What county demographics can tell us about poverty rates

Group Members: Rachel Stevenson, Yash Chaturvedi, Elizabeth Thompson, Glenn Klahre,
Adam Tarko

Statistics 512 -- July 30, 2018 -- Dr. Nandy

1 Introduction

County demographic information (CDI) is often collected on a county-population scale to determine various facets of a population such as per capita income, average number of children per household, population density, and crime rates. However, CDI can also be further analyzed to provide guidance into future plans and policies regarding more global complex issues such as poverty (Brookhaven National Laboratory). For instance, Australian demographic data has shown that low income, poor education, and poor health are good indicators of poverty rates (Callander and Schofield, 2015). Comparatively, South African poverty has been correlated with gender of the head of household, type of dwelling, and unemployment (Meyer, 2016). In the United States, education, family income, immigration status, and health all are shown to be correlated with poverty (Hoynes and Stevens, 2006). Although poverty is a complex problem that could be affected by many aspects of a society, better understanding of these relationships can be useful tools in combating the problem. For instance, if low high school graduation rates are found to be an indicator of future poverty rates, then future policies can be focused upon raising education rates and in turn lowing poverty rates (Lacour and Tissington, 2011).

Poverty rates can also be affected by macroeconomic factors, so one could assume that the historical year the data was collected may also have an effect on poverty rates, for example, during a recession or following the implementation of large scale governmental policies (Hoynes and Bitler, 2016). Most research concerning indicators of poverty are analyzing census or CDI data from more recent years, while less attention is being brought to years prior to 2000 (FAO Statistical Development Series). The purpose of this report is to analyze previously collected county demographic (CDI) data from 440 of the most populous counties in the United States from 1990 to 1992 to determine if a relationship between a variety of demographic information and poverty rates exists through statistical regression analysis.

The main objectives are to find the best model of poverty through regression analysis in SAS 9.4 from the following CDI variables: high school graduation rates, percent of population with a bachelor's degree, unemployment, and per capita income, population density and land area. We also aim to compare the results to more recent studies of poverty predictors to see if education and income are still strongly associated with poverty.

2 Methods

2.1 Data Description

The dataset used in the analysis of this project was adapted from the Applied Linear Statistical Models, 5th ed. textbook appendix C.2 dataset (Kutner et al. 2012). The 17 variable dataset was reduced to analyze 7 different explanatory variables (**Table 1**) with the response variable as poverty rate. Where poverty rate was defined as “the percent of 1990 CDI population with income below the poverty line” (Kutner et al. 2012). The sample size of the CDI data comes from the 440 most populated counties in the United States (n=440) with county used as the experimental unit.

Explanatory Variable Name	**Variable Short Names (if given)**	**Unit**	**Description**	**Discrete vs. Continuous**
High School Graduation rate	*HS*	Percent	Percent of adult population who completed 12 or more years of school.	Continuous
Percent Bachelor’s degrees	*College*	Percent	Percent of adult population with a bachelor’s degree.	Continuous
Per capita income	*IPC, Income*	Dollars	Per capita income of 1990 CDI population (in dollars).	Continuous
*Population Density***	*PD*	People per square mile	Estimated 1990 population density. Population/Land Area	Continuous
Unemployment	--	Percent	Percent of 1990 CDI labor force that is unemployed	Continuous
Total Population	*Pop, Population*	Number of people	Estimated 1990 population	Continuous
Land Area	--	Square Miles	Land Area (sq miles).	Continuous

Table 1: descriptions of each of the explanatory variables from Kutner et al. 2012
** indicates an adaption to the original dataset

2.2 Preliminary Analysis

A preliminary analysis of the data was performed before any model selection or diagnostics were taken. First, each of the variables was analyzed for abnormalities in SAS 9.4 including non-normality, unequal variance, linearity relationships, independence, and outliers. Each variable was analyzed by overseeing bivariate histograms, qqplots, and residual plots (**Figures 1, 12, 20**). By checking for assumptions on each variable, one can safely and accurately use various regression analysis later in the project to choose the best model of poverty. If any variables were found to possess any of the previously mentioned issues, a transformation was applied to that variable to remedy the problem.

The explanatory variable “population” was found to possess an outlier as well as displaying behavior of a non-normally distributed variable (right skewed)(**Figure 20 and 12**). Population density was calculated by dividing Population by Land Area, however the data was found to also be non-normally distributed as well (**Figure 20 and 12**). Because of this, a Box-Cox transformation statement was applied to both variables to which the output suggested a logarithmic transformation for population density, and a square root transformation of population.

Each of the variables were also run through a correlation analysis to determine if there was any high correlations between X variables. See appendix for correlation matrix (**Figure 21**) and Pearson Correlation output (**Figure 13**). Most variables did not appear to be highly dependent on each other with the largest correlation coming from high school graduation and college graduation (0.70). However, because these were under the 0.8 cutoff for Pearson Correlation Outputs, we still used both high school and college within the model.

Next we plotted the raw data points on a scatterplot for the variable “income per capita” (IPC) to determine how this particular variable behaved. Upon examining the plot we found that the IPC was great for piecewise regression due to the shape of the plot (**Figure 2, 4, 5**). We split our original scatter plot into two pieces and the regression point of 20,000 was taken as this was the inflection point on the graph. This seems to do a much better job of describing the trend in the data. To determine if the two pieces were similar or not, we ran the F value test, which had a value of 124.12 which was a high value and thus different from one another. However, upon further analysis it was determined that this regression change did not effect R^2 in the model for IPC. Therefore, we did not include the piecewise regression addition in the full model for simplicity sake. We observed this same effect for each of the X variables, so no piecewise regression changes were added to the model.

Due to the higher correlation between high school and college, we decided to run addition tests by creating a new variable SUM = HS + College. We ran 2 regression models without using HS and College to predict the response using all other explanatory variables, while the second model we also included the variable SUM. Comparing the model sum of squares between the first model (4754.45) and the second model (4842.114) we find the extra sum of squares to be 87.664. The extra sum of square was derived from subtracting Model sum of squares of the first model from the second model. Alternatively, we could have also derived this number from SS of the SUM variable. **Figures 6, 7, 8,** and **9** depict the output for these steps.

Next, we ran a regression to predict the response using all predictor variables except SUM. We added the Type I sum of squares for the predictors and then did same for Type II sum of squares. We then ran regression to predict response using different variables, including SUM as an explanatory variable. We made 10 different models, with different variables and calculated their R^2 and R^2 adjusted values and summarize them in **Table 2**.

2.3 Appropriate Model Selection

A model selection process was performed using the Cp criterion process. A stepwise selection technique was used in addition to the criterion statement in SAS 9.4 (**Table 3 and 4**).

Before running those models, histograms were studied to see if any of the variables need transformation. While many were slightly skewed, after some deliberation, only two were determined to need transformation: Population and Population density. The others, while slightly skewed were not transformed out of preference to pure linearity. Unfortunately, even after transformation, Population appeared to violate the Homoscedastic principal and was removed from analysis (Figure 25). Upon removing the variable from the model selection, we also see that Cp rises, however, we found this to be necessary in order to meet the assumptions of the model (**Table 3**). As mentioned before, land area was also left out of the model due to the non-linear relationship with poverty.

