STAT 512 Final Project

Group 1

Natalie A Ehmke, M.Reza Moini, Upasana Angara, Yunmei Bai

Introduction

In the dataset, 62 observations are provided with the Olympic Medals numbers in both 1992 Barcelona Summer Olympics Games and 1994 Norway Winter Olympic Games. Also, the data states the latitude and the population for each observed country. Based on the dataset, some controversial arguments are over the relationship between the two explanatory variables, namely latitude and population, and the explained variable medal numbers.

One implication is that larger countries are expected to gain more Olympic medals than smaller countries. Holding this perspective, some viewers believe that it is more reasonable to standardize the total medals numbers with respect to the range of population and to compare the per capita number of medals; Another implication would be that countries further from equator are expected to have better performance in winter olympic games.

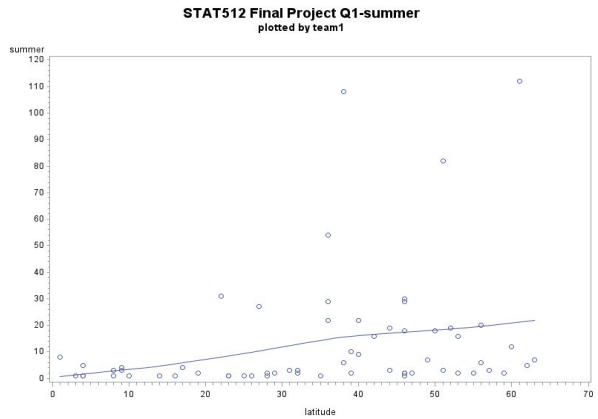
The purpose of this project is to figure out the firm relationship among the three variables and to justify whether the above statements hold credits.

The limitation of this study is that there are some missing data for medals of winter olympic games, which could possibly lead to inaccurate interpretation and requires manipulation of original dataset.

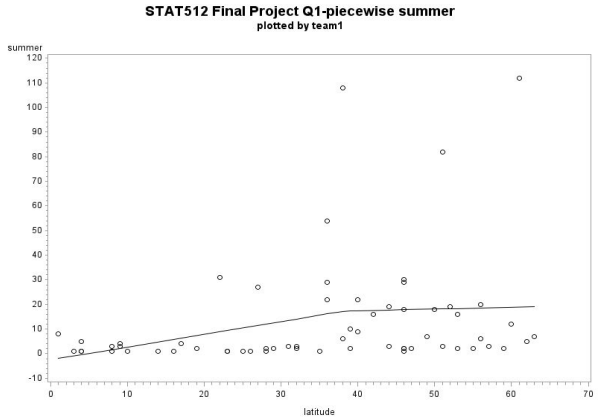
PART I

Question 1

In this question, we first decide to choose *'latitude'* as predictor to perform piecewise SLR for both summer and winter olympic games.

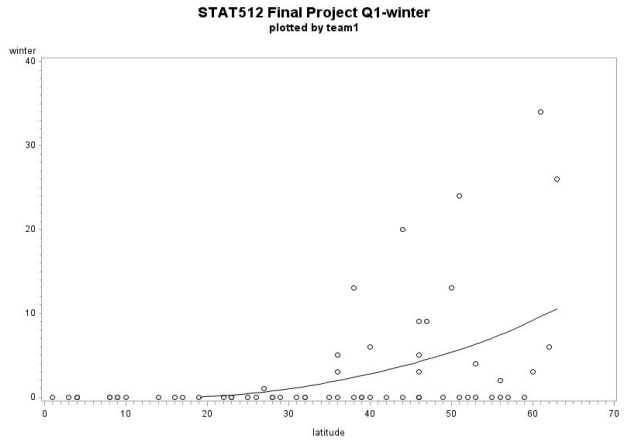


Graph 1

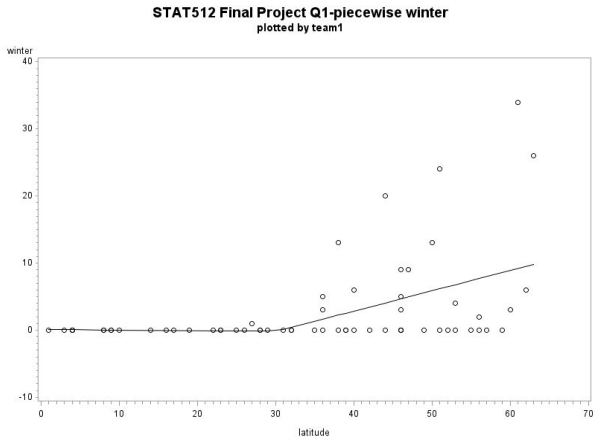


Graph 2

For summer olympic games, we first plot the scatterplot with smoothing line 70 between the summer medals and the corresponding country's population. From the Graph 1 above, we can apparently see that the slope shows an upward trend till the point of latitude 38 and then the slope demonstrate a more flatten trend. Consequently, we choose to use latitude 38 as the piecewise point. From the scatterplot of graph 2 above, we can see that the slope coincide with latitude 38.



Graph 3



Graph 4

For winter olympic games, we first plot the scatterplot with smoothing line 70 between the winter medals and the corresponding country's population. From the Graph 3, the slope starts to increase at the latitude around 30. In order to test for the most accurate piecewise latitude point, we tried a couple of latitude ranging from 28 to 32 and spotted latitude 30 has the most accurate matching degree.

STAT512 Final Project Q1-piecewise summer

plotted by team1

The REG Procedure
Model: MODEL1
Dependent Variable: summer

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2681.89810	1340.94905	2.75	0.0724
Error	59	28816	488.39867		

Corrected Total	61	31497

Root MSE	22.09974	R-Square	0.0851
Dependent Mean	13.09677	Adj R-Sq	0.0541
Coeff Var	168.74188		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-2.50723	7.62919	-0.33	0.7436
latitude	1	0.51954	0.27930	1.86	0.0679
cslope	1	-0.44555	0.63343	-0.70	0.4846

STAT512 Final Project Q1-piecewise winter

plotted by team1

The REG Procedure
Model: MODEL1
Dependent Variable: winter

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	621.59260	310.79630	7.98	0.0009
Error	59	2298.40740	38.95606		
Corrected Total	61	2920.00000			

Root MSE	6.24148	R-Square	0.2129

Dependent Mean	3.00000	Adj R-Sq	0.1862
Coeff Var	208.04929		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.09700	2.38617	0.04	0.9677
latitude	1	-0.00922	0.10822	-0.09	0.9324
cslope	1	0.31102	0.17773	1.75	0.0853

From the regression output from SAS, we can see both summer and winter models have p-values for '*cslope*' over 0.05. In conclusion, the piecewise for both summer and winter are not significantly different.

