Regression Analysis - STAT 512 Final Project

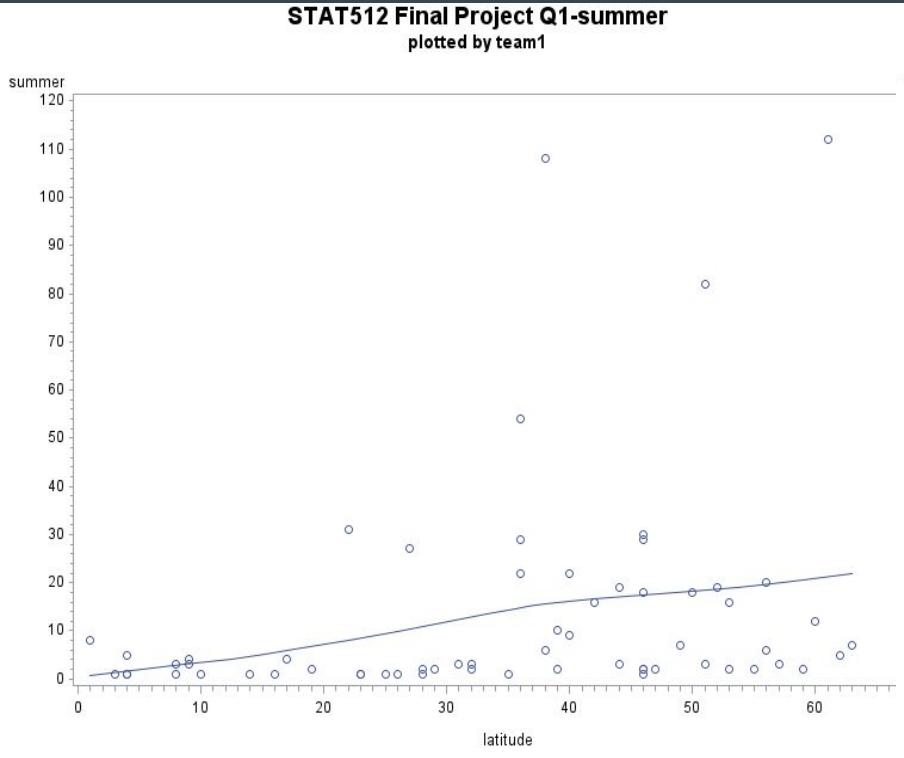
Group 1
Yunmei Bai, M. Reza Moini,
Upsana Angara, Natalie Ehmke

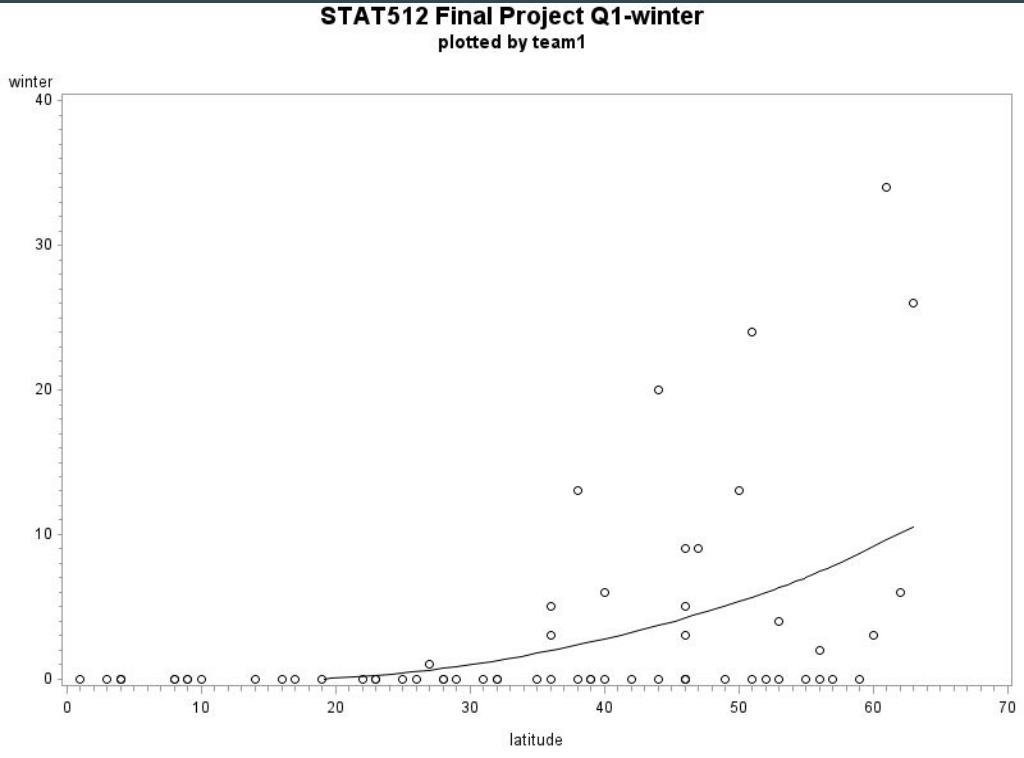
‘Olympic Medals’ Data Set

- Data contains the number of medals won by countries in:
  - 1992 Summer Olympic Games in Barcelona, Spain
    - Response 1
  - 1994 Winter Olympic Games in Lillehammer, Norway
    - Response 2
  - 2 Predictors: Population and Latitude of each country
    - Incomplete data set as countries that did not win medals were not included.
- Latitude and Population were used to predict the number of medals won in each Olympic Game