2.4 Diagnostics of Model Assumptions

After deciding upon the model, analysis was run and assumptions checked. The residuals didn't appear to be homoscedastic, and the qq plot showed an outlier. The notable outlier, due to it's unusually high residual and Student value (above six), on the qqplot was removed to see the possible changes (**Figure 15**), and it was determined that it did not have enough effect on the model to warrant removal, and the final model stood.

Multicollinearity could be an issue as we had previously seen that high school and college had a strong correlation, and if there is too much correlation between the variables this could pose issues with regression coefficients. However, after observing the output of the ANOVA table of the model, we did not see any indication that multicollinearity was an issue (**Table 6 and 7**). If there *was* an issue, we would expect see abnormally large regression coefficients despite a small p value. In addition, a Pearson Correlation and Variance of Inflation test was conducted on the best model (**Figures 13** and **14**).

Next, we checked to see if there was a linear relationship between the X variables and Y variable. This was done by creating scatterplots for each of the variables in relation to poverty (**Figure 26**). As mentioned before, land area and population were removed due to fail to meet regression assumptions. We noticed that for the variables used in the final model, a linear relationship could be observed.

Using the scatter and qqplots from the previous assumption, we can also check that the residuals have constant variance. We did notice that the residuals on each variable were clustered, but the variance within the cluster seemed approximately equal. Cook's D, studentized residual, and partial residual plots were also analyzed in order to check assumptions of the model.

A single outlier was found within the dataset which was removed to observe the effect this point had on the model. However, it was determined that the removal of the outlier posed no substantial change in the behavior of the model therefore it was left in the data (**Figure 15)**.

Thankfully, all assumptions appeared to be met aside from the outlier issue. Therefore, no additional transformations were needed.

2.5 Inferential Methods to Reach Hypothesized Conclusion

For the main model we used a pure F-test and t-tests for the individual parameters. We kept a p-value of .005 in mind in that regard.
We kept an eye on the t-tests, using a 90% total confidence interval, correcting with the Bonferroni correction, came to 99% intervals. In other words, and alpha of .01 was used for the Mean CI, and Parameter CIs. In addition to parameter CIs, a 90% prediction interval was also created for each individual observation. **Figures 16, 17, 18**, and **19**.

3 Results

3.1 Preliminary Analysis Results

3.1.1 Analysis of Individual Variables

From **Figure 1**, we can see that Land is not necessarily a good indicator for the variable poverty. Both histograms and residual plots of each variable can also be found in the appendix section for more visual information. We determined that land area, population, and population density were to be excluded from further analysis.

3.1.2 Correlations Between Variables

From the output shown in **Table 2** and the correlation matrix (**Figure 21**) we can see the relationship each variable has with one another. It seems college and high school (0.71), high school and poverty (-0.69), and college and income (0.70) all have high correlations.

3.1.3 Piecewise Analysis

The scatterplot of Poverty (Dependent) vs IPC (Independent) was plotted, and the regression point of 20,000 was chosen for the model; this resulted in a slope of -0.00126 until point 20,000 and a slope of -0.0002 after, as seen in **Figures 3, 4,** and **5**. Comparing the two slopes, we found there to be a significant difference between the two: F-value = 63.68 and p < .0001, it is safe to say they are different (**Figure 3, 4**)

3.1.4 Sum of Squares and the Variable SUM

Q1 P2) a)

Extra sum of squares is: 4842.114 - 4754.45 = 87.664.
The degrees of freedom are 1 and 434.
F(1,434) = 87.664/10.778 = 8.133

Q1) P2) b)
H_0 - coefficient of SUM variable is 0. H_a - coefficient of SUM variable is not 0.
Using the test statement in proc reg, the test statistic was obtained and is shown in **Figures 6** and **7**.
F value = 8.13   P value = 0.0046   Degrees of freedom are 1 and 434
The P value is not 0 and therefore we can reject the null hypothesis and therefore we can say that the coefficient of SUM is not 0.

Q1) P2) c) The test statistic = 8.13 and the P value = 0.0046 from the test statement for the SUM variable. The SUM variable output from the full model had the same P value of 0.0046 and test statistic = 8.13. Therefore as these values are consistent we can reject the null hypothesis. (**Figures 6** and **7**)

Q1 P3)
The Type I Sum of Squares for Predictors:
13.76 + 3758.9 + 554.75 + 427.01 + 862.10 + 898.52 = 6515.04
The Type II Sum of Squares for Predictors:
32.91 + 1798.10 + 108.28 + 72.84 + 1576.57 + 898.52 = 4487.22
(**Figures 6** and **7**)

The Type I sum of squares match the model sum of squares. The only predictor for which the two sums of squares are same is the College predictor. The Type I sum of squares is the sequential sum of squares therefore the order that the variables are input into the model is important. The SSI values will change if the order is changed. Type II sum of squares considers all predictors in the model. Regardless of the order of input, we will always get the same values for each predictor. The College predictor had the same value for both Type I and Type II sum of squares tests because it was the last variable in the sequence. If we had another variable at the last in SSI calculations, it would have same SSII value.

Q1 P4) The R^2 and R^2 adjusted values are summarized in **Table 2**.
The R^2 value for each predictor is always greater than its R^2 adjusted value because the latter only takes into account the predictors that affect the accuracy of the model.

The R^2 value increases with more variables, but the adjusted R^2 is better for comparing models with different number of predictors.

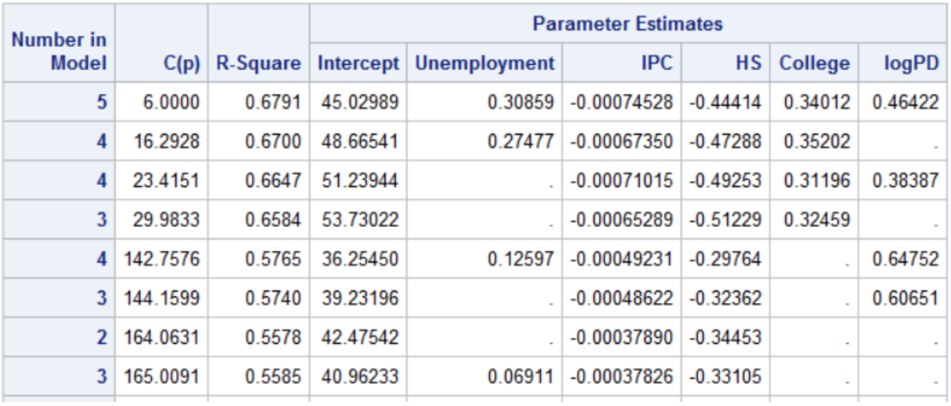
Note - The w/o Unemployment, SUM. This means that Unemployment rate has little to no effect on Poverty rates once variables like HS grad rates are included.

3.2 Variable Transformations

There were several variables we looked seriously at transforming. A majority of our variables are slightly skew on way or another. (See histograms in Appendix) For the most part we decided that they weren't bad enough to warrant transformation. With two exceptions: Population and Population density. Both were heavily and obviously right skewed.
Using Boxcox analysis we decided upon our transformations. Log transform for Population Density and Inverse Square Root for Population. The Boxcox graphs used to make the decision are included in the Appendix B section (**Figures 10** and **11**).

