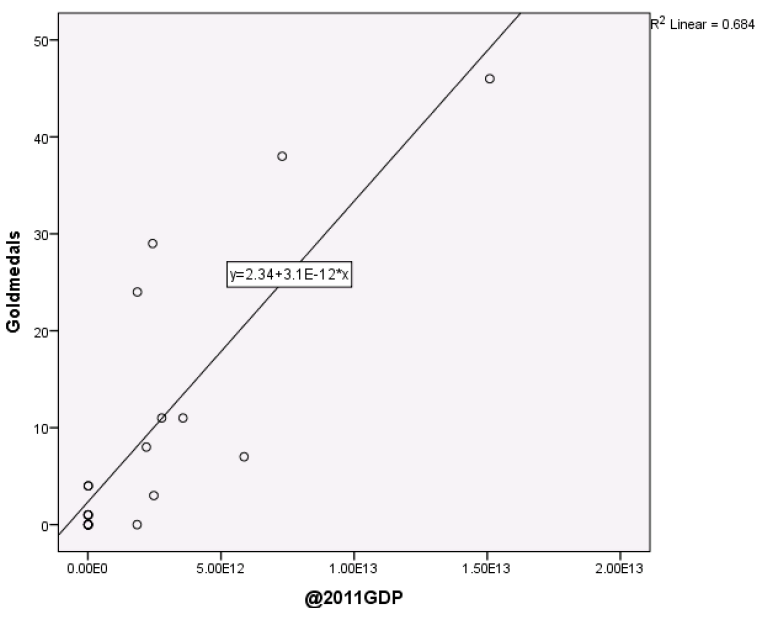
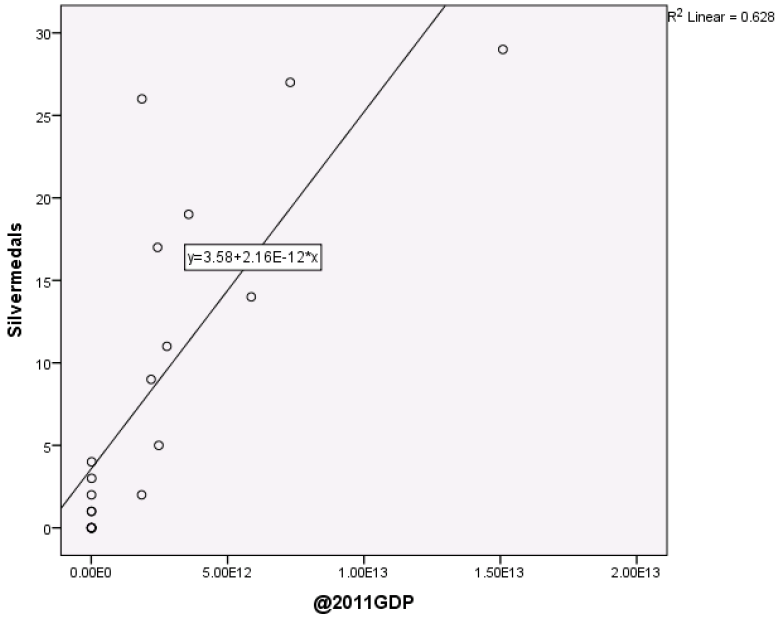
CSC424 Assignment 1
Sarah Cummings

Problem 1- Olympics data:

What are we competing for? The olympics have historically measured the strength and athleticism of different countries. Viewers believe that the countries that win the most medals are the strongest and most dedicated— something that inspires much patriotism. However, it appears as though this conclusion is not necessarily accurate. Beyond strength and athleticism, there is another potentially crucial factor determining a country's wins and losses: money.





As seen in the scatterplots above, there appears to be a positive relationship between a country's GDP and its medal counts. The countries with more money have more medals. The countries whose citizens can afford all of the training, traveling, and various expenses of being a professional athlete win more medals. This means that countries do not win the olympics for being stronger and more determines— they win because they have the money to do so.