Question 2

In this part, we create a new variable SUM which is the summation of two of the explanatory variables. Since we only have two predictors in our case i.e Latitude and population, we sum these two. SUM = Latitude + Population. Since we do not have other explanatory variables apart from these two, we will include these two for the following exercises.

a. i. Model ran with 2 variables population and latitude, each for winter and summer medals as follows. The predicted response for summer and winter is:

STAT512 Final Project Q2ai-Summer
plotted by team1

The REG Procedure
Model: MODEL1
Dependent Variable: summer

Number of Observations Read	62
Number of Observations Used	62

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	8659.88491	4329.94245	11.19	<.0001
Error	59	22838	387.07685		
Corrected Total	61	31497			

Root MSE	19.67427	R-Square	0.2749
Dependent Mean	13.09677	Adj R-Sq	0.2504
Coeff Var	150.22225		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS
Intercept	1	-4.35774	5.72234	-0.76	0.4494	10635	224.47776
population	1	0.06478	0.01616	4.01	0.0002	5807.62613	6219.63153
latitude	1	0.38838	0.14307	2.71	0.0087	2852.25878	2852.25878

STAT512 Final Project Q2ai-Winter
plotted by team1

The REG Procedure
Model: MODEL1
Dependent Variable: winter

Number of Observations Read	62
Number of Observations Used	62

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	572.33689	286.16844	7.19	0.0016
Error	59	2347.66311	39.79090		
Corrected Total	61	2920.00000			

Root MSE	6.30800	R-Square	0.1960
Dependent Mean	3.00000	Adj R-Sq	0.1688
Coeff Var	210.26676		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS
Intercept	1	-3.21654	1.83471	-1.75	0.0848	558.00000	122.30010
population	1	0.00688	0.00518	1.33	0.1897	51.84968	70.04807
latitude	1	0.16591	0.04587	3.62	0.0006	520.48720	520.48720

$$Y\,summer = -4.357 + 0.064 * Population + 0.388 * Latitude$$

$$Y\,winter = -3.216 + 0.00588 * Population + 0.16591 * Latitude$$

a. ii We only have 2 variables population and latitude so we cannot run the model without them along with sum. For both summer and winter, the model is not full rank. Least-squares solutions for the parameters are not unique. Some statistics will be misleading. A reported DF of 0 or B means that the estimate is biased.

STAT512 Final Project Q2aii-Summer
plotted by team1

The REG Procedure
Model: MODEL1
Dependent Variable: summer

Number of Observations Read	62
Number of Observations Used	62

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	8659.88491	4329.94245	11.19	<.0001
Error	59	22838	387.07685		
Corrected Total	61	31497			

Root MSE	19.67427	R-Square	0.2749
Dependent Mean	13.09677	Adj R-Sq	0.2504
Coeff Var	150.22225		

STAT512 Final Project Q2aii-Winter
plotted by team1

The REG Procedure
Model: MODEL1
Dependent Variable: winter

Number of Observations Read	62
Number of Observations Used	62

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	572.33689	286.16844	7.19	0.0016
Error	59	2347.66311	39.79090		
Corrected Total	61	2920.00000			

Root MSE	6.30800	R-Square	0.1960
Dependent Mean	3.00000	Adj R-Sq	0.1688
Coeff Var	210.26676		

latitude = sum - population

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS
Intercept	1	-4.35774	5.72234	-0.76	0.4494	10635	224.47776
sum	B	0.38838	0.14307	2.71	0.0087	6681.95510	2852.25878
population	B	-0.32360	0.14315	-2.26	0.0275	1977.92981	1977.92981
latitude	0	0	.	.	.	.	.

latitude = sum - population

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS
Intercept	1	-3.21654	1.83471	-1.75	0.0848	558.00000	122.30010
sum	B	0.16591	0.04587	3.62	0.0006	94.61375	520.48720
population	B	-0.15903	0.04590	-3.46	0.0010	477.72314	477.72314
latitude	0	0	.	.	.	.	.

b.

STAT 512 Final Project 2b-Summer
plotted by team1

The REG Procedure
Model: MODEL1

Test test1 Results for Dependent Variable summer				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	2852.25878	7.37	0.0087
Denominator	59	387.07685		

STAT 512 Final Project 2b-Winter
plotted by team1

The REG Procedure
Model: MODEL1

Test test1 Results for Dependent Variable winter				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	520.48720	13.08	0.0006
Denominator	59	39.79090		

For this part, we calculate extra sum of squares for explanatory variables:Population and Latitude along with SUM.

The degrees of freedom (DFn, DFd) for the reduced models Summer are (1,59) and for winter are also (1,59).
The hypothesis is
H0 : Slope(given by coefficient of SUM)= 0
Ha : Slope $\neq 0$
The p value for summer is $0.0087 < 0.05(\alpha)$ there the p-value is significant and we can reject the null hypothesis i,e the coefficient of slope is not zero i.e there is a linear relationship between summer medals and SUM.
The case for winter is also similar i.e p-value is significant , hence we reject the null hypothesis. There is a linear relationship between winter medals and SUM.

C. Comparing the individual t-test from the full model and F statistic from the reduced model, we find that the relationship between the two follow as
$F \sim F_{n-p,p-1} = (t^*)^2$

For summer Medals: (2.71)^2 = 7.37
For winter Medals: (3.62)^2 = 13.08

Question 3

STAT 512 Final Project 2b-Summer
plotted by team1

The REG Procedure
Model: MODEL1
Dependent Variable: summer

Number of Observations Read	62
Number of Observations Used	62

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	8659.88491	4329.94245	11.19	<.0001
Error	59	22838	387.07685		
Corrected Total	61	31497			

Root MSE	19.67427	R-Square	0.2749
Dependent Mean	13.09677	Adj R-Sq	0.2504
Coeff Var	150.22225		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS
Intercept	1	-4.35774	5.72234	-0.76	0.4494	10635	224.47776
population	1	0.06478	0.01616	4.01	0.0002	5807.62613	6219.63153
latitude	1	0.38838	0.14307	2.71	0.0087	2852.25878	2852.25878

STAT 512 Final Project 2b-Winter
plotted by team1

The REG Procedure
Model: MODEL1
Dependent Variable: winter

Number of Observations Read	62
Number of Observations Used	62

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	572.33689	286.16844	7.19	0.0016
Error	59	2347.66311	39.79090		
Corrected Total	61	2920.00000			

Root MSE	6.30800	R-Square	0.1960
Dependent Mean	3.00000	Adj R-Sq	0.1688
Coeff Var	210.26676		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS
Intercept	1	-3.21654	1.83471	-1.75	0.0848	558.00000	122.30010
population	1	0.00688	0.00518	1.33	0.1897	51.84968	70.04807
latitude	1	0.16591	0.04587	3.62	0.0006	520.48720	520.48720

The sum of Type I error for population and latitude adds up to SSM for both summer and winter.
Extra sum of squares as per our model will be
SSM(sum | population latitude) = SSE(population latitude) - SSE(sum population latitude).
The given two SSE are equal so we cannot use this method to calculate the F statistic.

The Values of Type I and type II SS for summer and winter as follow:
Sum of type I errors for summer= 5807.62+2852.25 = 8569.87 = SSM for summer
Sum of type I errors for winter = 51.84+520.48=572.32 = SSM for Winter
They add up to their model sum of squares (SSM).

Question 4

Summer

Predictor #s	R^2	P-value	Intercept	Population	Latitude	Sum	MSE
2	0.2749	<0.0001	-4.358	0.0648	0.3884		387.077
2	0.2749	<0.0001	-4.358	-0.3236		0.3883	387.077
2	0.2749	<0.0001	-4.358		0.3236	0.0648	387.077
1	0.1844	0.0005	9.371	0.0625			428.163
1	0.0078	0.0285	0.5403		0.3588		484.286
1	0.2121	0.0002	6.756			0.0670	413.589

Winter

Predictor #s	R^2	P-value	Intercept	Population	Latitude	Sum	MSE
2	0.1960	0.0016	-3.217	0.0069	0.1659		39.791
2	0.1960	0.00016	-3.217	-0.1590		0.1659	39.791
2	0.1960	0.0016	-3.2165		0.0459	0.0069	39.791
1	0.0178	0.3018	2.6479	0.0059			47.803
1	0.1720	0.0008	-2.6967		0.1628		40.295
1	0.0324	0.1615	2.2455			0.0056	47.089