Part 1- Question 1

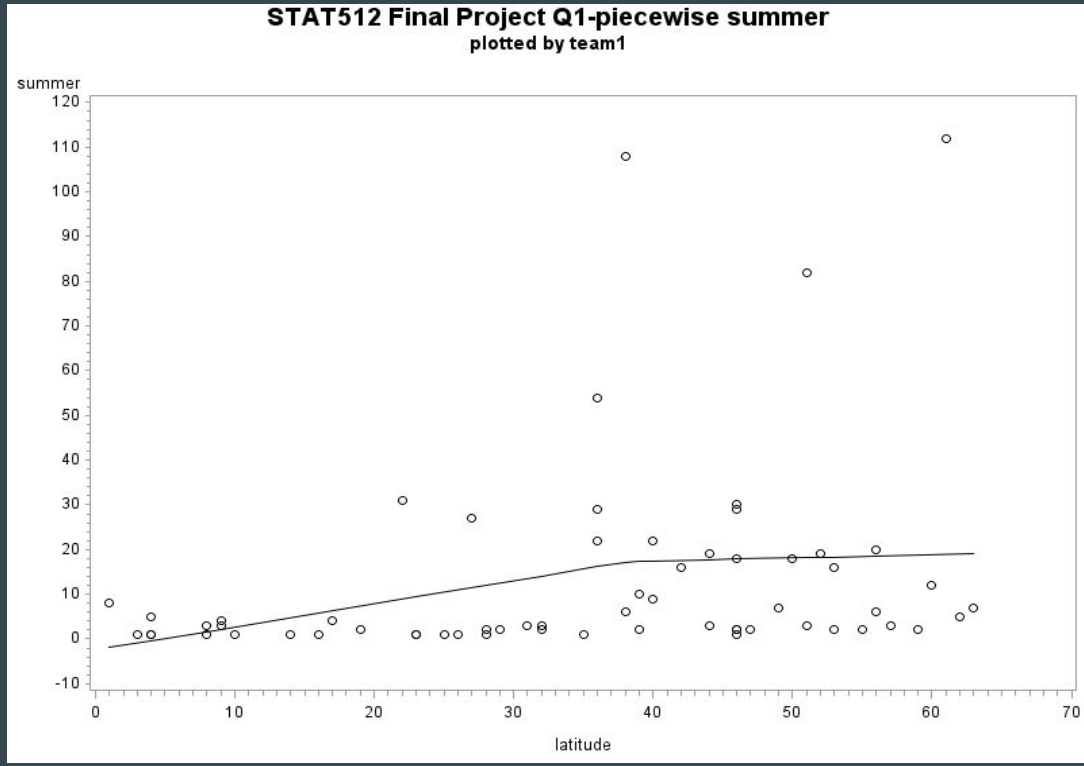
Preliminary Scatter plots with smoothing lines, latitude is chosen predictor





