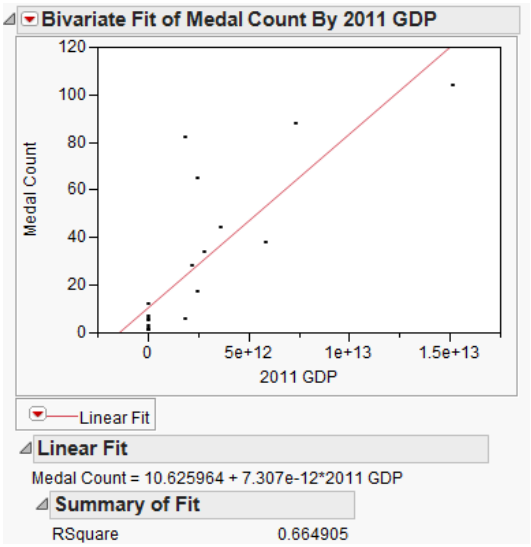
Assignment 1 – Answer Key

Problem 1(5 points – data exploration, visualization, and interpretation): Every four years, many of the world's greatest athletes gather to participate in the Summer Olympics. In addition to individual (or team) prowess, the Olympics is also a highly-watched pageant of national pride and competition. The data set (Olympics.xls under the course documents for week 2) for this problem concerns the performance of various countries in the 2012 London Summer Olympics. For each included country, the data contains medal counts, number of athletes (by gender), national population figures, and national GDP (gross domestic product).

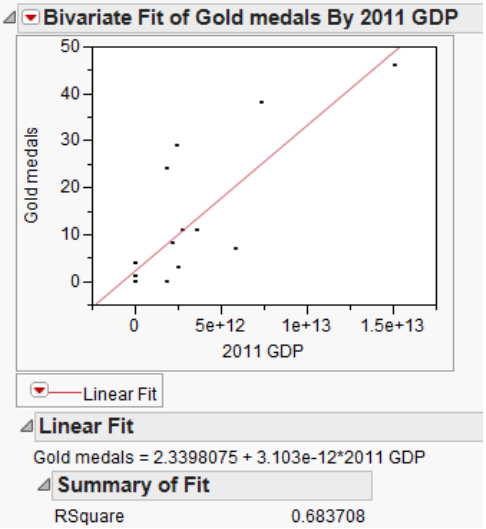
It is your job to distill an interesting story or insight in this data and present it to the general public. You must choose the message you would like to communicate. Is there an important trend or lesson that you would like the public to understand? For example, are there ways to evaluate a country's "performance" beyond raw medal counts, and if so, do any surprises emerge? Is there any relationship between the success in Olympics game and the wealth of the people in country? How good/bad are they compared to the peers?

In your write-up, be sure to include the graph(s) you are using to see the relationships and clearly indicate the intended message of your graphic.

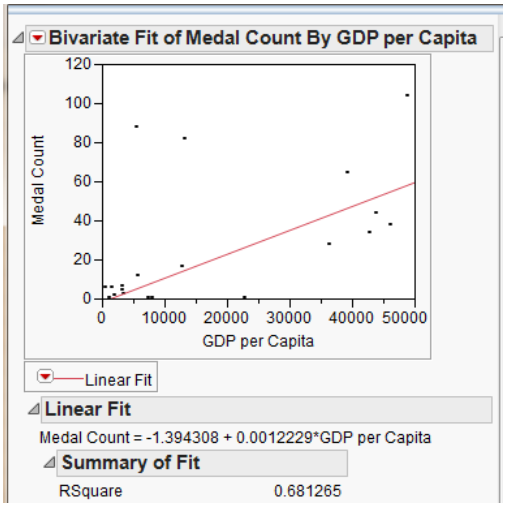
There are many different answers that could have been accepted. Mainly we were looking for a story with graphs and statistics that were sound. Some of the relationships that may have been seen are:



