for testing the null hypothesis that B=0 shows that it can be rejected.
$t_0 = (0.8898 - 0) / \text{sqrt}(0.0012) = 25.6306$

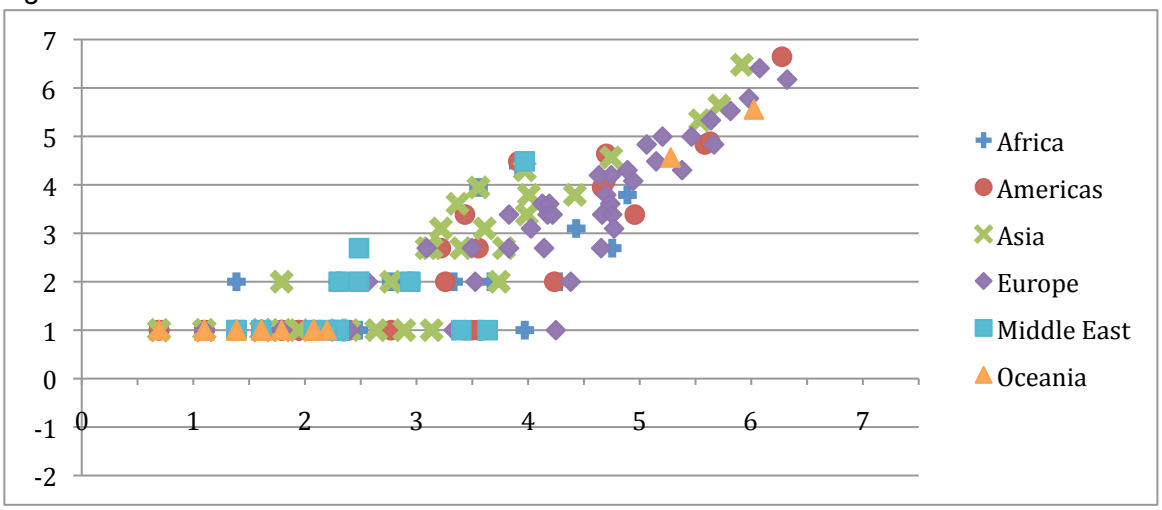
The standard error of the intercept and slope are
$SE(A) = (S_E * \text{sqrt}(\Sigma\ x_i\text{^}2)) / \text{sqrt}(n * \Sigma\ (x_i - \bar{X})\text{^}2) = 0.1112$
$SE(B) = S_E / \text{sqrt}(\ \Sigma\ (x_i - \bar{X})\text{^}2) = 0.0347$

For 204 - 2 degrees of freedom, $t_{0.025} = 1.9718$ so the 95% confidence intervals are
$\alpha = -2.4528 \pm 1.9718 * 0.1112 = -2.4528 \pm 0.2193$
$\beta = 0.8898 \pm 1.9718 * 0.0347 = 0.8898 \pm 0.0684$

Using the multiple regression model, the omnibus null hypothesis, B1=B2=0 has F statistic
$F_0 = (\text{RegSS} / k) / (\text{RSS} / (n - k - 1)) = (354.1746 / 204) / (108.1719 / 201) = 3.2260$

For dummy variable regression, regions of the world are useful as a qualitative explanatory variable. Countries within certain regions may be more competitive or have a culture of athletics. A preliminary look shows that Asian countries seem to be on the higher side of medal count across all team sizes but the slope looks like it would be in line with the whole data set. Africa seems to be in the lower end of the spectrum. Middle Eastern countries may have a different slope. For this purpose, Y is set to ln(medal count) + 2 so that there are no zero values.