There is one surprising country in this set: Russia. In the silver medal count scatterplot (right), it appears as though their silver medal count is very high given their GDP. Looking back at the data, we see they have 26 silver medals. This is close to the top silver medal winning countries (USA with 29 and China with 27), despite the fact that their GDP is much lower than that of USA and China.

Problem 2- Genetic variation of maple trees.

a) Regression of LeafIndex on Latitude. Is latitude a useful predictor of leaf index?

A simple linear regression of leafIndex on latitude is:
LeafIndex= -1.667 +0.454(Latitude)
The Latitude variable is a useful predictor of LeafIndex with its t stat, 6.108, having p value of 0.00. The model itself is useful with F stat 37.310 and again p value 0.00.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	104.406	1	104.406	37.310	.000[b]
	Residual	83.949	30	2.798		
	Total	188.355	31			

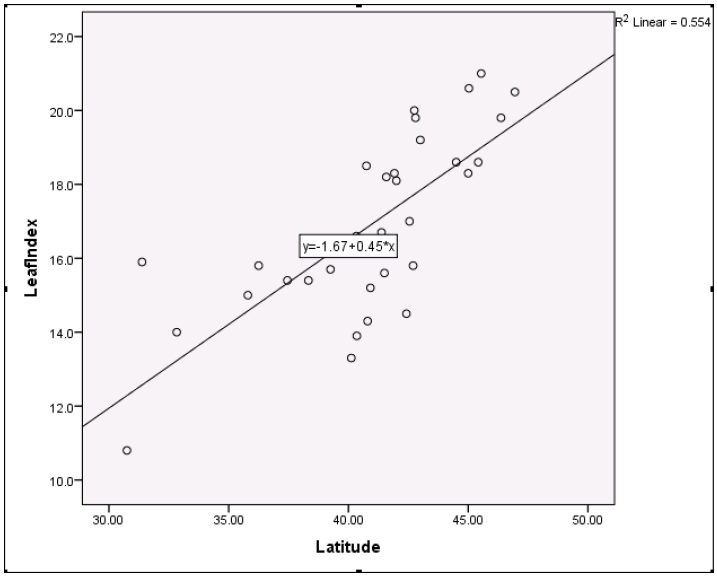
a. Dependent Variable: LeafIndex

b. Predictors: (Constant), Latitude

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1.667	3.052		-.546	.589
	Latitude	.454	.074	.745	6.108	.000

a. Dependent Variable: LeafIndex



b) What about LeafIndex on July temp?

A simple linear regression of leafIndex on JulyTemp is:
LeafIndex= 40.743 -0.333(JulyTemp)
The JulyTemp variable is a useful predictor of LeafIndex with its t stat, 9.145, having p value of 0.00. The model itself is useful with F stat 28.819 having p value 0.00.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	92.287	1	92.287	28.819	.000[b]
	Residual	96.068	30	3.202		
	Total	188.355	31			