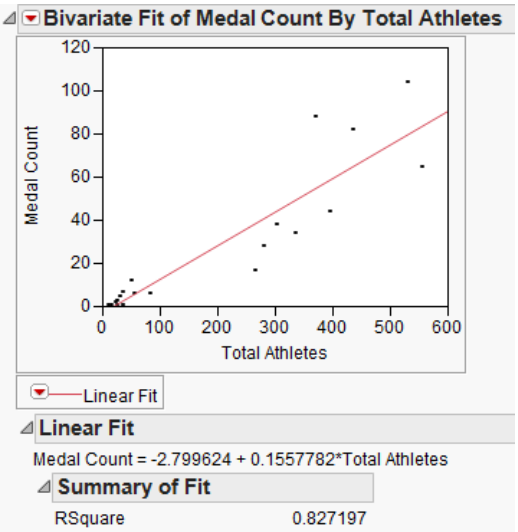
Correlation between GDP and Medal Count = .82



Correlation between GDP and Gold Medal Count = 0.83



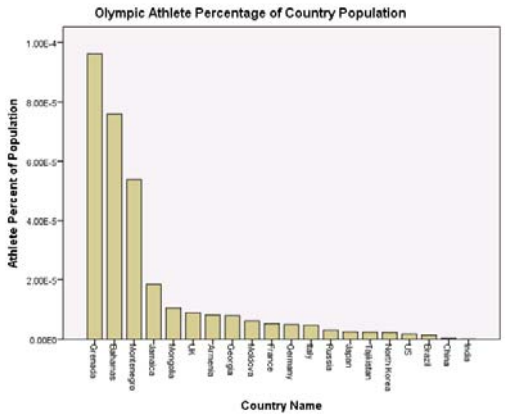
Linear fit between GDP /Capita and Medal Count with China and Russia removed from analysis (top left points)



Correlation between # of Athletes sent /Capita and Medal Count = 0.91

Things to keep in mind are that the variable that is to be predicted (dependent variable) should be on the y-axis and the dependent on the x-axis. In addition, items can still have a pattern but not be 'linearly correlated' or have low r-squared values. This is why graphs are important to look at.

Also, some interesting relationships existed with proportions of the number of Olympians /per population.



Problem 2: (**10 points – regression analysis**) In a study of genetic variation in sugar maple, seeds were collected from native trees in the eastern United States and Canada and planted in a nursery in Wooster, Ohio. The time of leafing out of these seedlings can be related to the latitude and mean July temperature of the place of origin of the seed. The variables are X1 = latitude, X2 = July mean temperature, and Y = weighted mean index of leafing out time. (Y is a measure of the degree to which the leafing out process has occurred. A high value is indicative that the leafing out process is well advanced.) *The data is in the file maple.txt* on the course web page under the documents for week 2.

(a) (2 points) Find the regression of LeafIndex on Latitude. Is latitude a useful predictor of leaf index?

(b) (2 points) Repeat part (a) for the regression of LeafIndex on JulyTemp.

(c) (6 points) Find the regression of LeafIndex on Latitude and JulyTemp. Compare the results of this analysis with your results from (a) and (b). How different are the slope coefficients in each case? What best explains the differences in their values?

Part a -

Regression Analysis: LeafIndex versus Latitude

The regression equation is
LeafIndex = - 1.67 + 0.454 Latitude

Predictor	Coef	SE Coef	T	P	VIF
Constant	-1.667	3.052	-0.55	0.589	
Latitude	0.45369	0.07427	6.11	0.000	1.000

S = 1.67282 R-Sq = 55.4% R-Sq(adj) = 53.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	104.41	104.41	37.31	0.000
Residual Error	30	83.95	2.80		
Total	31	188.36			

Unusual Observations

Obs	Latitude	LeafIndex	Fit	SE Fit	Residual	St Resid
30	31.4	15.900	12.570	0.766	3.330	2.24RX
32	30.8	10.800	12.284	0.810	-1.484	-1.01 X

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

The p-value is 0.000 therefore Latitude is a good predictor of Leaf Index. The R-squared is 55.4% (53.9% adjusted) and thus we are explaining over half of the variation of LeafIndex by using Latitude alone.

Part b –

Regression Analysis: LeafIndex versus JulyTemp

The regression equation is
LeafIndex = 40.7 - 0.333 JulyTemp

Predictor	Coef	SE Coef	T	P	VIF
Constant	40.743	4.455	9.15	0.000	
JulyTemp	-0.33318	0.06206	-5.37	0.000	1.000

S = 1.78949 R-Sq = 49.0% R-Sq(adj) = 47.3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	92.287	92.287	28.82	0.000
Residual Error	30	96.068	3.202		
Total	31	188.355			

Again, we have the p-value for July Temperature at 0.000 indicating it is also a good predictor. The R-square is 49% (47.3% adjusted). The ANOVA indicates the model is significant. We are explaining almost half of the variation in LeafIndex with JulyTemp.