3.2.1 Cp Criterion

Q2 P3) We use the □□ criterion to select the best subset of variables for the data. We use the original and transformed variables, not SUM.



Number in Model	C(p)	R-Square	Parameter Estimates					
			Intercept	Unemployment	IPC	HS	College	logPD
5	6.0000	0.6791	45.02989	0.30859	-0.00074528	-0.44414	0.34012	0.46422
4	16.2928	0.6700	48.66541	0.27477	-0.00067350	-0.47288	0.35202	.
4	23.4151	0.6647	51.23944	.	-0.00071015	-0.49253	0.31196	0.38387
3	29.9833	0.6584	53.73022	.	-0.00065289	-0.51229	0.32459	.
4	142.7576	0.5765	36.25450	0.12597	-0.00049231	-0.29764	.	0.64752
3	144.1599	0.5740	39.23196	.	-0.00048622	-0.32362	.	0.60651
2	164.0631	0.5578	42.47542	.	-0.00037890	-0.34453	.	.
3	165.0091	0.5585	40.96233	0.06911	-0.00037826	-0.33105	.	.

Table 3: Cp Criterion of Transformed and Normal Variables
We ran into issues with the transformed Income variable. It has the look of possibly failing the homoscedastic test (**Figure 25**). The highest R^2 value is of the model that contains the predictors unemployment, log(population density), income, high school, and college. The data sets that have a C_p value less than or almost equal to the number of predictors will give us less variability in the analysis and therefore will be best suited to us. Based on **Table 3** and **Table 4**, tt seems as if the best model would include those aforementioned variables with R^2 = 0.6791 and C_p = 6.000.

3.3 Diagnostics of Model Assumptions Results

The pearson correlations look fine (**Table 5**), all fall below the 0.8 standard. Some of them are a mite higher than we would like, but it shouldn't cause to much of an issue.

The residual plot for the best model (**Figure 23**) is not great looking. It isn't surprising that there is a condensed area in the middle, but the shape isn't as spread out as we would perhaps like. What is most concerning is the way it definitely appears to become less spread out as it gets smaller. This raises massive questions, but it is likely not substantial enough to warrant the analysis completely invalid.

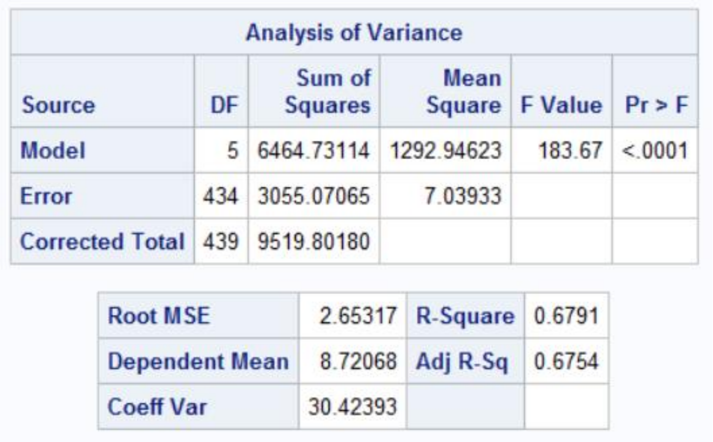
The qq plot(**Figure 24**) looks decent, with the exception of the outlier at top.
That outlier was removed (**Figure 15**), but none of the Parameters showed substantial changes, and there was minimal effect overall on the model, so it was left in.

3.4 Inferential Methods to Reach Hypothesized Conclusion Results

Below is the regression equation for the best model:
Poverty = 45.030 - 0.001(IPC) - 0.444(HS) + 0.340(College) + 0.309(Unemployment) + 0.464(logPD)

With the Confidence Limits of the mean response variable poverty: [8.355 , 9.087]

3.4.1 ANOVA and Parameter Estimates of Final Model

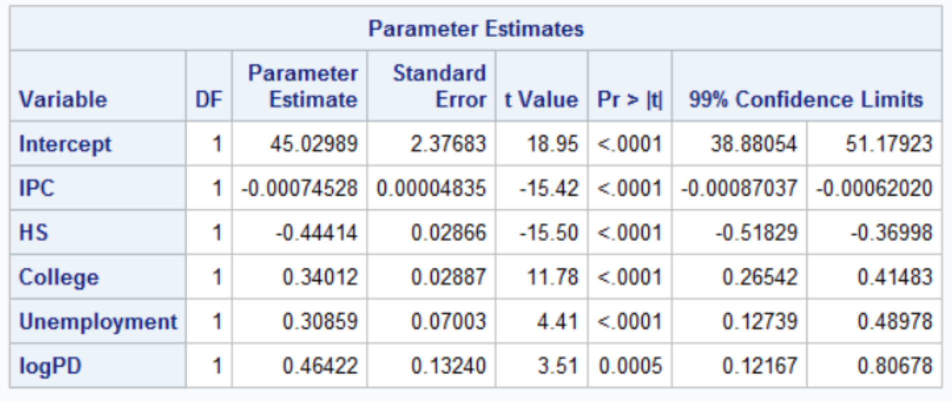


Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	6464.73114	1292.94623	183.67	<.0001
Error	434	3055.07065	7.03933		
Corrected Total	439	9519.80180			

Root MSE	2.65317	R-Square	0.6791
Dependent Mean	8.72068	Adj R-Sq	0.6754
Coeff Var	30.42393		

Table 6: ANOVA for Best Model



Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	99% Confidence Limits	
Intercept	1	45.02989	2.37683	18.95	<.0001	38.88054	51.17923
IPC	1	-0.00074528	0.00004835	-15.42	<.0001	-0.00087037	-0.00062020
HS	1	-0.44414	0.02866	-15.50	<.0001	-0.51829	-0.36998
College	1	0.34012	0.02887	11.78	<.0001	0.26542	0.41483
Unemployment	1	0.30859	0.07003	4.41	<.0001	0.12739	0.48978
logPD	1	0.46422	0.13240	3.51	0.0005	0.12167	0.80678

Table 7: Parameter Estimates for Best Model.

In **Figures 12** and **13** in the appendix are the 90% confidence interval graphs for Mean prediction and Individual points. Bonferroni correction was used for the mean prediction graph bringing it to a 99% confidence interval.

4 Discussion

Poverty is a global problem that has been the basis for many political and economic debates throughout history (Hoynes and Bitler 2016). Knowing indicators of poverty can help mitigate or possible reduce the overall rate within a population. Through our analysis of CDI data from 1990 and 1992, we were able to use the variables high school graduation rate, percent of population with a bachelor's degree, income per capita, population density, and unemployment rates to serve as an indicator of poverty rates.