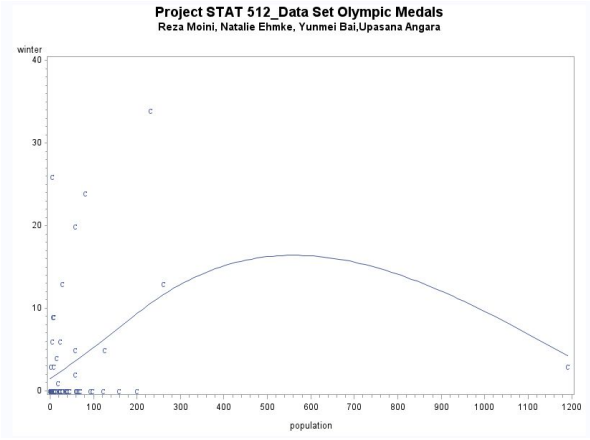
PART II

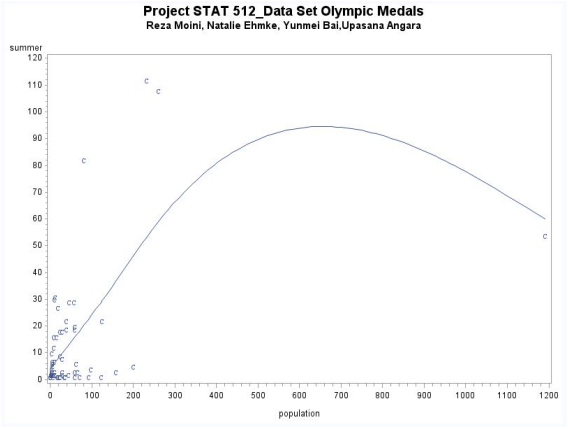
Question 1

Transformation and Addition of Variables:

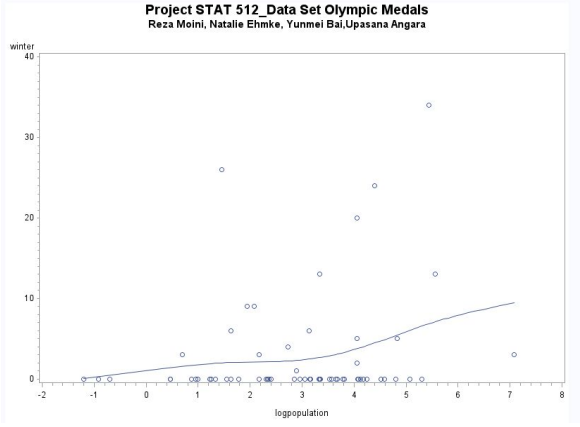
The initial model included two variables latitude and population. Looking at residual plot of latitude and population, latitude seem well distributed but residual plot for population seem not very well distributed and seems nonlinear. Therefore, a log transformation for the variable population is suggested and applied. Figure below shows the the scattered plots for summer and winter before and after log transformation of population:.

Before:





After:



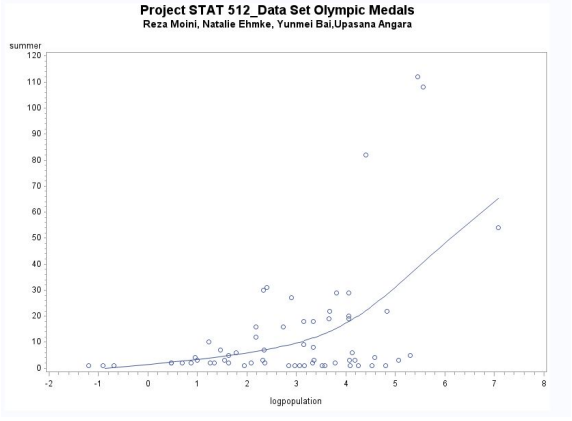
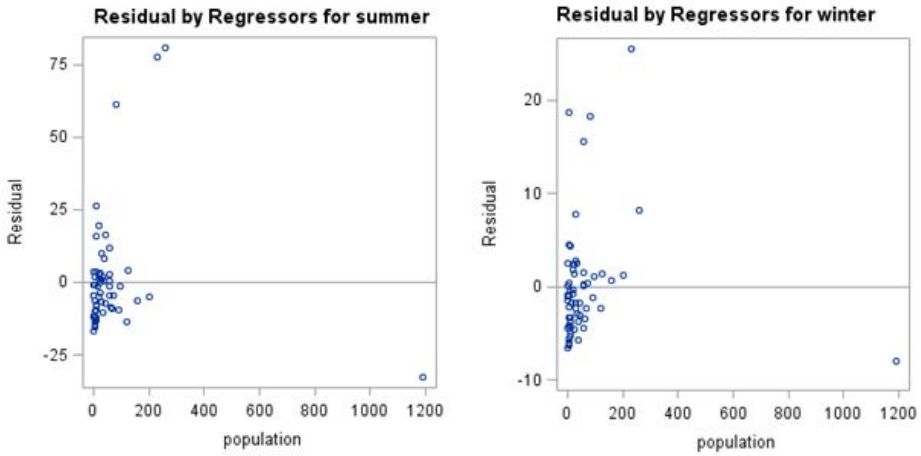
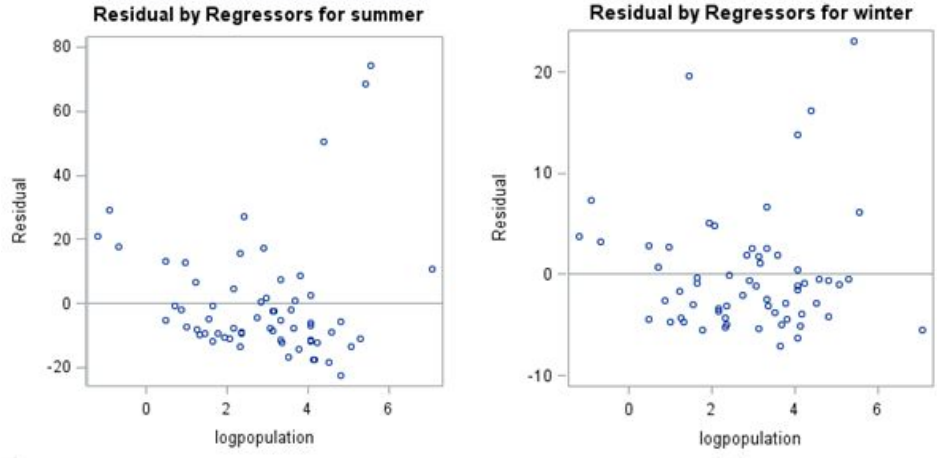


Figure below shows the the residual plots for summer and winter before and after log transformation of population. The latitude is not transformed:

Before:



After:



In addition to existing two continuous variables latitude and population, we created **One** dummy new **Binary** variables, **season**. to be able to have summer and winter medals both in one regression model.

Three new **Interaction** variables sealat, sealogpop, and latpop as follow and the data file is updated to reflect transformation of population and the following additions to the variables:

- **season** represents winter and summer olympic medals by introducing binary data (0=summer Olympics and 1=winter Olympics).
- **sealat** represents the interaction between the binary variable season and latitude.
- **sealogpop** represents the interaction between the binary variable season and log of population.
- **latpop** represents the interaction between.