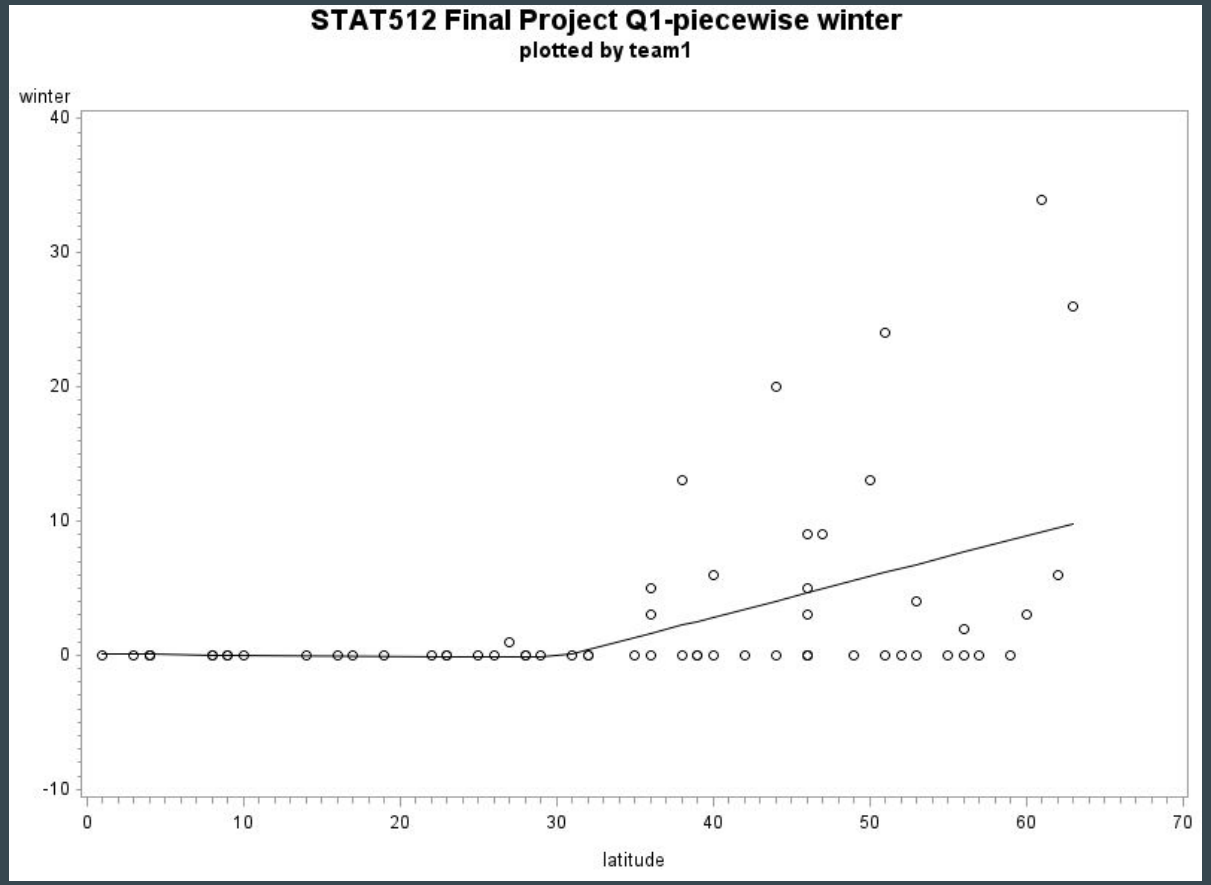
Part 1-Question 1

- SLR using latitude as a predictor
- Summer Medals
- Piecewise plot shows change in slope at x=38



Part 1- Question 1

- SLR using latitude as a predictor
- Winter Medals
- Piecewise plot shows change in slope at x=30



The SAS System

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	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

Testing slopes of different pieces

Ho: $\beta 1 = \beta 2$

Ha: The slopes are different

Summer: Ho, Slopes are equal

Winter: Ha, Slopes are not equal

The SAS System

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	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

Part 1- Question 2a

Model with predictors population and latitude

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

$$F = \frac{(SSE(R) - SSE(F)) / (df_E(R) - df_E(F))}{SSE(F)/df_E(F)}$$

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

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

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

Part 1- Question 2a

Sum=Latitude+Population

- Model is not full rank due to need to run model with sum AND predictors latitude and population
- B" and "0" under degrees of freedom indicate bias

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	.	.	.	.	.

Part 1- Question 2b

Proc Reg

- The degrees of freedom for reduced model:
- Summer=(1,59)
- Winter=(1,59).
- H0 : Slope(given by coefficient of SUM)= 0
- Ha : Slope ≠ 0
- The p value for summer is 0.0087 < 0.05(α)
  - Significant; Reject the null hypothesis
  - There is a linear relationship between summer medals and SUM
- The pvalue for winter is 0.0006< 0.05 (α)
  - Significant; Reject the null hypothesis
  - There is a linear relationship between winter medals and SUM

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		

Part 1- Question 2c

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~Fn-p,p-1 = (t*)^2

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

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		

Part 1- Question 3

The Values of Type I and type II SS for summer:

- Sum of type I errors for summer = 5807.62+2852.25 = 8569.87 = SSM(summer)
- T1SS(latitude)=T2SS(latitude)
  - So... 2852.25 = 2852.25

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

Part 1- Question 3

The Values of Type I and type II SS for winter:

- Sum of type I errors for winter = 51.84+520.48=572.32 = SSM(winter)
- T1SS(latitude)=T2SS(latitude)
  - So... 520.4872=52.4872

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

Part 1- Question 4

Comparison of Regression Models with variety of variables including SUM:

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

Part 1- Question 4

Comparison of Regression Models with variety of variables including SUM:

Winter:

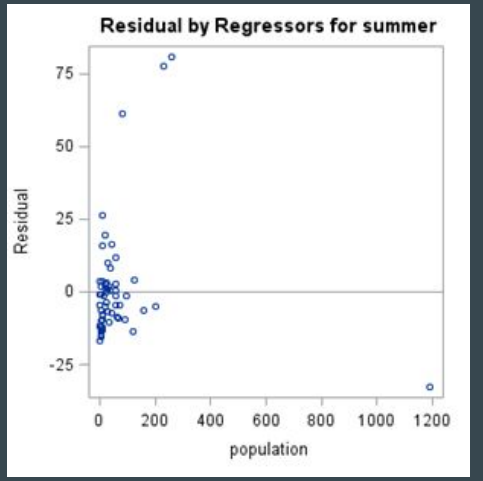
Predictor #s	R^2	P-value	Intercept	Population	Latitude	Sum	MSE
2	0.1960	0.0016	-3.217	0.0069	0.1659		39.791
2	0.1960	0.00016	-3.217	-0.1590		0.1659	39.791
2	0.1960	0.0016	-3.2165		0.0459	0.0069	39.791
1	0.0178	0.3018	2.6479	0.0059			47.803
1	0.1720	0.0008	-2.6967		0.1628		40.295
1	0.0324	0.1615	2.2455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

Part 2- Question 1

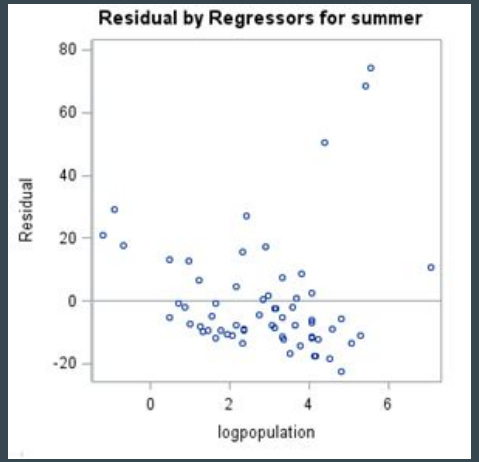
Transformation of Variable (Population) - Summer:

- Residual/Scatter plot suggested transformation of “population” variable
- “logpopulation” was then used in lieu of “population” for the rest of the model.