From what we have gathered from the previous research, United States poverty has been shown to be related to income and education throughout history (Lacour and Tissington 2011) . Interestingly, it seems our 1990 and 1992 United State CDI model on poverty is also comparable to a 2015 study on Australian census data regarding poverty’s predictors (Callander and Schofield 2015). Therefore, we could infer that first world countries may have similar predictors of poverty compared to second or first world countries (Meyer 2016). Although, the data was from the 1990s, we can also see that it is still similar to more recent indicators of poverty too (Brookhaven National Laboratory ; Kutner et al. 2005).

However, there were several limitation with this project analysis. We are unable to research the methods for which the CDI data was collected, and therefore we cannot confirm the design of the study was properly conducted. We were also unable to know whether or not the outlier discovered in the dataset was due to a human error such as a typo or an intentional outlier. One could also argue that our reasoning for omitting certain aspects of the model was flawed. For instance, we chose to run the variable IPC without a piecewise function even though we stated that this function better described the data. Since there was little change in R^2 we omitted this, but one could argue that this piecewise function should have been included. Other actions that could be debated were our decision to leave the outlier in the data and choosing to omit population density from the model.

There is also the issue at the tails of the prediction. At high predicted values the model became imprecise, though still useful enough to tell that poverty would be high. Still, it is apparent enough that the model could use revision at high levels of poverty. Not only that, but the problem with lack of homoscedasticity in the final model is not to be ignored.

While the model may be of some use within pure retrospective analysis, as a predictive measure, it is highly suspect at best.

5 References

Brookhaven National Laboratory, and United States. Dept. of Energy. Office of Scientific Technical Information. Use of County Level Data in Health, Energy, Demographic, Environmental, and Economic Analysis. Upton, N.Y. : Oak Ridge, Tenn.: Brookhaven National Laboratory ; Distributed by the Office of Scientific and Technical Information, U.S. Dept. of Energy, 1979.

Callander, E. J., Schofield, D. J. (2015). Multidimensional poverty and health status as a predictor of chronic income poverty. *Health Economics, 24* (12), 1638-643.

FAO Statistical Development Series. Importance of the census of agriculture. 11(2010): 9-16.

Hoynes, H. & Bitler, M. (2016). The more things change, the more they stay the same? The safety net and poverty in the great recession. *Journal of Labor Economics, 34* (S1), S403-S444.

Hoynes, H. W., Page, M. E., & Stevens, A. H. (2006). Poverty in america: Trends and explanations. *Journal of Economic Perspectives, 20* (1), 47-68.

Kutner, M., Nachtsheim, C., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed., McGraw-Hill/Irwin series operations and decision sciences).

Lacour, M. & Tissington, L. D. (2011). The effects of poverty on educational achievement. *Educational Research and Review, 6* (7), 522-527.

Meyer, D. F. (2016). Predictors of poverty: A comparative analysis of low income communities in the northern free state region, South Africa. *International Journal of Social Sciences and Humanity Studies*, *8*(2), 132-149.

6 Appendix A -- SAS Code

Part 1 Question 1 piecewise framework;

```
if IPC le 20000 then pieceslope=0;
if IPC gt 20000 then pieceslope=IPC - 20000;
run;

*Use to find non-linear, curved function;

proc sort data=Proj;by IPC;
proc gplot data=Proj;
symbol1 v=circle i=sm70;
plot Poverty*Pop Poverty*HS Poverty*College Poverty*IPC Poverty*Unemployment Poverty*PD;
run;

*After this I decided upon IPC, because it looked like the most optimal one for a piecewise function.
I decided upon 20,000 as the break, because that is where the inflection point appeared to be.;

*Get the piecewise slopes;
proc reg data=Proj;
model Poverty=IPC pieceslope;
output out=pieceout p=povhat;
```

```
test pieceslope;
run;

*Set up and plot piecewise;
proc sort data=pieceout; by IPC;
proc gplot data=pieceout;
plot (Poverty povhat)*IPC /overlay;
run;
```

Part 1 Question 2;

```
proc reg data=Proj;
Model Poverty = Pop IPC Unemployment PD / ss1;
model Poverty = Pop IPC Unemployment PD SUM/ ss1;
Sum: test SUM=0;
run;
```

Part 1 Question 3;

```
proc reg data=Proj;
Model Poverty = Pop IPC Unemployment PD HS College/ ss1 ss2;
*Note: The reason why the ss1 and ss2 will be the same for the last value is because ss2 simply reruns the regression to find the SSE if the variable is the last variable.
The last variable in ss1 is already last in that regression, so it won't change.;
run;
```

Part 1 Question 4;

```
proc reg data=Proj;
*First set of attempts: Is SUM better than HS College? Is there an individual which is best?;
JustSum: model Poverty= Pop IPC PD Unemployment SUM;
JustHS:model Poverty= Pop IPC PD Unemployment HS;
JustCollege:model Poverty= Pop IPC PD Unemployment College;
NoSum:model Poverty= Pop IPC PD Unemployment HS College;
*Second set of attempts: Is Pop better than PD? Feel free to change these based upon the first set of models.;
NoPD:model Poverty= Pop IPC Unemployment HS College;
NoPop:model Poverty= PD IPC Unemployment HS College;
NoPDPop:model Poverty= IPC Unemployment HS College;
*Third set of attempts: Does income matter? Does unemployment? Are the two of them tied super closely?;
NoI:model Poverty= Pop PD Unemployment HS College;
NoU:model Poverty= Pop PD IPC HS College;
NoUI:model Poverty= Pop PD HS College;
Run;
```

***Part 2 Question 1;**

```
proc corr data=ProjReduced noprob plot(maxpoints=NONE)=matrix(nvar=7);
run;
```

Part 2 Question 2

```
proc reg data=ProjReduced;
model Poverty=Pop HS College Unemployment IPC /r influence;
run;

title1 'Poverty';
proc reg data=ProjReduced;
model Poverty=Pop HS College Unemployment IPC/r partial;
plot r.*(Poverty Pop HS College Unemployment IPC);
run;

proc transreg data=ProjReduced test;
model boxcox(PD)= identity(Poverty);
run;

proc transreg data=ProjReduced test;
model boxcox(Pop) = identity(Poverty);
run;

data ProjR;
set ProjReduced;
isrPop = Pop**(-0.5);
logPD = log(PD);
run;

proc univariate data=ProjR;
var Poverty Unemployment IPC HS College logPD isrPop;
qqplot / normal;
histogram;
Run;
```

Part 2 Question 3;

```
proc reg data=ProjR;
title1 'Best Variables';
model poverty=land pop hs college unemployment income/ selection = cp b best=5;
run;
```

Part 2 Question 4;

```
proc reg data=ProjR;
model Poverty = Unemployment IPC HS College logPD/ selection=stepwise;
run;
```

Part 2 Question 5;