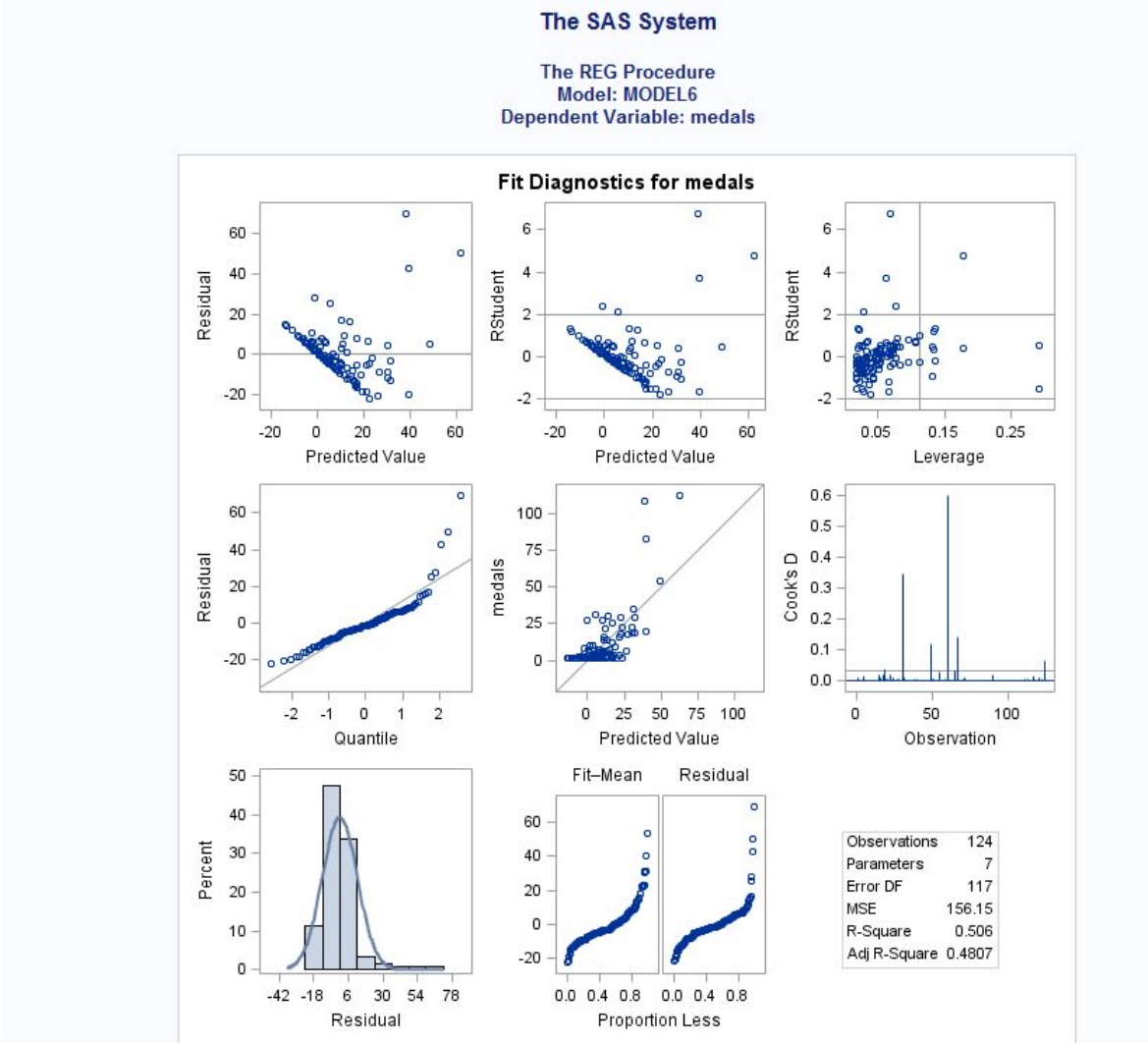
The season variable was added to have a single model for the summer and winter data and the interaction terms were added to look at interaction of variables. The table below represents the Adjusted R^2 for the initial model and the model with transformed population and subsequently added binary and interaction variables.

# of Variables	Model	R^2	Adjusted R^2
2	latitude population	0.1726	0.1589
2	latitude logpopulation	0.2211	0.2082
3	latitude logpopulation season	0.2904	0.2727
4	latitude logpopulation season latpop	0.4218	0.4024
5	latitude logpopulation season latpop sealat	0.4317	0.4076
6	latitude logpopulation season latpop sealat sealogpop	0.5060	0.4807

The addition of interaction and season variables doesn't hurt the model. The sixth model that includes all variable demonstrate highest Adjusted R^2 and so it is used prior to transformation of Y.

Transformation of Response - Y (Olympic Medals):

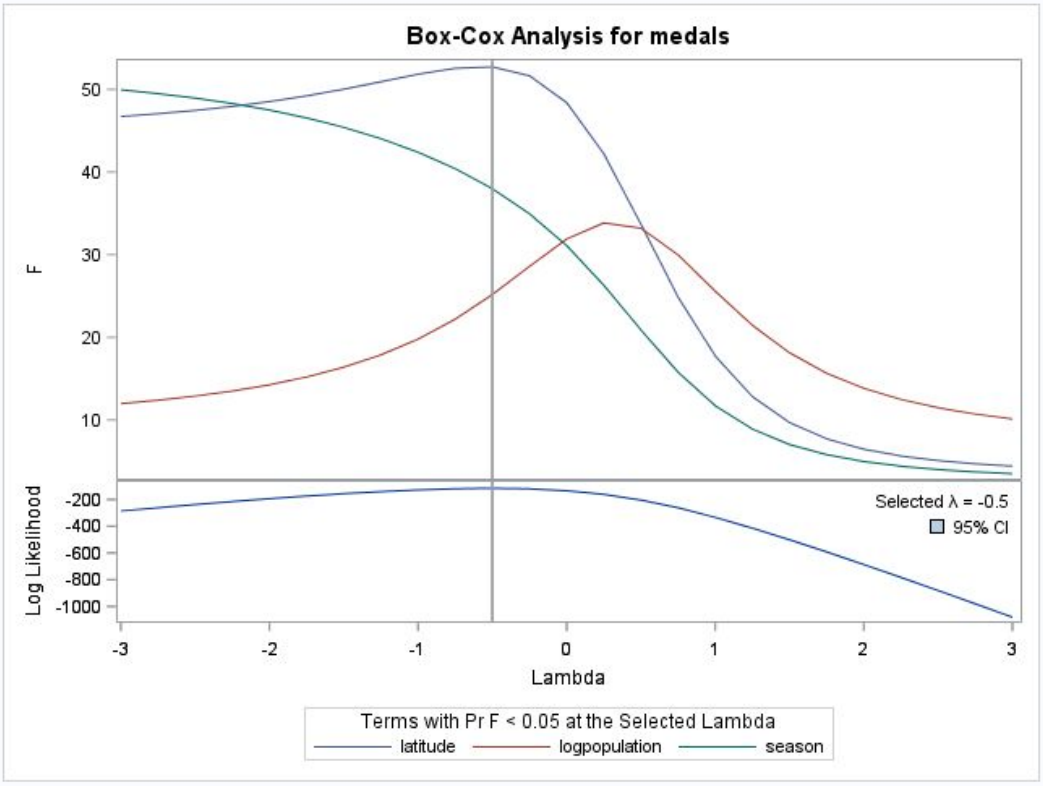
The 6th model (with variables latitude logpopulation season latpop sealat sealogpop) is taken and the normality of errors is checked using Q-Q plot is shown below. As demonstrated from Q-Q plot the errors don't seem normal. Therefore a transformation for Y (response) is considered.