Figure 11

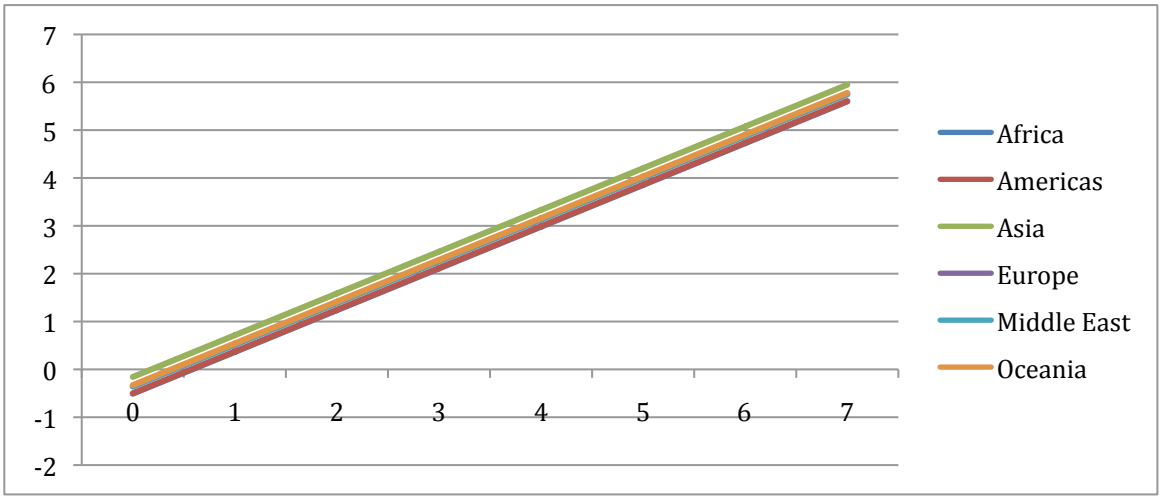


Assuming no interactions between the qualitative explanatory variable of region and the quantitative explanatory variable of team size, we get the following regression formula where D_1 is 1 for Africa, D_2 is 1 for Americas, D_3 is 1 for Asia, D_4 is 1 for Europe, and D_5 is 1 for Middle East.

$Yi = \alpha + \beta X + \gamma_1 D_1 + \gamma_2 D_2 + \gamma_3 D_3 + \gamma_4 D_4 + \gamma_5 D_5$

Using the regression analysis in excel, we get intercept $\alpha = -0.3289$, slope $\beta = 0.8725$ and the vertical separations between regression lines $\gamma_1 = -0.1785$, $\gamma_2 = -0.1711$, $\gamma_3 = 0.1706$, $\gamma_4 = -0.0272$, and $\gamma_5 D = -0.2364$.

Figure 12



The model with no interactions has a regression sum of squares of 357.0522, a residual sum of squares of 90.4453, and 6 degrees of freedom. The mean square for groups is 59.5087 and for residuals is 0.5345 which produces an F-ratio of 111.3376 and a p-value of 1.5298.

When we include interaction between region and team size, the slope of each region's regression changes and we get the following regression formula and slope/intercept results.