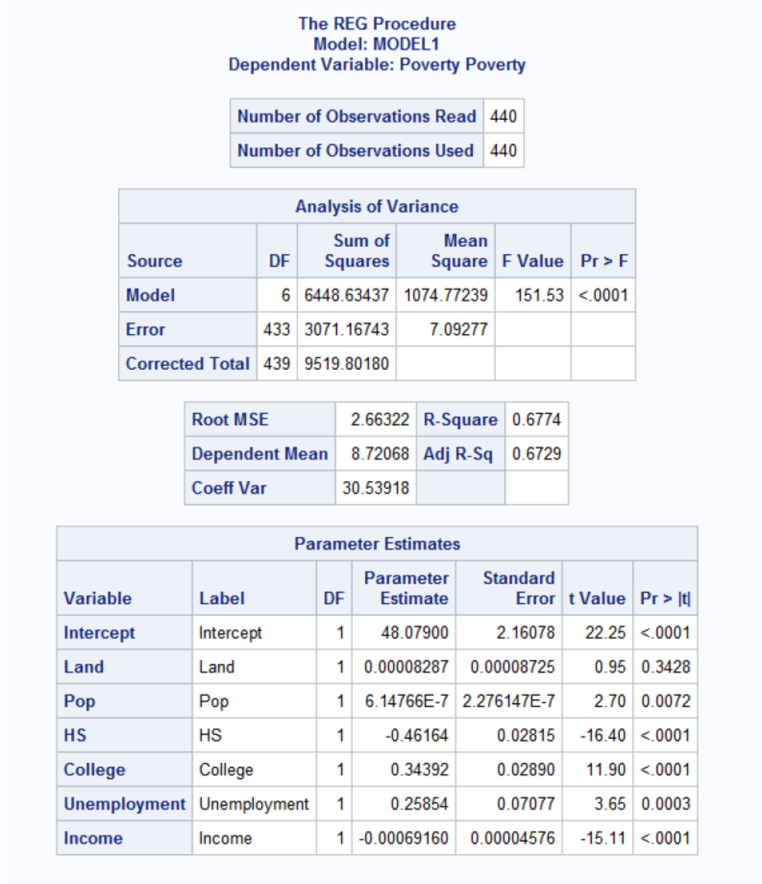
```
proc corr data=ProjR noprob;
var Unemployment IPC HS College logPD;
run;
```

Part 2 Question 6:7;

```
proc reg data=ProjR;
model Poverty = IPC HS College Unemployment logPD/r p vif;
```

```
run;

data ProjRd;
set ProjR;
if Poverty = 27.3 then delete; *Removing the outlier through brute force;
run;

proc reg data=ProjRd;
model Poverty = IPC HS College Unemployment logPD;
run;

proc reg data=ProjR;
model Poverty = IPC HS College Unemployment logPD /alpha=.01 clb;
output out=outp p=phat;
run;

proc reg data=ProjR;
model Poverty = IPC HS College Unemployment logPD /alpha=.1 clb cli;
output out=outp p=phat;
run;
**this gives us the 90% CI needed for question 7 a b and c (above)

proc gplot data=outp;
title1 '90% individual confidence interval';
symbol1 v=star i=rlcli90;
plot Poverty*phat;
run;

proc gplot data=outp;
title1 '90% mean confidence interval';
symbol1 v=star i=rlclm99;
plot Poverty*phat;
run;

proc means data=proj lclm uclm alpha=.1;
var poverty;
run;
**above code shows the CL for mean response variable (question 7 part b)
```

7 Appendix B -- SAS Output



The REG Procedure
Model: MODEL1
Dependent Variable: Poverty Poverty

Number of Observations Read	440
Number of Observations Used	440

Analysis of Variance

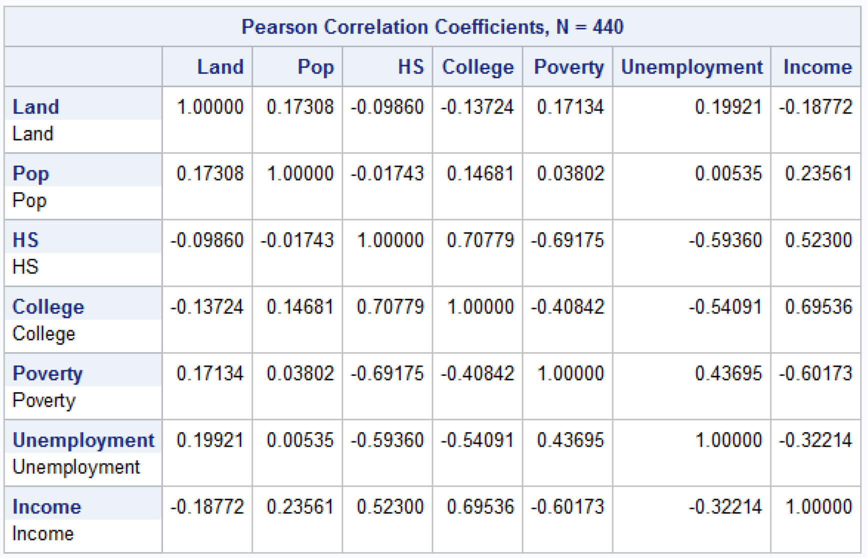
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	6448.63437	1074.77239	151.53	<.0001
Error	433	3071.16743	7.09277		
Corrected Total	439	9519.80180			

Root MSE	2.66322	R-Square	0.6774
Dependent Mean	8.72068	Adj R-Sq	0.6729
Coeff Var	30.53918		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	48.07900	2.16078	22.25	<.0001
Land	Land	1	0.00008287	0.00008725	0.95	0.3428
Pop	Pop	1	6.14766E-7	2.276147E-7	2.70	0.0072
HS	HS	1	-0.46164	0.02815	-16.40	<.0001
College	College	1	0.34392	0.02890	11.90	<.0001
Unemployment	Unemployment	1	0.25854	0.07077	3.65	0.0003
Income	Income	1	-0.00069160	0.00004576	-15.11	<.0001

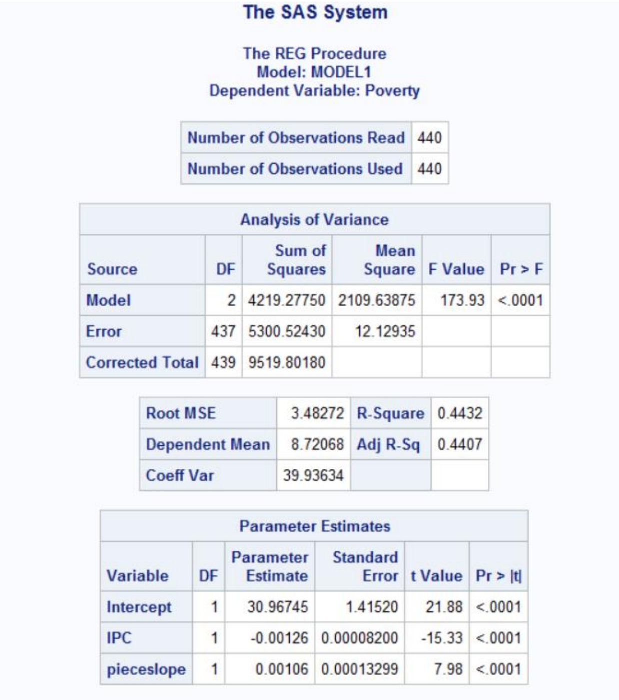
Figure 1: Indicates ANOVA and Parameter Estimates of each of the variables.