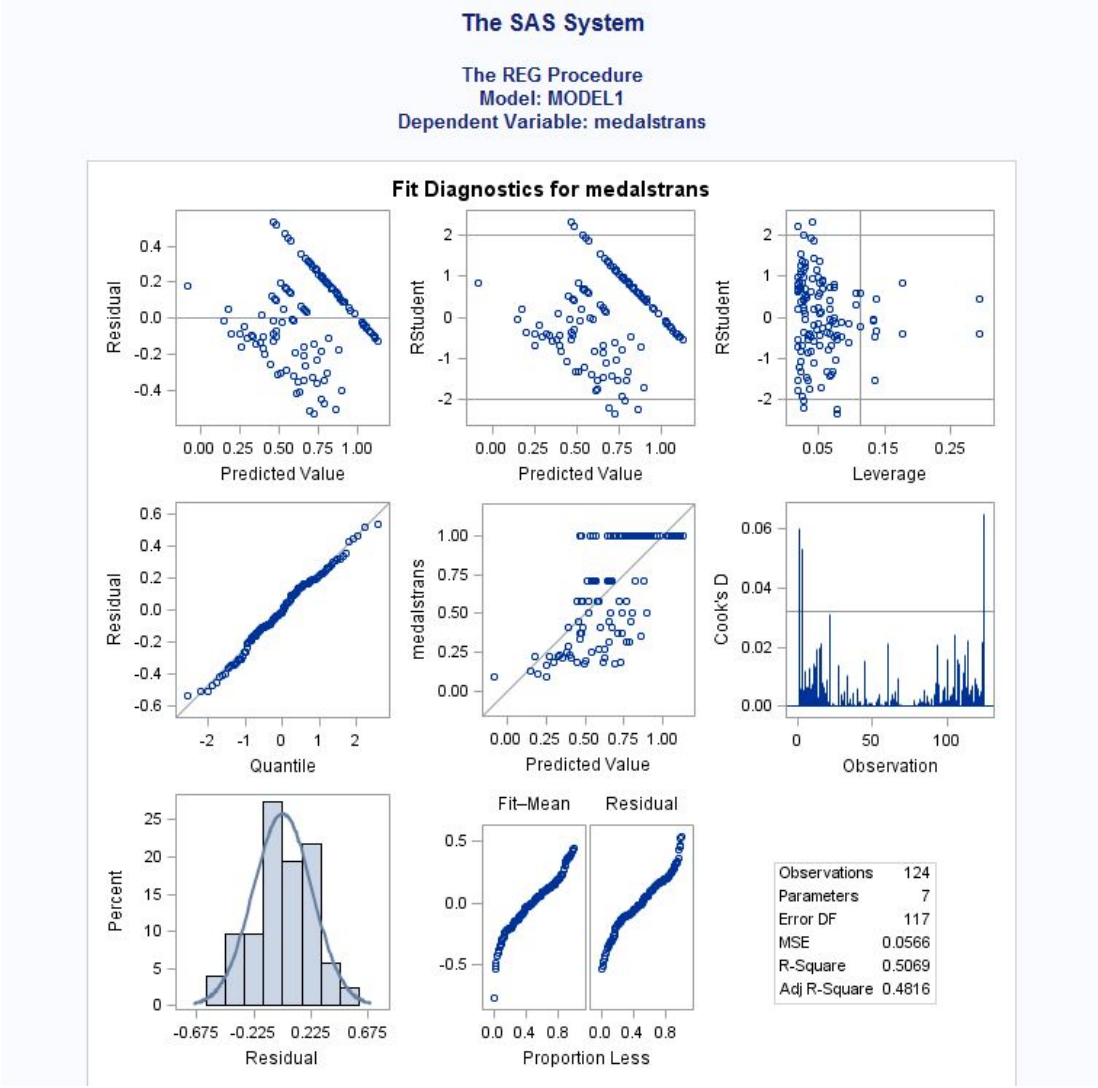
To determine the right transformation for Y, Box-Cox was used to estimate λ using latitude, population and season for variables (identity). The best lamda = -0.5 as shown in Figure below. This lambda is used in the model for the rest of the project.

The SAS System

The TRANSREG Procedure



Upon transformation of Y, the residuals appear to be normal as shown below:



Question 2

For model selection procedure , we first used the Mallow's Cp criterion to evaluate the best subset of the full model. SAS reported the below values based on our data set we got values as listed for the table below.

As per Mallow's Cp, a good model fulfills the condition $Cp \leq p$.

We have 6 predictors and for the models highlighted below the Cp is less than p for that model.

We found the models highlighted below best fulfill the condition. We know that a good model should have a high R2, high adjusted R2 and a Mallow's Cp value less than and closest to the number of predictors(+intercept). Based on this criteria we can pick the model

Model with variables Latitude logpopulation season and the interaction terms latitude*logpopulation, season*logpopulation and season*latitude.

Question 3

Based Selection Criterion, stepwise selection, SAS reported the model with predictors:latitude, season and latpop(interaction between latitude and log population) as the best subset and based on the F values of each predictors, all the predictors are significant. The model with highest R2 would be our best pick. For the subset of data with 3 variables: season, the model with the highest R2 contains all the 3 variables,latitude and latitude*population,.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	latpop		1	0.3153	0.3153	42.4567	56.17	<.0001
2	season		2	0.1682	0.4835	4.5516	39.40	<.0001
3	latitude		3	0.0201	0.5036	1.7771	4.86	0.0293

Based on our observations and models reported by SAS, we pick the model with 3 variables,latitude and latitude*population, The R2 values for this model is comparable to the R2 value of the full model and the Cp criterion also suggests it is a potentially good model without the complication of more variables.
Medals = 0.8688 - 0.00324*latitude + 0.26982*season - 0.00217*season*logpopulation

R-Square = 0.5036 and C(p) = 1.7771 and Adjusted R2 = 0.4912

Question 4

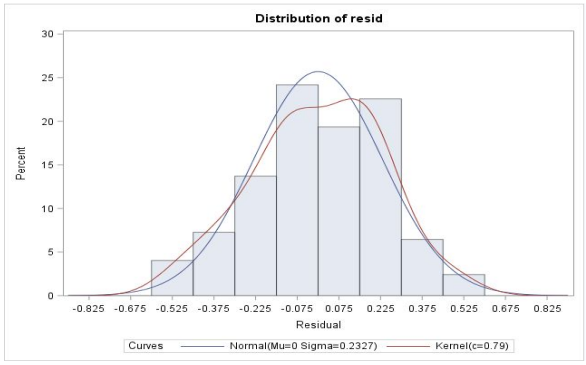
Based on the best regression model :
Medals = 0.8688 - 0.00324*latitude + 0.26982*season - 0.00217*season*logpopulation

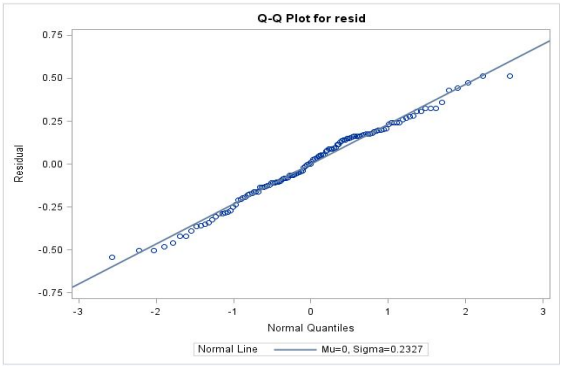
We have to check whether these model violated the four assumptions:

1) Normality Assumptions
2) Constant Variance
3) Linearity Assumptions
4) Independence Assumptions

<u>Normality Assumption</u>

Please find the below histogram graph and the residual quantile report (qqplot) for detail information:
From the histogram graph below, we can see the qqplot for residual shows apparent linearity, therefore, we can confirm the Normal Assumption