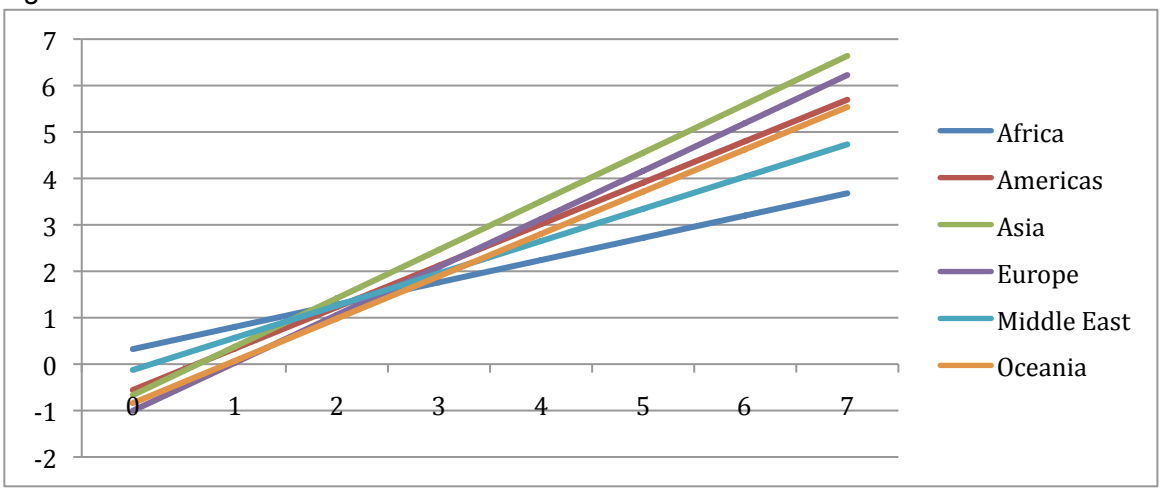
$Yi = \alpha + \beta X + \gamma_1 D_1 + \gamma_2 D_2 + \gamma_3 D_3 + \gamma_4 D_4 + \gamma_5 D_5 + \delta_1 D_1 X + \delta_2 D_2 X + \delta_3 D_3 X + \delta_4 D_4 X + \delta_5 D_5 X$

$\alpha = -0.4088$
$\beta = 0.9111$
$\gamma_1 = 0.7323$
$\gamma_2 = -0.1493$
$\gamma_3 = -0.2588$
$\gamma_4 = -0.5954$
$\gamma_5 = 0.2826$
$\delta_1 = -0.4317$
$\delta_2 = -0.0178$
$\delta_3 = 0.1327$
$\delta_4 = 0.1217$
$\delta_5 = -0.2169$

The model with interactions, shown in Figure 13, has a regression sum of squares of 371.9012, a residual sum of squares of 90.4453, and 11 degrees of freedom. The mean square for groups is 33.8092 and for residuals is 0.4711 which produces an F-ratio of 71.7712 and a p-value of 6.4871.

Both models have a total sum of squares of 462.3465.

Figure 13



Comparing these two models can test the significance of adding regional interactions. The null hypothesis is $\delta_1 = + \delta_2 = \delta_3 = \delta_4 = \delta_5$ and the F-ratio is

F-ratio = ((RegSS1 – RegSS2) / (k1 – k2)) / (RSS / (n-k-1))
= ((371.9012 - 357.0522)/(11-6)) / (90.4453/(204-11-1)) = 6.3044

Using deviation regressors (where S_1 is 1 for Africa, S_2 is 1 for Americas, S_3 is 1 for Asia, S_4 is 1 for Europe, and S_5 is 1 for Middle East and all S_i are -1 for Oceana) produces these intercepts for the model with no interactions.

	with qualitative explanatory variable	no qualitative explanatory variable
a	-0.4028	1.9915
g1	-0.1056	-0.6533
g2	-0.0971	-0.0482
g3	0.2446	0.4440
g4	0.0467	1.1787
g5	-0.1625	-0.4074

As they should be, the ANOVA results based on dummy regressors are the same as those based on deviation regressors, as long as the other explanatory variables and interactions are the same.