Skewed before transformation:



After Transformation:

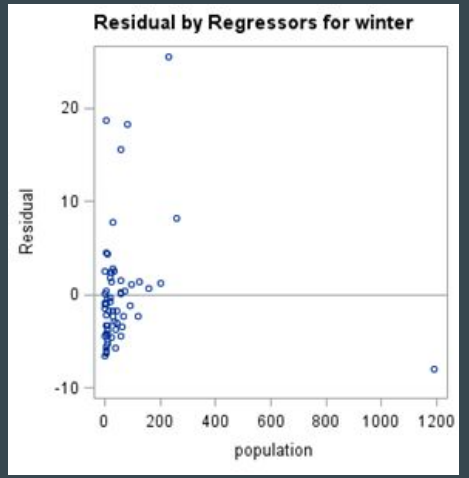


Part 2- Question 1

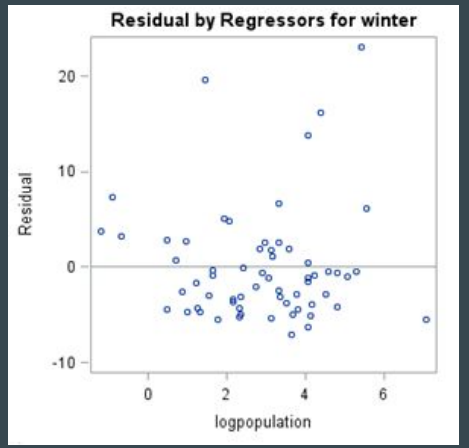
Transformation of Variable (Population) - Winter

- Residual/Scatter plot suggested transformation of “population” variable
- “logpopulation” was then used in lieu of “population” for the rest of the model.

Skewed before transformation:



After Transformation



Part 2- Question 1

Additional Season Variable & Interaction Variables:

- Variable "**Season**" was added to the variables as binary (0=for summer, 1 = Winter Olympics medals) to account for the binary nature of the data set in 1 single regression model.

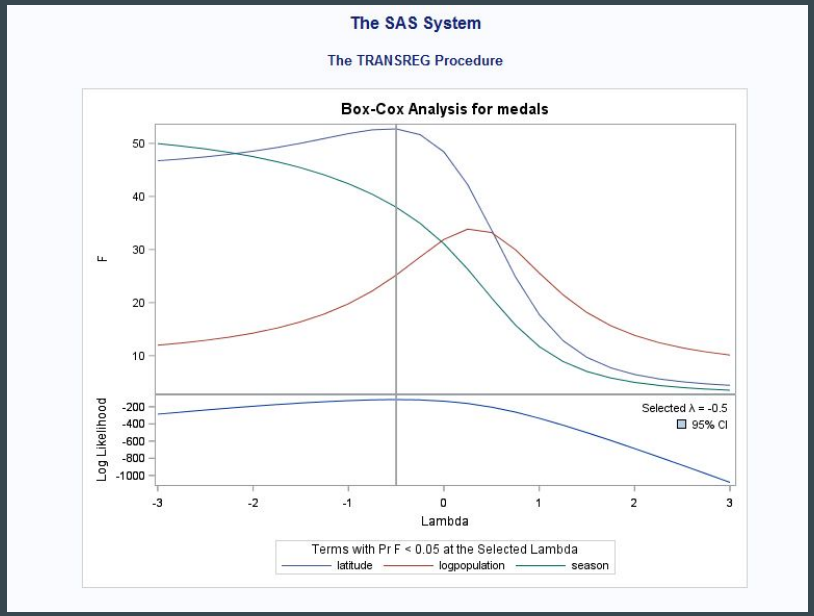
- Following interaction models were also taken into account:
- logPopulation & Latitude
- Season & LogPopulation
- Season & Latitude

# of Variables	Model	R^2	Adjusted R^2
2	latitude population	0.17	0.15
2	latitude logpopulation	0.22	0.20
3	latitude logpopulation season	0.29	0.27
4	latitude logpopulation season latpop	0.42	0.40
5	latitude logpopulation season latpop sealat	0.43	0.40
6	latitude logpopulation season latpop sealat sealogpop	0.50	0.48

Part 2- Question 1

Transformation of Response (Y) - Olympic Medals:

Best $\lambda = -0.5$

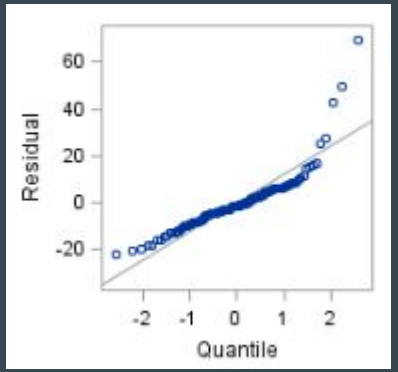


Model then includes:

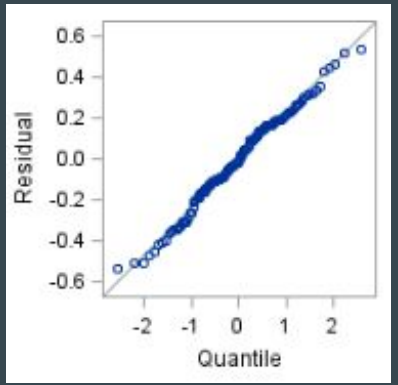
Y: 1 / Sqrt (Medals)

X's : (latitude logpopulation season latpop sealat sealogpop)

Before Transformation:



After Transformation:



Part 2-Question 2

As per Mallow's Cp condition Cp ≤ p would be a good model.