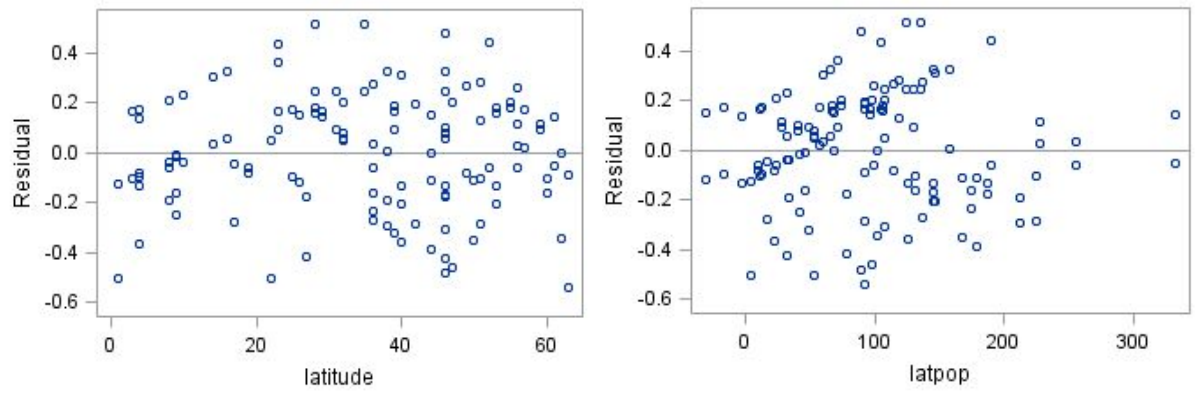
/*qqplot for residual*/
proc reg data=trans_new;
model medalstrans=latitude season latpop; output out = diagtrans_new r = residual;
run;

proc univariate data = diagtrans_new plot normal;
var resid;
histogram resid/ normal kernel(L=2);
qqplot resid / normal (L=1 mu=est sigma=est);
run;

<u>Constant Variance & Linearity Assumption:</u>

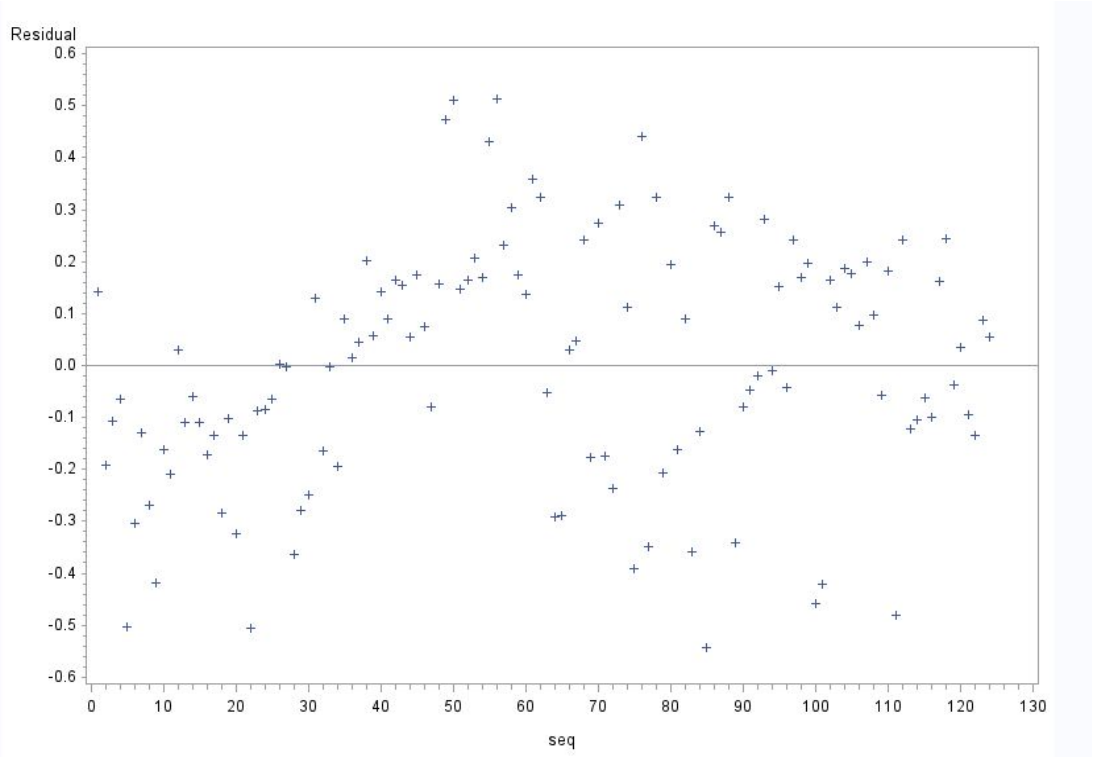
We use Residuals plot to check if the variances are constant and whether the linearity assumption holds for the linear term or not. From the graph, the residual plots for both logpopulation and latitude show an

approximately constant variance without any clear pattern (the scatter points are approximately random), which indicates that the linearity and constant variance assumption hold.



Independence Assumption:

We use the residual vs sequence plot to show if the residuals are distributed independently, and based on the assumption, we can conclude that there is no unusual pattern in the Residuals v.s. sequence plot, which implies that the residuals are independent.



Question 5

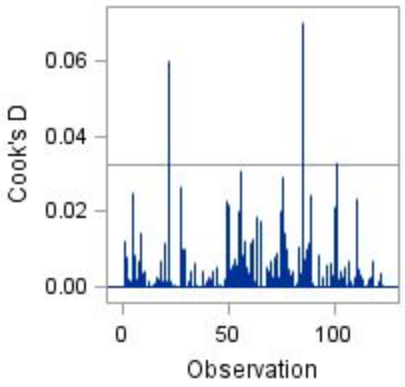
The following tests were performed to asses our data for potential outliers and influential data points.

Studentized Residual & Studentized Deleted Residual

Studentized residuals search data sets for outliers on the basis of difference between a given response for a specific value and the predicted response for the same value given the calculated model. While the actual cutoff to compare this difference can be calculated based off of the t distribution and the sample size, as a general rule, residuals larger than 3 are outliers. Using this general rule of $e_i < 3$, we found that there were no outliers.

Cook's Distance

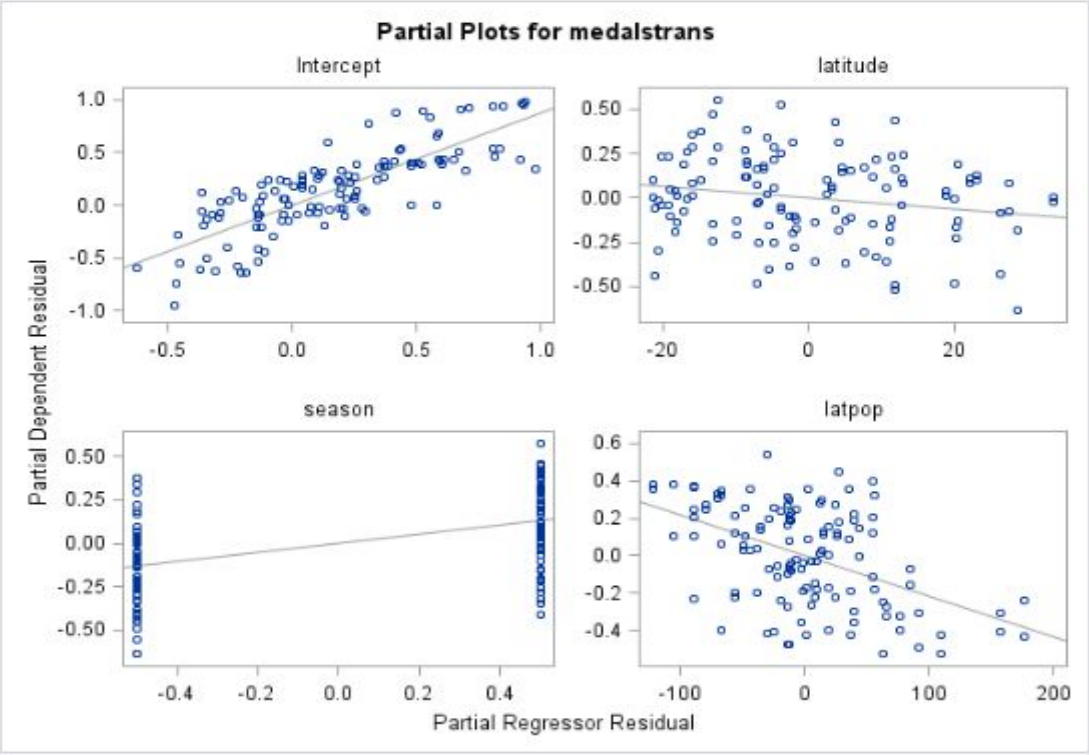
Cook's Distance is a measure to test the influence of data points and is a standardized version of the sum of squares of differences between predicted values with and without case "i." Large values indicate that the observation has high influence on the data. Larger values are defined as being larger than F(4,120) α=0.5, which here was found to be 0.844. The SAS graph below shows that there were three observations with high influence on the data.