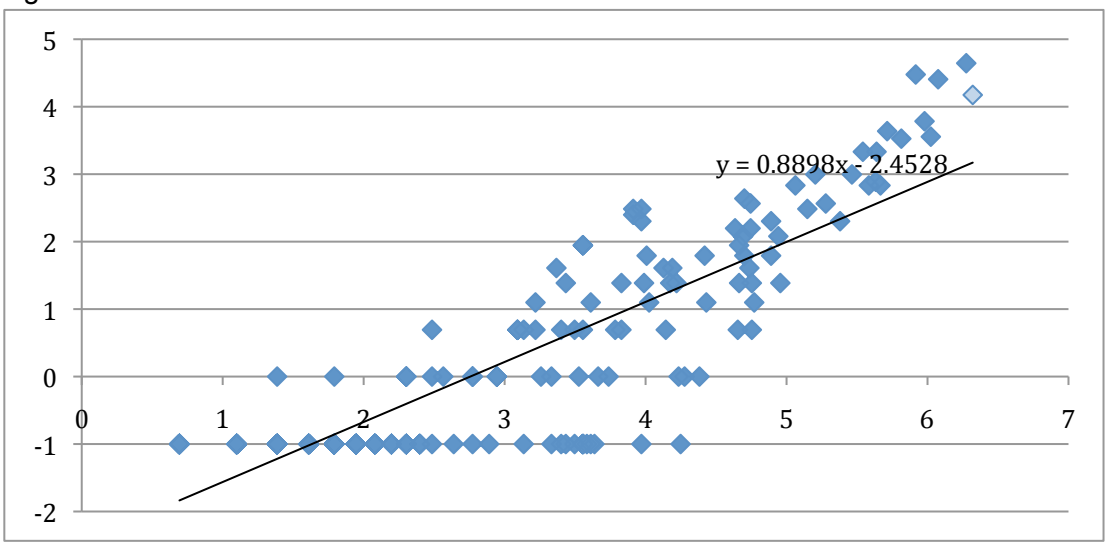
Checking the data visually, there do not appear to be outliers with a great deal of influence. The data points at high-leverage Xs, further outside the center of the distribution, are in line with the rest of the data. Looking back at the single regression model for simplicity, hat-values will be bounded between 1/204 and 1, and the average hat value is $\bar{h} = (k + 1) / n = 2/204$. The actual values calculated in excel range from 0.00490736 to 0.03201028.

It would be difficult to test the impact on the coefficients of deleting each observation in turn, even under the single regression model with no interactions, because of the large number of observations. Although there are no obvious outliers, an example of the studentized residual calculation can be demonstrated by deleting the light blue data point in Figure 14. This turns

out to be the point for the United Kingdom, which hosted the Games and thus (not coincidentally) had the largest team size. It may even make sense to exclude the United Kingdom since prior statisticians have shown that countries tend to win more medals in years when they host the Olympics then in years where they travel to another country.

The regression line *before* deleting the data for the United Kingdom has also been added to the chart. This makes it clear that after the deletion, the slope should decrease

Figure 14



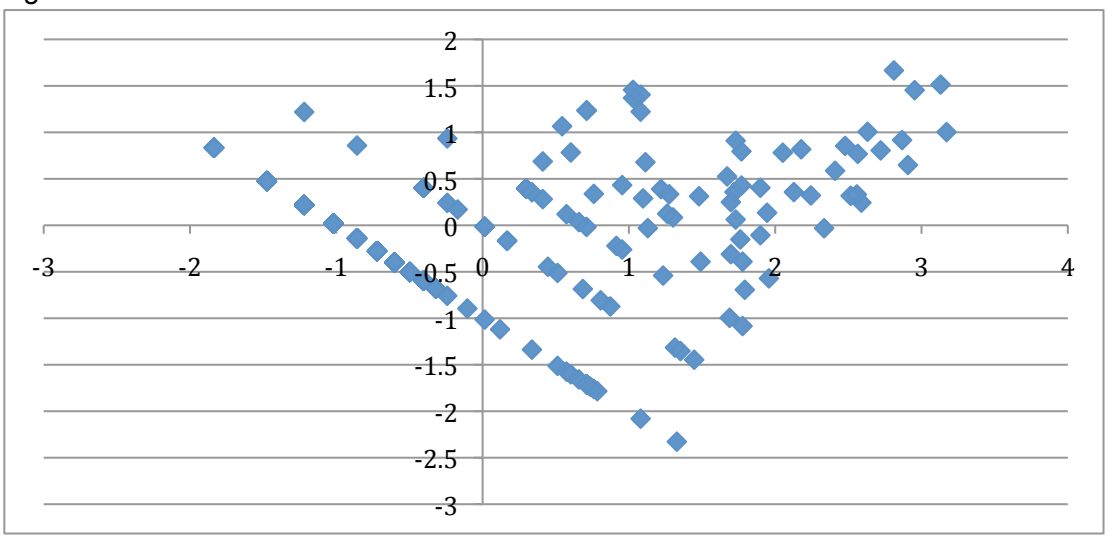
Recall that the slope and intercept for the single regression model including all data points were 0.8898 and -2.4528, respectively. The slope and intercept for the remaining 203 data points after the UK is excluded are 0.8817 and -2.4350.