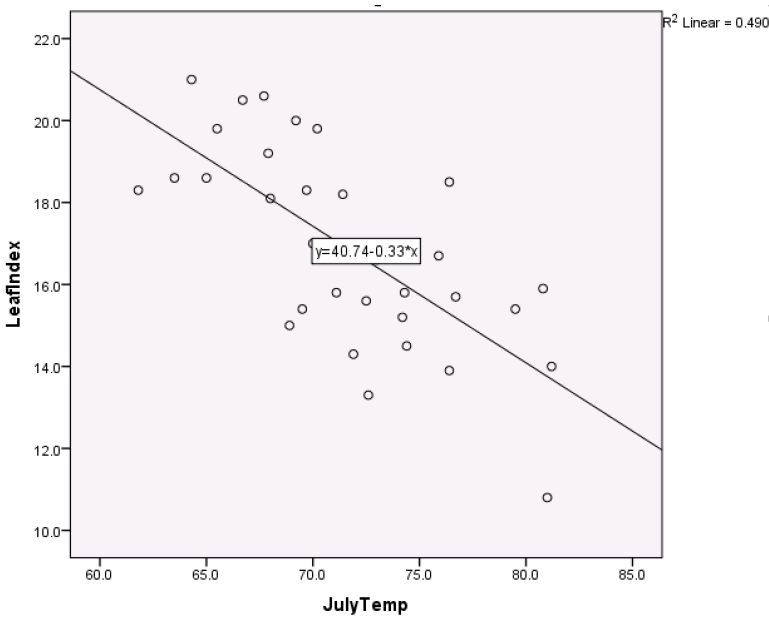
a. Dependent Variable: LeafIndex

b. Predictors: (Constant), JulyTemp

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	40.743	4.455		9.145	.000
	JulyTemp	-.333	.062	-.700	-5.368	.000

a. Dependent Variable: LeafIndex



c) What about both Latitude and July temp on LeafIndex?

A linear regression for Latitude and July temp on LeafIndex is:
LeafIndex= 13.732 -0.135(JulyTemp) +0.314(Latitude)
The overall model is useful, with F stat 20.25 having p=0.00. The individual coefficients, however, are not both significant predictors. The JulyTemp t-stat -1.398 is not significant with p=.173. The Latitude t-stat 2.534 is significant with 0.017 .

In comparing these three models, we see that the JulyTemp variable consistently has a negative coefficient and the Latitude variable generally has a positive coefficient. The slopes in our third model, however, are smaller (in absolute value). In the first model, the Latitude coefficient is 0.454 and in the third, the Latitude coefficient is 0.314. In the first model, the JulyTemp coefficient is -0.333 and in the third, the coefficient is -0.135. These coefficient do not necessarily indicate the relationship between the variables is weaker than we thought, but rather that they are adjusting to the fact that we now have two independent variables rather than one.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	109.703	2	54.852	20.225	.000[b]
	Residual	78.652	29	2.712		
	Total	188.355	31			

a. Dependent Variable: LeafIndex

b. Predictors: (Constant), Latitude, JulyTemp

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	13.732	11.420		1.202	.239
	JulyTemp	-.135	.097	-.284	-1.398	.173
	Latitude	.314	.124	.515	2.534	.017

a. Dependent Variable: LeafIndex

Problem 3- significant Predictors of insurance policies issues

a) Predicted signs of independent variables:

(PCTMINOR): Negative
(FIRES): Negative
(THEFTS):Negative
(PCTOLD):Negative
(INCOME): Positive

Correlation matrix:

Correlations

		pctminor	fire	thefts	pctold	newpol	fairpol	income
pctminor	Pearson Correlation	1	.593**	.255	.251	-.759**	.714**	-.704**
	Sig. (2-tailed)		.000	.084	.089	.000	.000	.000
	N	47	47	47	47	47	47	47
fire	Pearson Correlation	.593**	1	.556**	.412**	-.686**	.703**	-.610**
	Sig. (2-tailed)	.000		.000	.004	.000	.000	.000
	N	47	47	47	47	47	47	47
thefts	Pearson Correlation	.255	.556**	1	.318*	-.312*	.150	-.173
	Sig. (2-tailed)	.084	.000		.030	.033	.315	.245
	N	47	47	47	47	47	47	47
pctold	Pearson Correlation	.251	.412**	.318*	1	-.606**	.476**	-.529**
	Sig. (2-tailed)	.089	.004	.030		.000	.001	.000
	N	47	47	47	47	47	47	47
newpol	Pearson Correlation	-.759**	-.686**	-.312*	-.606**	1	-.746**	.751**
	Sig. (2-tailed)	.000	.000	.033	.000		.000	.000
	N	47	47	47	47	47	47	47
fairpol	Pearson Correlation	.714**	.703**	.150	.476**	-.746**	1	-.665**
	Sig. (2-tailed)	.000	.000	.315	.001	.000		.000
	N	47	47	47	47	47	47	47
income	Pearson Correlation	-.704**	-.610**	-.173	-.529**	.751**	-.665**	1
	Sig. (2-tailed)	.000	.000	.245	.000	.000	.000	
	N	47	47	47	47	47	47	47