Pearson Correlation Coefficients, N = 440

	Land	Pop	HS	College	Poverty	Unemployment	Income
Land Land	1.00000	0.17308	-0.09860	-0.13724	0.17134	0.19921	-0.18772
Pop Pop	0.17308	1.00000	-0.01743	0.14681	0.03802	0.00535	0.23561
HS HS	-0.09860	-0.01743	1.00000	0.70779	-0.69175	-0.59360	0.52300
College College	-0.13724	0.14681	0.70779	1.00000	-0.40842	-0.54091	0.69536
Poverty Poverty	0.17134	0.03802	-0.69175	-0.40842	1.00000	0.43695	-0.60173
Unemployment Unemployment	0.19921	0.00535	-0.59360	-0.54091	0.43695	1.00000	-0.32214
Income Income	-0.18772	0.23561	0.52300	0.69536	-0.60173	-0.32214	1.00000

Figure 2: Output of Pearson Correlation Coefficients between each of the variables



The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: Poverty

Number of Observations Read	440
Number of Observations Used	440

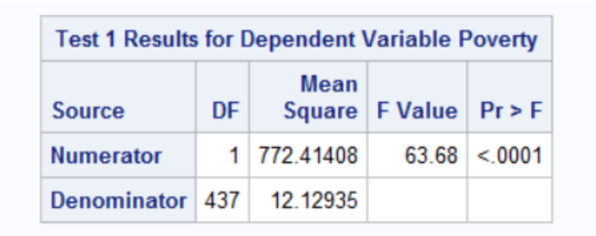
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	4219.27750	2109.63875	173.93	<.0001
Error	437	5300.52430	12.12935		
Corrected Total	439	9519.80180			

Root MSE	3.48272	R-Square	0.4432
Dependent Mean	8.72068	Adj R-Sq	0.4407
Coeff Var	39.93634		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	30.96745	1.41520	21.88	<.0001
IPC	1	-0.00126	0.00008200	-15.33	<.0001
pieceslope	1	0.00106	0.00013299	7.98	<.0001



Test 1 Results for Dependent Variable Poverty

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	772.41408	63.68	<.0001
Denominator	437	12.12935		

Figure 3 (left): Output of the comparison between slopes of the inflection points
Figure 4 (right): Significance test to determine whether both slopes were different from one another

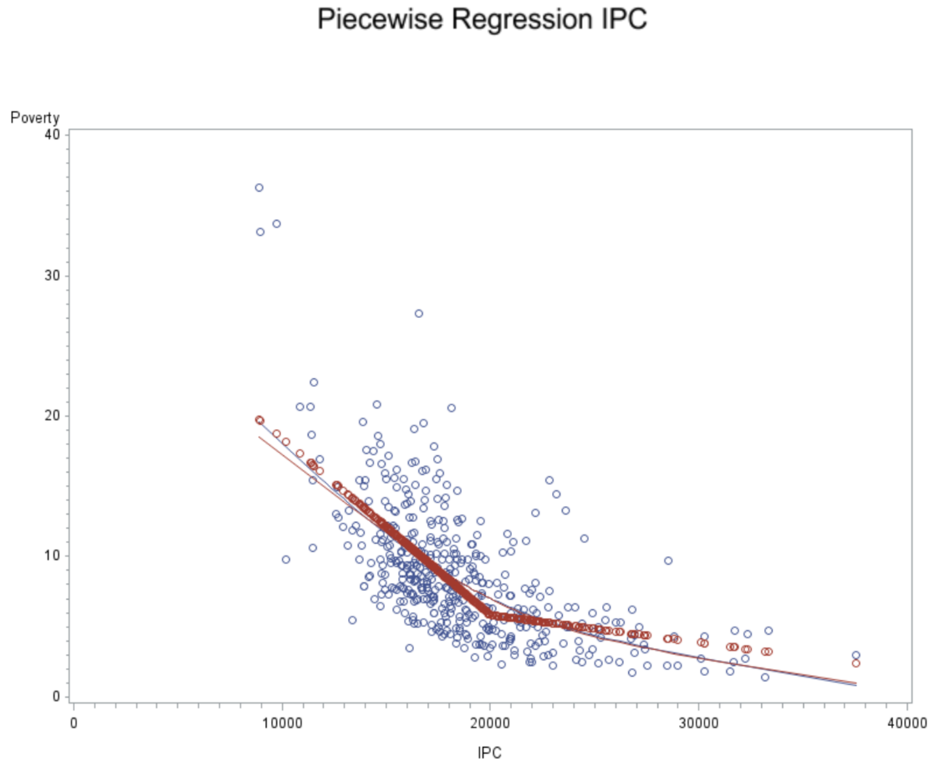
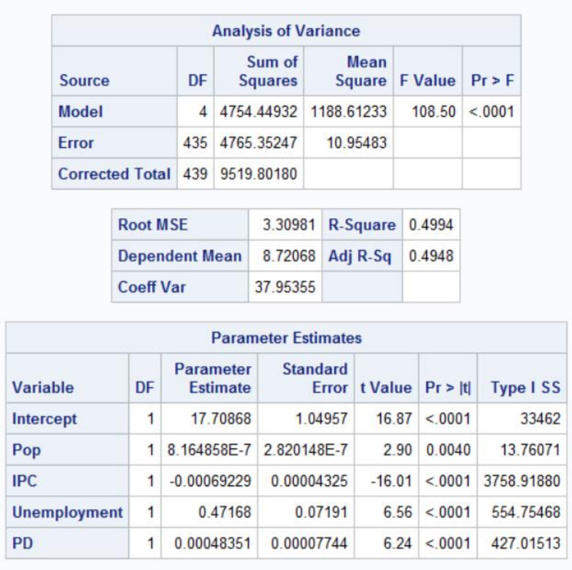


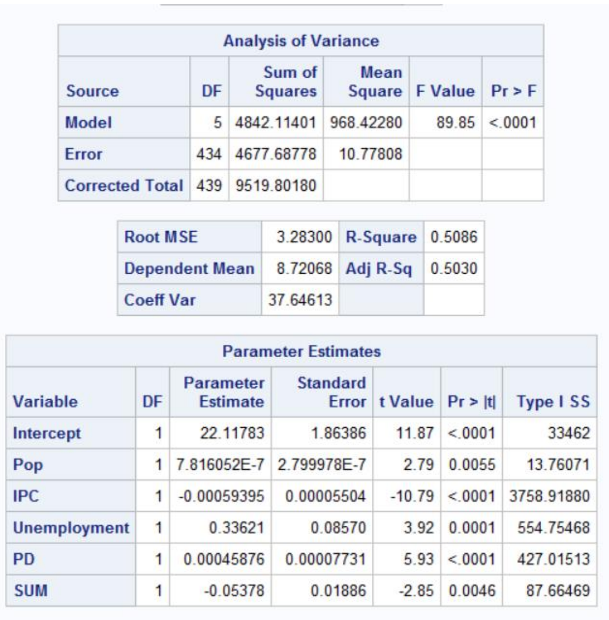
Figure 5: piecewise regression output at point 20000 for the variable income per capita (IPC)
Note: This is the output from Part 1, Question 1



Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	4754.44932	1188.61233	108.50	<.0001
Error	435	4765.35247	10.95483		
Corrected Total	439	9519.80180			