Part c

Regression Analysis: LeafIndex versus JulyTemp, Latitude

The regression equation is
LeafIndex = 13.7 - 0.135 JulyTemp + 0.314 Latitude

w

Predictor	Coef	SE Coef	T	P	VIF
Constant	13.73	11.42	1.20	0.239	
JulyTemp	-0.13524	0.09676	-1.40	0.173	2.870
Latitude	0.3139	0.1239	2.53	0.017	2.870

S = 1.64685 R-Sq = 58.2% R-Sq(adj) = 55.4%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	109.703	54.852	20.22	0.000
Residual Error	29	78.652	2.712		
Total	31	188.355			

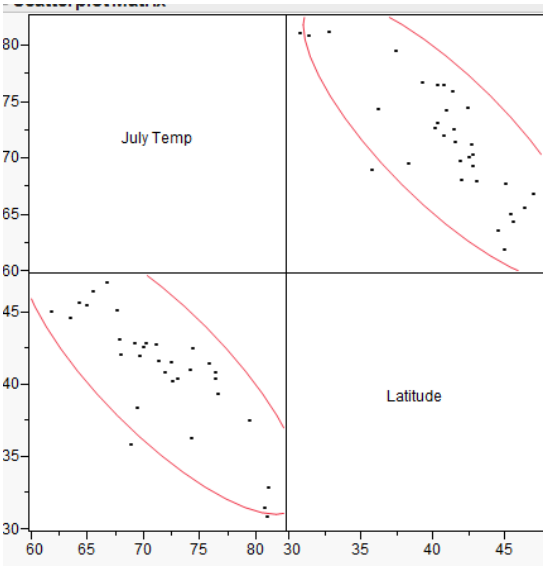
Source	DF	Seq SS
JulyTemp	1	92.287
Latitude	1	17.417

Unusual Observations

Obs	JulyTemp	LeafIndex	Fit	SE Fit	Residual	St Resid
29	68.9	15.000	15.652	0.905	-0.652	-0.47 X
30	80.8	15.900	12.655	0.757	3.245	2.22R

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

By looking at the relationship between our two factors, there is a correlation of -0.81. This is a fairly significant correlation between these two variables. Further investigation indicates that the best model would be to eliminate the JulyTemp variable as the P-value is 0.173 (above 0.05). The correlation between our variables is what is explaining the changing in the slope.



Problem 3: (**30 points – regression analysis**) The data in the file *chicinsur.txt* are collected from 47 zip-code areas in the Illinois area. There are 8 columns in the data file but not all are relevant here. The response variable of interest is the number of new home insurance policies (NEWPOL) (minus canceled policies) per 100 housing units. The predictor variables are the percent minority population living in the area (PCTMINOR), the number of fires per 1000 housing units (FIRES), the number of thefts per 1000 in population (THEFTS), the percent of housing units built before 1940 (PCTOLD), and the median income (INCOME). We are interested in which predictors are significant predictors of insurance policies issued.

(a) (5 points) Before running any regressions make a prediction as to what the sign of the coefficient of each predictor should be expected to be. (5 points) Obtain the correlation matrix for the variables PCT-MINOR FIRES THEFTS PCTOLD INCOME NEWPOL. Do the simple correlations support your predictions about the signs?

(b) Run a multiple regression of NEWPOL on the variables listed above.

i. (5 points) Comment on the overall significance of the regression fit.

ii. (5 points) Which predictors have coefficients that are significantly different from zero at the .05 level?

iii. (5 points) Do any of the predictors have signs that are different than suggested by their simple correlations? If so, explain what may be happening. If not, explain how such a thing can happen.

iv. (5 points) Examine a plot of residuals versus predicted values. Do you see any problems?

Part a)

Predictions could be any predictions that you may have about the variables.