Number in Model	C(p)	R-Square	Parameter Estimates						
			Intercept	latitude	logpopulation	season	latpop	sealat	sealogpop
3	1.7771	0.5036	0.86888	-0.00324	.	0.26982	-0.00217	.	.
4	3.1374	0.5063	0.86745	-0.00257	.	0.21588	-0.00240	.	0.01885
3	3.4252	0.4966	0.82396	.	.	0.16587	-0.00288	.	0.03632
3	3.4627	0.4965	0.75550	.	0.02925	0.26982	-0.00304	.	.
4	3.7758	0.5036	0.86586	-0.00317	0.00100	0.26982	-0.00219	.	.
4	3.7771	0.5036	0.86900	-0.00324	.	0.26959	-0.00217	0.00000659	.
2	4.5516	0.4835	0.79899	.	.	0.26982	-0.00262	.	.
3	4.6251	0.4916	0.78146	.	.	0.36137	-0.00244	-0.00262	.
4	4.7082	0.4997	0.78795	.	0.01791	0.20492	-0.00304	.	0.02268
4	4.8982	0.4989	0.80929	.	.	0.23761	-0.00273	-0.00150	0.02961
5	5.0213	0.5068	0.89831	-0.00317	-0.01034	0.20492	-0.00219	.	0.02268
5	5.1227	0.5063	0.87267	-0.00271	.	0.20416	-0.00240	0.00030012	0.01928
4	5.1269	0.4979	0.75541	.	0.02360	0.31411	-0.00287	-0.00127	.
5	5.7758	0.5036	0.86598	-0.00317	0.00100	0.26959	-0.00219	0.00000659	.
4	5.7816	0.4951	0.93693	-0.00337	.	.	-0.00271	0.00340	0.04438
5	6.4985	0.5005	0.78537	.	0.01428	0.24533	-0.00291	-0.00101	0.02092
2	6.6358	0.4747	0.87869	.	.	.	-0.00315	.	0.07426
3	6.6508	0.4830	0.86835	.	.	.	-0.00320	0.00173	0.06312
6	7.0000	0.5069	0.90554	-0.00335	-0.01065	0.19047	-0.00219	0.00036151	0.02331

Conclusion:
We pick the model with Cp=1.771 based since it fulfills the criteria with the smallest Cp

Part 2 Question 3

Project STAT 512_Data Set Olympic Medals
Reza Moini, Natalie Ehmke, Yunmei Bai,Upasana Angara

The REG Procedure
Model: MODEL1
Dependent Variable: medalstrans

Number of Observations Read	124
Number of Observations Used	124

Stepwise Selection: Step 1

Variable latpop Entered: R-Square = 0.3153 and C(p) = 42.4567

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4.23022	4.23022	56.17	<.0001
Error	122	9.18814	0.07531		
Corrected Total	123	13.41837			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	0.68365	0.02464	57.95470	769.52	<.0001
latpop	-0.00262	0.00034963	4.23022	56.17	<.0001

Stepwise selection in SAS reported the following.

The test was run at significance level alpha=0.15

Running the test at a lower significance level would mean that the likelihood of rejection is smaller and vice versa

Part 2 Question 3

Stepwise Selection: Step 3

Variable latitude Entered: R-Square = 0.5036 and C(p) = 1.7771

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	6.75719	2.25240	40.58	<.0001
Error	120	6.66118	0.05551		
Corrected Total	123	13.41837			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	0.68365	0.02116	57.95470	1044.04	<.0001
latitude	-0.00324	0.00147	0.27003	4.86	0.0293
season	0.26982	0.04232	2.25693	40.66	<.0001
latpop	-0.00217	0.00036403	1.96532	35.40	<.0001

Bounds on condition number: 1.4708, 11.825

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

At this point, the remaining variables did not make it into the model and therefore the best model was reported

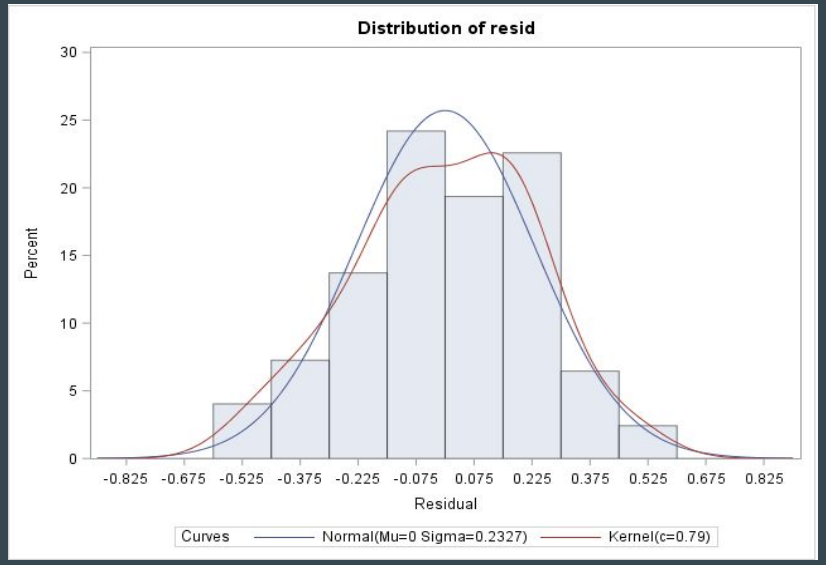
Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	latpop		1	0.3153	0.3153	42.4567	56.17	<.0001
2	season		2	0.1682	0.4835	4.5516	39.40	<.0001
3	latitude		3	0.0201	0.5036	1.7771	4.86	0.0293

Conclusion:Based on both selection criteria, the model with the 3 predictors was picked.