Figure 14 also highlights that the distribution of errors is not perfectly normal and seems to be skewed in different directions at different places on X. This compromises the efficiency of the least squares estimation and the interpretation of the least squares fit, even after the data has been transformed. Because of the large number of countries who won no medals at all, the data will always be positively skewed.

The errors are also definitely not constant which further impairs the efficiency of the least squares estimator. Olympic medal data has much more error variance at the middle range of expected Y values. A country cannot win less than zero medals of course, however there appears to be a strong correlation between the increase in expected Y and the decrease in error distribution, once Y is past a certain point. This scenario is less common than error variances that increase as expectation of Y grows, but is due to the transformation of the data described previously.

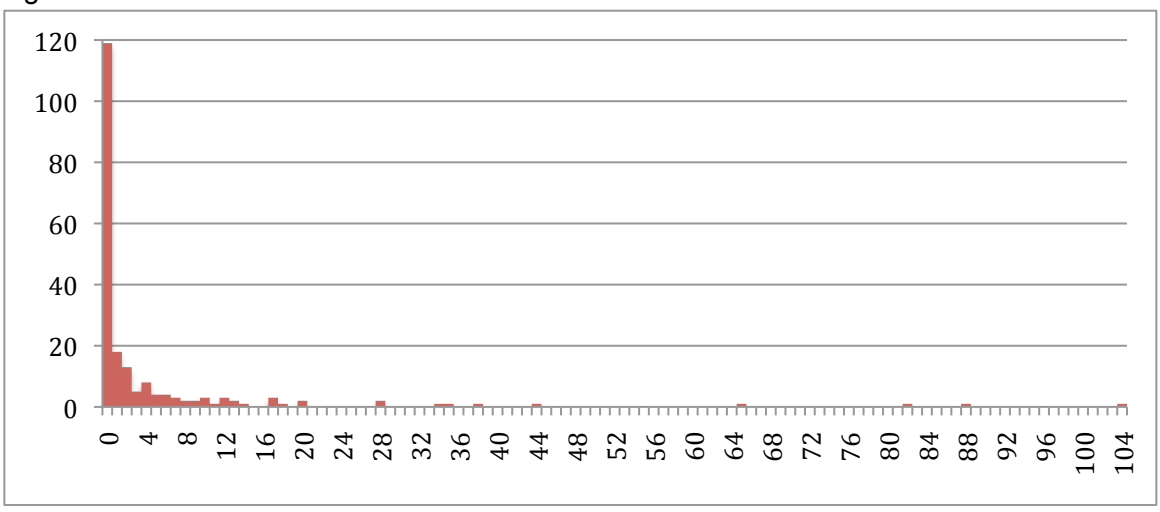
Figure 15 shows the residuals plotted against the fitted values.

Figure 15



Given the total distribution of medal count, a Poisson GLM for counts may be the best model. Figure 16 shows the distribution of pre-transformed medal data, which is highly positively skewed and has a huge majority of zero counts. The conditional distribution of medal count could differ but the lack of symmetry seems to indicate that least squares regression may be an oversimplified model.

Figure 16



Overall, there is not perfect answer to describe the response variable based on explanatory variables but basic regression analysis can help explore the relationships. This particular data set on Olympic medal count by country highlights both the strengths and capabilities of the statistical models described, as well as some of the drawbacks and limitations often encountered in social science studies. Regression analysis is a powerful tool.