Correlation matrix of variables is shown below:

	pctminor	**fires**	**thefts**	**pctold**	**income**	**newpol**
pctminor	1.0000	0.5928	0.2551	0.2505	-0.7037	-0.7594
fires	0.5928	1.0000	0.5562	0.4122	-0.6104	-0.6865
thefts	0.2551	0.5562	1.0000	0.3176	-0.1729	-0.3116
pctold	0.2505	0.4122	0.3176	1.0000	-0.5287	-0.6057
income	-0.7037	-0.6104	-0.1729	-0.5287	1.0000	0.7510
newpol	-0.7594	-0.6865	-0.3116	-0.6057	0.7510	1.0000

Part b-

Regression Analysis: newpol versus pctmin, fires, thefts, pctold, income

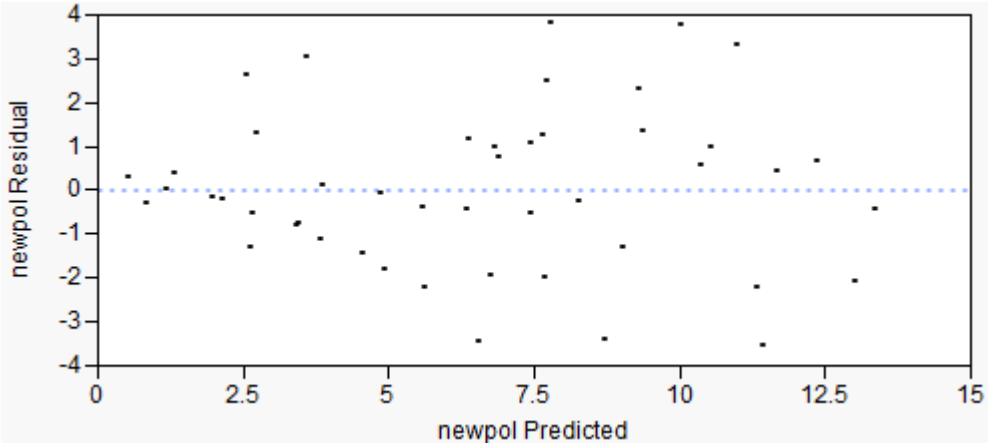
```
The regression equation is
newpol = 12.1 - 0.0595 pctmin - 0.102 fires + 0.0136 thefts - 0.0644 pctold
         + 0.000116 income


Predictor       Coef    SE Coef      T      P    VIF
Constant      12.061      2.819   4.28  0.000
pctmin      -0.05947    0.01318  -4.51  0.000  2.333
fires       -0.10185    0.04801  -2.12  0.040  2.522
thefts       0.01356    0.01624   0.84  0.408  1.656
pctold      -0.06437    0.01583  -4.07  0.000  1.615
income     0.0001164  0.0001804   0.64  0.523  3.121


S = 1.90738   R-Sq = 79.4%   R-Sq(adj) = 76.9%


Analysis of Variance

Source          DF      SS      MS      F      P
Regression       5  574.56  114.91  31.59  0.000
Residual Error  41  149.16    3.64
Total           46  723.72
Regression Equation:
```

i) The overall significance of the model can be seen in the ANOVA table. Here we see a p-value of 0.000. This value is under .05 alpha value we set. Therefore, we conclude that the overall model is significant. The r-sq is 79.4% (adjusted r-sq is 76.9%) which indicates that about 79% of the variation in newpol can be explained by our model.

ii) By looking at the p-value for each predictor we can conclude that ***pctminor, fires, and pct old*** are statistically significant as their p-value is less than .05.

iii) We can see that thefts has a negative correlation with newpol but is positive in our regression equation. This may be of a concern but we learned in part ii) that it is not statistically significant so it should be removed from the model. Remember, that if a factor is not statistically significant, it means that it is not statistically significant from 0. When this happens, switching signs could happen.

iv) The predicted vs. residual plot is shown below. It shows a random distribution around 0 and is not indicative of any problems. There may be a slight cone indicating that at low predictions we have a lower residual but as the newpol increases, our residual increases.



Problem 4 (5 points –regression application): Briefly describe an application for the multiple regression in a field of interest to you. Identify possible independent variables and the dependent variable for your application. If you read about the application from a research paper or news article, please provide a reference to it.

Many different answers were acceptable. As long as I saw the problem statement and a list of variables applicable, I gave full credit.