Root MSE	3.30981	R-Square	0.4994
Dependent Mean	8.72068	Adj R-Sq	0.4948
Coeff Var	37.95355		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS
Intercept	1	17.70868	1.04957	16.87	<.0001	33462
Pop	1	8.164858E-7	2.820148E-7	2.90	0.0040	13.76071
IPC	1	-0.00069229	0.00004325	-16.01	<.0001	3758.91880
Unemployment	1	0.47168	0.07191	6.56	<.0001	554.75468
PD	1	0.00048351	0.00007744	6.24	<.0001	427.01513

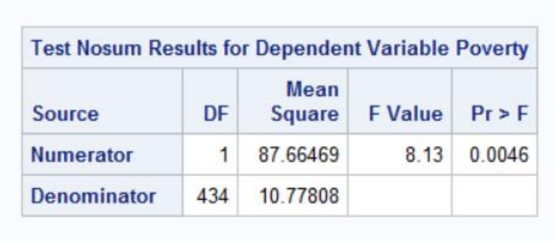


Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	4842.11401	968.42280	89.85	<.0001
Error	434	4677.68778	10.77808		
Corrected Total	439	9519.80180			

Root MSE	3.28300	R-Square	0.5086
Dependent Mean	8.72068	Adj R-Sq	0.5030
Coeff Var	37.64613		

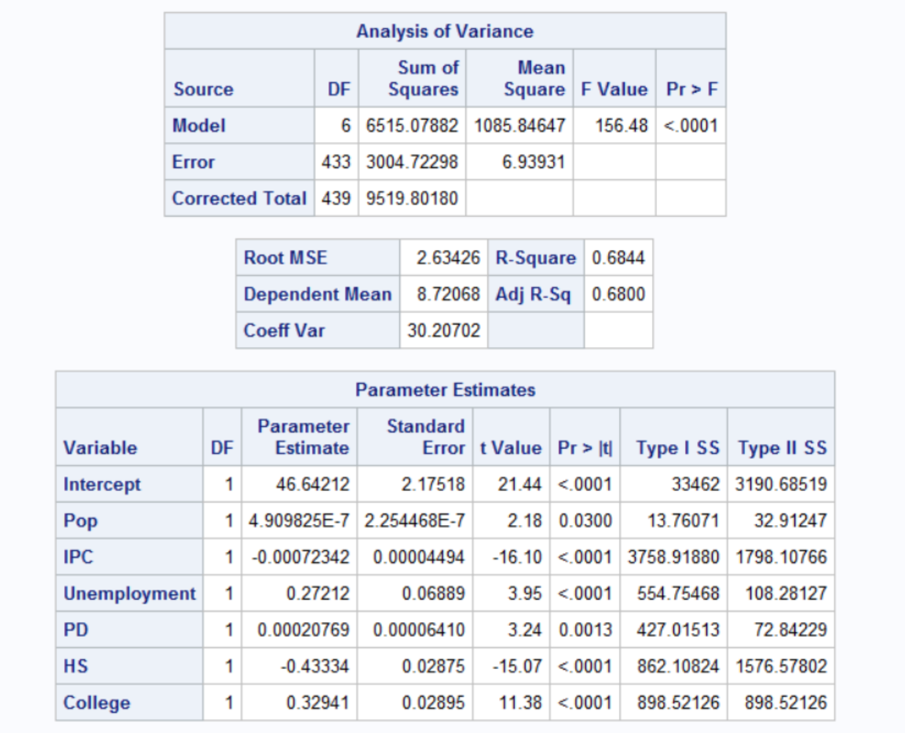
Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS
Intercept	1	22.11783	1.86386	11.87	<.0001	33462
Pop	1	7.816052E-7	2.799978E-7	2.79	0.0055	13.76071
IPC	1	-0.00059395	0.00005504	-10.79	<.0001	3758.91880
Unemployment	1	0.33621	0.08570	3.92	0.0001	554.75468
PD	1	0.00045876	0.00007731	5.93	<.0001	427.01513
SUM	1	-0.05378	0.01886	-2.85	0.0046	87.66469

Figure 6 (left): Output before SUM
Figure 7 (right): Output after SUM



Test Nosum Results for Dependent Variable Poverty				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	87.66469	8.13	0.0046
Denominator	434	10.77808		

Figure 8: Output of P and F values to determine that we reject the null



Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	6515.07882	1085.84647	156.48	<.0001
Error	433	3004.72298	6.93931		
Corrected Total	439	9519.80180			

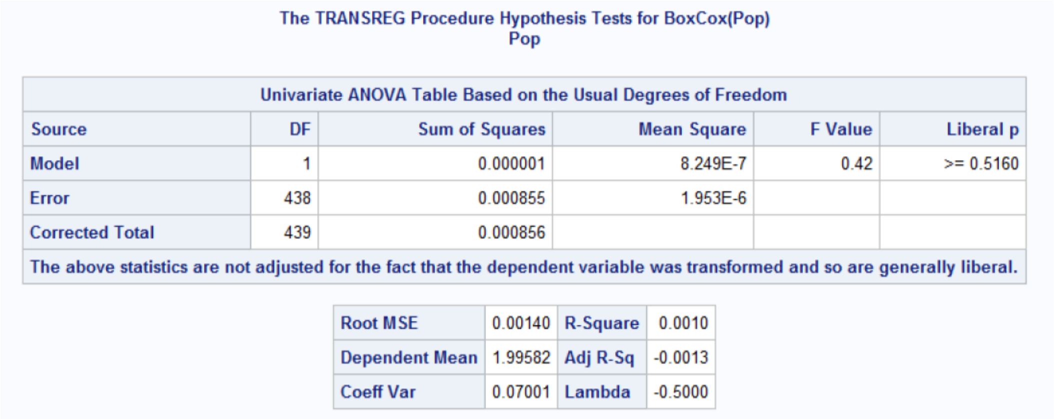
Root MSE	2.63426	R-Square	0.6844
Dependent Mean	8.72068	Adj R-Sq	0.6800
Coeff Var	30.20702		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS
Intercept	1	46.64212	2.17518	21.44	<.0001	33462	3190.68519
Pop	1	4.909825E-7	2.254468E-7	2.18	0.0300	13.76071	32.91247
IPC	1	-0.00072342	0.00004494	-16.10	<.0001	3758.91880	1798.10766
Unemployment	1	0.27212	0.06889	3.95	<.0001	554.75468	108.28127
PD	1	0.00020769	0.00006410	3.24	0.0013	427.01513	72.84229
HS	1	-0.43334	0.02875	-15.07	<.0001	862.10824	1576.57802
College	1	0.32941	0.02895	11.38	<.0001	898.52126	898.52126

Figure 9: Output of Sum of Squares

Model	R^2	R^2_{adj}
w/o HS, College	.5086	.5030
w/o SUM, College	.5900	.5853
w/o SUM, HS	.5188	.5132
w/o SUM	.6844	.6800
w/o PD, SUM	.6767	.6730
w/o Pop, SUM	.6809	.6772
w/o Pop, PD, SUM	.6700	.6670
w/o IPC, SUM	.4955	.4897
w/o Unemployment, SUM	.6730	.6692
w/o IPC, Unemployment, SUM	.4921	.4874