Other tests to measure influence are the Hat Matrix Diagonals, where we compare the leverage of each individual case on the data. A large value for an observation indicates that that obersation is distant from the center of all other observations. Hat matrix values were compared to 0.0645 and those whose values were greater than this number were found to have high leverage on our data. Observations 28, 58, 60, 90, 120, 122 were found to have high influence in this method. Defits also measures influence and values larger than 1 are considered influential. Analysis of our Defits numbers showed no influential observations. DFBetas is a procedure to measure the influence of case i on each regression coefficient. Similar to Defits, values larger than 1 have high influence; our DFBetas showed no value larger than 1. The Variance Inflation Factor/Tolerance are both tests looking at the variance of the estimated regression coefficients. Since both VIF and tolerance are related by tolerance=1/VIF, a VIF equal to or greater than 10 and likewise a Tolerance greater than 0.1 indicate multicollinearity. The variance showed no multicollinearity.

Partial Regression Plots

We use the partial regression plots to determine whether each explanatory variable is useful and should be included in the model or not. From the SAS output, we can figure out that there is a linear relationship between the residual of Y and residual of X , we can make the conclusion that the variable is useful in the model.



Question 6

(a) Based on the previous SAS analysis, we can produce our best model as:
1/sqrt(medals)= 0.8688- 0.0032*latitude + 0.270season -0.0022(latitude*population)

(b) & (c)

90% confidence interval for the mean of the response variable and the 90% prediction interval for individual observation:

The table below presents the output of the first 10 observations which include the confidence mean of response variable & Individual observations

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	90% CL Mean		90% CL Predict		Residual
1	0.3536	0.8584	0.0511	0.7736	0.9432	0.4588	1.2581	-0.5049
2	1.0000	0.8360	0.0491	0.7545	0.9174	0.4370	1.2349	0.1640
3	0.4472	0.8100	0.0480	0.7304	0.8896	0.4114	1.2086	-0.3628
4	1.0000	0.8303	0.0482	0.7504	0.9102	0.4316	1.2289	0.1697
5	1.0000	0.8639	0.0491	0.7825	0.9452	0.4649	1.2628	0.1361
6	0.5774	0.7724	0.0443	0.6990	0.8459	0.3750	1.1698	-0.1951
7	1.0000	0.7937	0.0445	0.7199	0.8674	0.3962	1.1911	0.2063
8	0.5000	0.7504	0.0435	0.6783	0.8224	0.3532	1.1475	-0.2504
9	0.5774	0.7410	0.0435	0.6688	0.8132	0.3438	1.1381	-0.1636
10	1.0000	0.7680	0.0426	0.6973	0.8386	0.3711	1.1648	0.2320

(d) *90% confidence interval for the regression coefficients* prediction interval for individual observation:

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	90% Confidence Limits	
Intercept	1	0.86888	0.05216	16.66	<.0001	0.78241	0.95535
latitude	1	-0.00324	0.00147	-2.21	0.0293	-0.00567	-0.00080400
season	1	0.26982	0.04232	6.38	<.0001	0.19968	0.33997
latpop	1	-0.00217	0.00036403	-5.95	<.0001	-0.00277	-0.00156

Codes:

PART I

Q1:

```
data medals;
input country$ summer winter population latitude;
datalines;
UnifiedTeam    112    34    231.5    61
UnitesStates 108      13    260.7    38
Germany        82     24    81.1     51
```

China	54	3	1190.4	36
Cuba	31	0	11.1	22
Hungary	30	0	10.3	46
France	29	5	57.8	46
SouthKorea	29	0	45.1	36
Australia	27	1	18.1	27
Japan	22	5	125.1	36
Spain	22	0	39.3	40
Britain	20	2	58.1	56
Italy	19	20	58.1	44
Poland	19	0	38.7	52
Canada	18	13	28.1	50
Romania	18	0	23.2	46
Netherlands	16	4	15.4	53
Bulgaria	16	0	8.8	42
Sweden	12	3	8.8	60
NewZealand	10	0	3.4	39
NorthKorea	9	6	23.1	40
Kenya	8	0	28.2	1
Norway	7	26	4.3	63
Czechoslovakia	7	0	10.4	49
Denmark	6	0	5.9	56
Turkey	6	0	62.1	38
Finland	5	6	5.1	62
Indonesia	5	0	200.4	4
Jamaica	4	0	2.6	17
Nigeria	4	0	98.1	9
Belgium	3	0	10.1	51
Brazil	3	0	158.7	9
Croatia	3	0	4.7	44
Ethiopa	3	0	58.7	8
Iran	3	0	65.6	31
Latvia	3	0	2.7	57
Morocca	3	0	28.6	32
Austria	2	9	8	47
Slovenia	2	3	2	46
Algeria	2	0	27.9	29
Estonia	2	0	1.6	59
Greece	2	0	10.6	39
Ireland	2	0	3.5	53
Israel	2	0	5.1	32
Lithuania	2	0	3.8	55
Mongolia	2	0	2.4	46

```
Namibia         2       0       1.6     19
SouthAfrica     2       0       43.9    28
Switzerland     1       9       7       46
Argentina       1       0       33.9    35
Bahamas         1       0       0.3     26
Colombia        1       0       35.6    3
Ghana   1       0       17.2    8
Malaysia        1       0       19.3    4
Mexico 1        0       92.2    23
Pakistan        1       0       121.9   28
Peru    1       0       23.7    10
Philippines     1       0       69.8    14
Qatar   1       0       0.5     25
Suriname        1       0       0.4     4
Taiwan 1        0       21.3    23
Thailand        1       0       59.5    16
;
run;

proc print data = medals;
run;

symbol1 v=circle i=sm70;
proc sort data = medals;
by latitude;
proc gplot;
plot summer*latitude;
run;

symbol1 v=circle i=sm70 c=black;
proc sort data=medals;
by latitude;
proc gplot;
plot winter*latitude;
run;
/*Part 1, Question 1*/
data piecewise;
set medals;
if latitude le 38 then cslope=0;
if latitude gt 38 then cslope=latitude-38;
run;
proc print data=piecewise;
run;
```

```
proc reg data=piecewise;
model summer=latitude cslope;
output out=pieceout p=summerhat;
run;

symbol1 v=circle i=none c=black;
symbol2 v=none i=joint c=black;
proc sort data=pieceout;
by latitude;
proc gplot data=pieceout;
plot(summer summerhat)*latitude/overlay;
run;

data piecewise;
set medals;
if latitude le 30 then cslope=0;
if latitude gt 30 then cslope=latitude-30;
run;
proc print data=piecewise;
run;

proc reg data=piecewise;
model winter=latitude cslope;
output out=pieceout p=winterhat;
run;

symbol1 v=circle i=none c=black;
symbol2 v=none i=joint c=black;
proc sort data=pieceout;
by latitude;
proc gplot data=pieceout;
plot(winter winterhat)*latitude/overlay;
run;
```