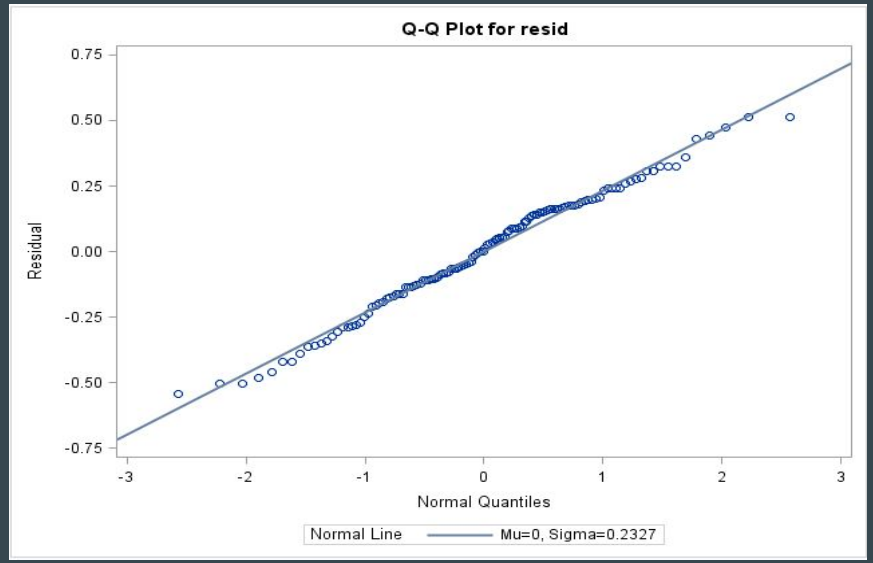
PART 2 - Q.4 Check the Best Model Assumptions

Normality Assumption:

- Histogram

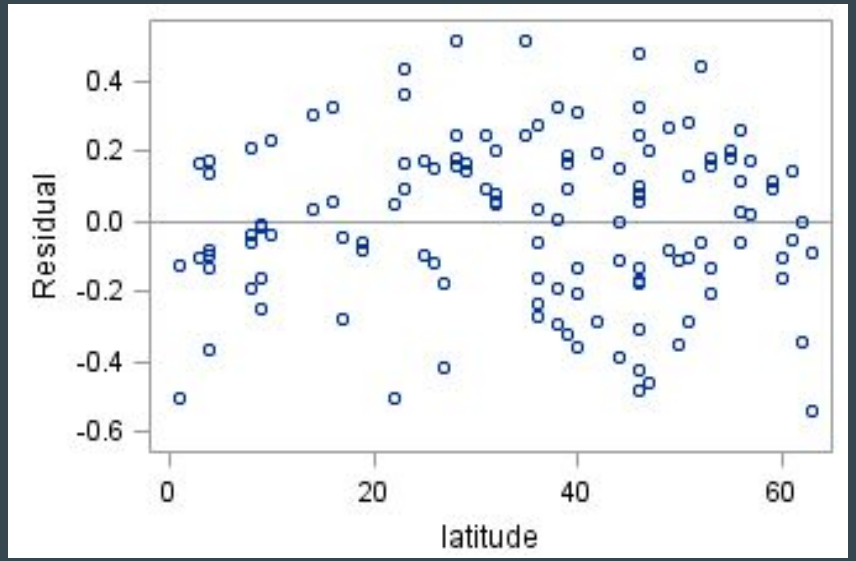


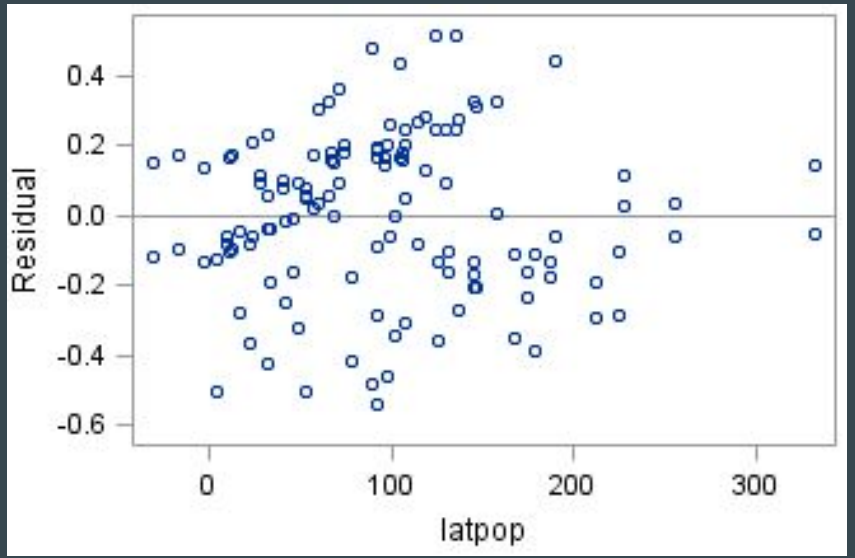
- QQ plot



Constant Variance and Linearity Assumption

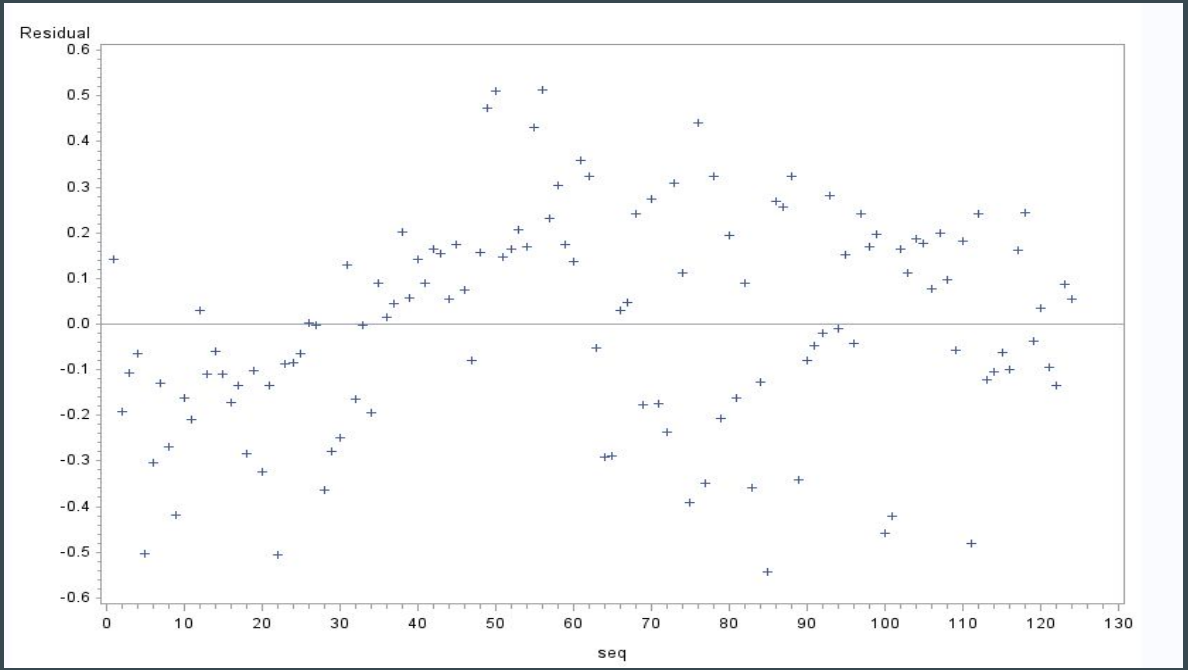
- Residual Plot





Independence Assumption

- Residual vs. Sequence Plot



PART 2 - Q.5 Outliers and Influential Observations

						Output Statistics		
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	Cook's D	RStudent
1	0.3536	0.8584	0.0511	-0.5049	0.230	-2.195	0.060	-2.2313
2	1.0000	0.8360	0.0491	0.1640	0.230	0.712	0.006	0.7104
3	0.4472	0.8100	0.0480	-0.3628	0.231	-1.573	0.027	-1.5827
4	1.0000	0.8303	0.0482	0.1697	0.231	0.736	0.006	0.7345
5	1.0000	0.8639	0.0491	0.1361	0.230	0.591	0.004	0.5891
6	0.5774	0.7724	0.0443	-0.1951	0.231	-0.843	0.007	-0.8420
7	1.0000	0.7937	0.0445	0.2063	0.231	0.892	0.007	0.8909
8	0.5000	0.7504	0.0435	-0.2504	0.232	-1.081	0.010	-1.0819
9	0.5774	0.7410	0.0435	-0.1636	0.232	-0.707	0.004	-0.7052
10	1.0000	0.7680	0.0426	0.2320	0.232	1.001	0.008	1.0014

Outlier Test

- Studentized Residual
- Studentized Deleted Residual

Ho: Case i is not an outlier

Ha: Case i is an outlier

Tc = +/- 3

Tc - t(n-p-1, alpha/2n)

Influential Observation Test

- Cook's D = Fp, n-p(50%) = F4,120(0.5) = 0.844
- Hat Matrix = 2p/n = 0.0645
- DeFITS = 2 sq(p/n) or 1 = 1(small & medium size)
- DFBetas = 2/sq(n) or 1 = 1(small & medium size)

Influential Observations :