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Focusing on the fifth row, we see that the simple correlations support my predictions, and these correlations are all significant at either the 0.01 level or better.

b) Multiple Regression with independent variables.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.891[a]	.794	.769	1.9074

a. Predictors: (Constant), pctold, pctminor, thefts, fire, income

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	574.557	5	114.911	31.586	.000[b]
	Residual	149.162	41	3.638		
	Total	723.718	46			

a. Dependent Variable: newpol

b. Predictors: (Constant), pctold, pctminor, thefts, fire, income

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	12.061	2.819		4.279	.000
	fire	-.102	.048	-.239	-2.121	.040
	pctminor	-.059	.013	-.489	-4.512	.000
	thefts	.014	.016	.076	.835	.408
	income	.000	.000	.081	.645	.523
	pctold	-.064	.016	-.366	-4.066	.000

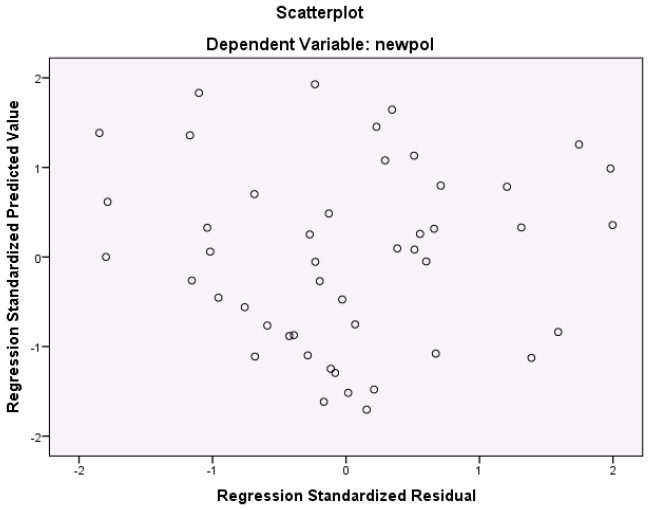
a. Dependent Variable: newpol

i. Overall, the model is significant with F stat, 32.586, having a p-value of 0.00. The model also have an adjusted r-sqared of 0.769 which means that the model accounts for 76.9% of the variability in the data.

ii. The following predictors are significantly different from zero at the .05 level: fire, pctminor, and pctold.

iii. Both thefts and income have signs that are different than suggested from their simple correlations— thefts is positive when expected to be negative and income is 0.00 when expected to be positive. This might have happened because thefts is correlated with another independent variable. Looking back at our correlation matrix, we see thefts and fire having a significant correlation of 0.556. Thus, thefts wasn't a crucial variable in our model as the same variability was explained by the fires variable. I am not sure why the coefficient for income is 0.00 when income has such a strong positive correlation with newpol.

iv. Residuals v fits



Looking at the residual v fit plot (left), it appears to fir our requirements. There is no real pattern to the residuals, they aren't in the dreaded cone shape, they seem ok.

Problem 4- Application of multiple linear regression with possible independent, dependent variables, reference.

Multiple linear regression can be used in environmental science and renewable energy research, which is really interesting to me. I spent a summer at the Pacific Northwest National Laboratory with a team who was wording on converting biomass (algae, wood from trees, etc) to mixed alcohols for use in liquid transportation fuels. In the conversion process, our goal was to discover how factors such as temperature, pressure, space velocity and H_2:CO ratios affect the converted carbon selectivity to various products. In general, we want to minimize the converted carbon selectivity to methane and other light hydrocarbons and maximize the converted carbon selectivity to useful products such as the C_2+ oxygenates. In doing this, we were able to optimize the amount of useful products from the conversions.

We had data from many runs of the experiment, and we ran multiple linear regressions with temperature, pressure, space velocity and H_2:CO ratios as the independent variables and converted carbon selectivity to C_2+ oxygenates as the dependent variable.

Here's an article the team recently published:
http://www.sciencedirect.com/science/article/pii/S0960852414013911