Q2:

```
/*Part 1, Question 2*/
data medals1;
set medals;
sum=population+latitude;
run;
proc print data=medals1;
```

```
run;

*Q2 ai for Summer and Winter;
*Summer;
title1 'STAT512 Final Project Q2ai-Summer';
title2 'plotted by team1';
proc reg data=medals1;
model summer=population latitude/clm ss1 ss2;
run;
*Winter;
title1 'STAT512 Final Project Q2ai-Winter';
title2 'plotted by team1';
proc reg data=medals1;
model winter=population latitude/clm ss1 ss2;
run;

*Q2 aii for Summer and Winter;
*Summer;
title1 'STAT512 Final Project Q2aii-Summer';
title2 'plotted by team1';
proc reg data=medals1;
model summer=sum population latitude/clm ss1 ss2;
run;
*Winter;
title1 'STAT512 Final Project Q2aii-Winter';
title2 'plotted by team1';
proc reg data=medals1;
model winter=sum population latitude/clm ss1 ss2;
run;

*Q2 b for Summer and Winter;
*Summer;
proc reg data=medals1;
title1 'STAT 512 Final Project 2b-Summer';
title2 'plotted by team1';
model summer=sum population latitude/ss1 ss2;
test1: test sum;
run;
*Winter;
proc reg data=medals1;
title1 'STAT 512 Final Project 2b-Winter';
title2 'plotted by team1';
model winter=sum population latitude/ss1 ss2;
```

```
test1: test sum;
run;
```

Q3:

```
/*Part 1, Question 3*/
*Q3 for Summer and Winter;
*Summer;
proc reg data=medals1;
title1 'STAT 512 Final Project 2b-Summer';
title2 'plotted by team1';
model summer= population latitude/ss1 ss2;
run;
*Winter;
proc reg data=medals1;
title1 'STAT 512 Final Project 2b-Winter';
title2 'plotted by team1';
model winter= population latitude/ss1 ss2;
run;
```

```
/*Part 1, Question 4*/
proc reg data=new;
model summer=sum;
model summer= latitude;
model summer= population;
model summer= population latitude;
model summer=population sum;
model summer=latitude sum;
output out=diag r=resid;
run;
proc reg data=new;
model winter= population;
model winter=latitude;
model winter= sum;
model winter=population latitude;
model winter=latitude sum;
model winter= population sum;
output out=diag2 r=resid2;
run;
```

PART2

Q1:

```
/*Part 2, Question 1*/
```

```
proc gplot data=medals;
plot summer*population;
plot summer*latitude;
plot winter*population;
plot winter*latitude;
run;
data trans;
set medals;
logpopulation =log(population);
run;
proc print data=trans;
run;
proc gplot data=trans;
plot summer*logpopulation;
run;
proc gplot data=trans;
plot winter*logpopulation;
run;
proc reg data=trans;
model winter= latitude logpopulation;
output out=diag1 r=resid1;
run;
proc reg data=trans;
model summer= latitude logpopulation;
output out=diag2 r=resid2;
run;
data summer;
set medals;
drop winter;
run;
proc print data=summer;
run;
data winter;
set medals;
drop summer;
run;
proc print data=winter;
run;
data summer1;
set summer;
season=0;
medals= summer;
drop summer;
```

```
run;
proc print data=summer1;
run;
data winter1;
set winter;
season=1;
medals=winter;
drop winter;
run;
/*adding constant to winter medals to run transformation*/
data winter2;
set winter1;
medals1=medals + 1;
medals=medals1;
drop medals1;
run;
proc print data=winter2;
run;
data combine;
set summer1 winter2;
run;
proc print data=combine;
run;
data trans;
set combine;
logpopulation =log(population);
latpop= latitude*log(population);
sealat= season*latitude;
sealogpop= season*log(population);
run;
proc print data=trans;
run;
proc univariate data=trans plot;
var medals logpopulation latitude;
run;
proc gplot data=trans;
plot medals*logpopulation;
run;
proc gplot data=trans;
plot medals*latitude;
run;
proc reg data=trans;
model medals= latitude population;
```

```
model medals= latitude logpopulation;
model medals= latitude logpopulation season;
model medals= latitude logpopulation season latpop;
model medals= latitude logpopulation season latpop sealat;
model medals= latitude logpopulation season latpop sealat sealogpop;
run;
proc standard data=trans out=std mean=0;
var latitude logpopulation season latpop sqlatitude sqlogpop sealat sealogpop;
run;
proc sgscatter data=trans;
matrix medals latitude logpopulation season sqlatitude sqlogpop;
run;
/*Transformation for Y*/
proc transreg data=trans;
model boxcox(medals)=identity(latitude);
run;
proc transreg data=trans;
model boxcox(medals)=identity(population);
run;
proc transreg data=trans;
model boxcox(medals)=identity(latitude logpopulation season);
Run;

/*Transformed model*/
data trans_new;
set trans;
medalstrans = 1/sqrt(medals);
run;
proc print data=trans_new;
run;
proc reg data=trans_new;
model medalstrans = latitude logpopulation season latpop sealat sealogpop;
run;
```

Q3:

```
/*Part 2, Question 3 Stepwise selection*/
proc reg data=trans_new;
model medals= latitude logpopulation season latpop sealat sealogpop/ selection=stepwise;
run;
proc reg data=trans_new;
model medals= logpopulation season latpop sealat sealogpop/ selection=stepwise;
Run;
```

Q2:

```
/*Question 2, testing assumptions*/
proc gplot data=trans_new;
plot medalstrans*latpop;
plot medalstrans*season;
plot medalstrans*latitude;
run;
proc univariate data=trans1 plot;
var latitude season latpop;
qqplot latitude season latpop/ normal(L=1 mu=est sigma=est);
run;
/*Mallow's Cp*/
proc reg data=trans_new;
model medalstrans = latitude logpopulation season latpop sealat sealogpop /selection=cp b;
run;
/*Best Subset*/
proc reg data=trans_new;
model medalstrans = latitude logpopulation season latpop sealat sealogpop/selection=stepwise;
run;
proc univariate data=trans_new plot;
var latitude logpopulation season latpop sealat sealogpop;
run;

proc gplot data=trans_new;
plot medalstrans*latpop;
plot medalstrans*season;
plot medalstrans*latitude;
run;

proc reg data=trans_new;
model medalstrans = latitude logpopulation season latpop sealat sealogpop;
run;
```

Q5:

```
/*Question 5, Outlier, Influential Observations*/

/*Studentized deledted residuals and Cooks D*/
proc reg data=trans_new;
model medalstrans=latitude season latpop/r influence;
run;
/*Hat matrix, DFITS DBETAS*/
proc reg data=trans_new;
model medalstrans=latitude season latpop/tol vif;
```

```
run;
/*Partial Residuals Plots*/
proc reg data=trans_new;
model medalstrans=latitude season latpop/r partial;
run;
```

Q6:

```
/*Question 6*/
/*b*/
proc reg data=trans_new;
model medalstrans=latitude season latpop/clm alpha=0.10;
run;
/*c*/
proc reg data=trans_new;
model medalstrans=latitude season latpop/cli alpha=0.10;
run;
/*d*/
proc reg data=trans_new;
model medalstrans=latitude season latpop/clb alpha=0.10;
run;
```