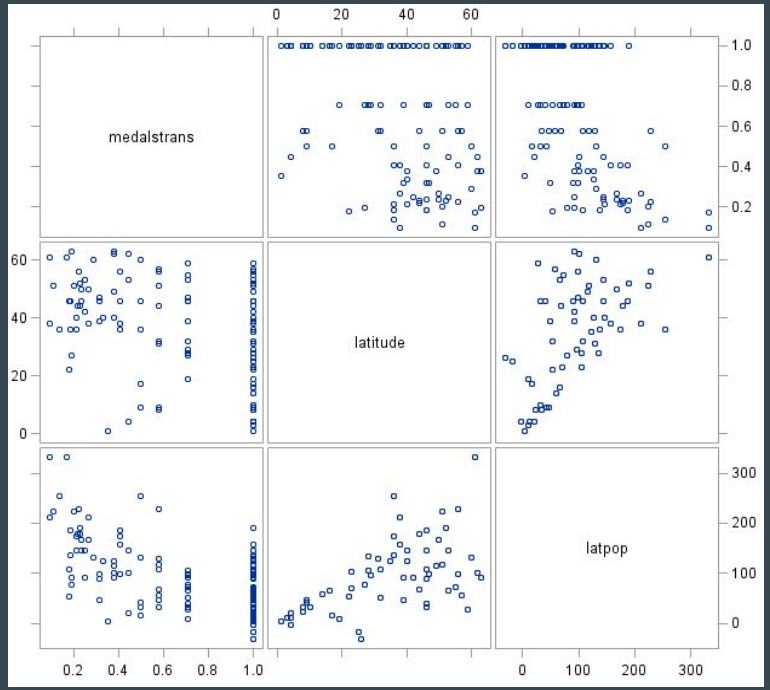
Cook's D - None

Hat Matrix - Observation 28,58,60,90,120,122

DeFITS - None

DFBetas - None

Case	latitude	season	latpop
58	59	0	27.73
59	60	0	130.49
60	61	0	332.12

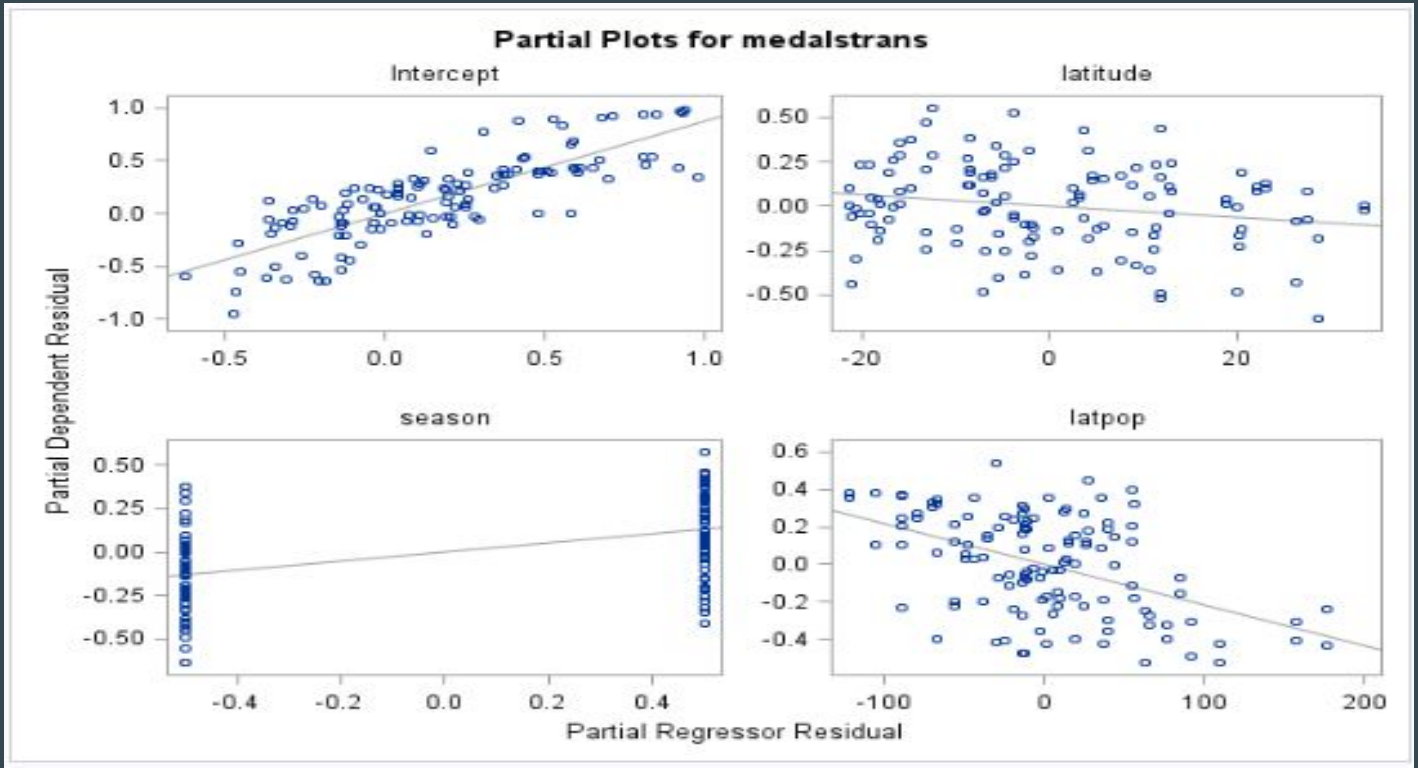


Multicollinearity

- Variance Inflation Factor(VIF) > 10 are considered as excessive multicollinearity
- Tolorance < 0.1

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	0.86888	0.05216	16.66	<.0001	.	0
latitude	1	-0.00324	0.00147	-2.21	0.0293	0.67989	1.47082
season	1	0.26982	0.04232	6.38	<.0001	1.00000	1.00000
latpop	1	-0.00217	0.00036403	-5.95	<.0001	0.67989	1.47082

Partial Regression Plot



PART 2 - Q.6

- 90% Confidence Interval for Mean and Predictions

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

90% Confidence Interval for Regression Coefficients

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