Table 2: calculated R^2 and R^2 adjusted values

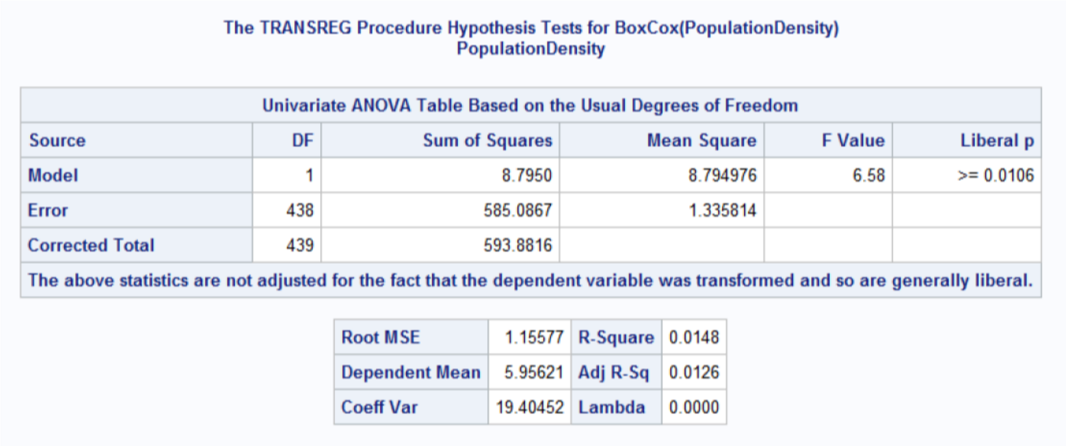


The TRANSREG Procedure Hypothesis Tests for BoxCox(Pop)
Pop

Univariate ANOVA Table Based on the Usual Degrees of Freedom					
Source	DF	Sum of Squares	Mean Square	F Value	Liberal p
Model	1	0.000001	8.249E-7	0.42	>= 0.5160
Error	438	0.000855	1.953E-6		
Corrected Total	439	0.000856			
The above statistics are not adjusted for the fact that the dependent variable was transformed and so are generally liberal.					

Root MSE	0.00140	R-Square	0.0010
Dependent Mean	1.99582	Adj R-Sq	-0.0013
Coeff Var	0.07001	Lambda	-0.5000

Figure 10: Box-Cox statement results for Population



The TRANSREG Procedure Hypothesis Tests for BoxCox(PopulationDensity)
PopulationDensity

Univariate ANOVA Table Based on the Usual Degrees of Freedom					
Source	DF	Sum of Squares	Mean Square	F Value	Liberal p
Model	1	8.7950	8.794976	6.58	>= 0.0106
Error	438	585.0867	1.335814		
Corrected Total	439	593.8816			
The above statistics are not adjusted for the fact that the dependent variable was transformed and so are generally liberal.					

Root MSE	1.15577	R-Square	0.0148
Dependent Mean	5.95621	Adj R-Sq	0.0126
Coeff Var	19.40452	Lambda	0.0000

Figure 11: Box-Cox statement results for Population Density

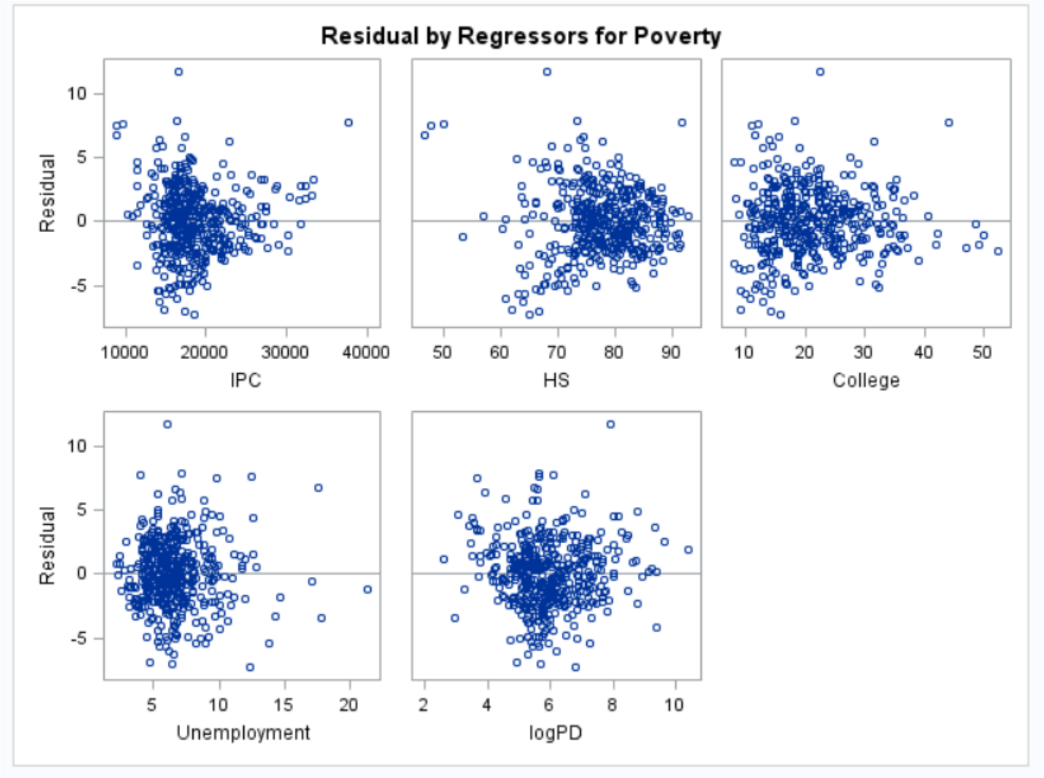
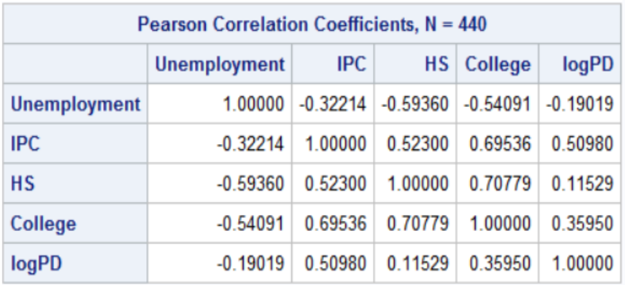
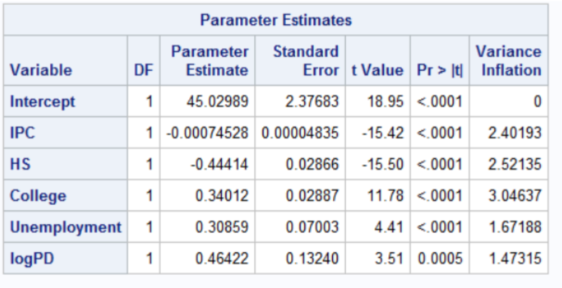


Figure 12: Residual plots of each of the variables

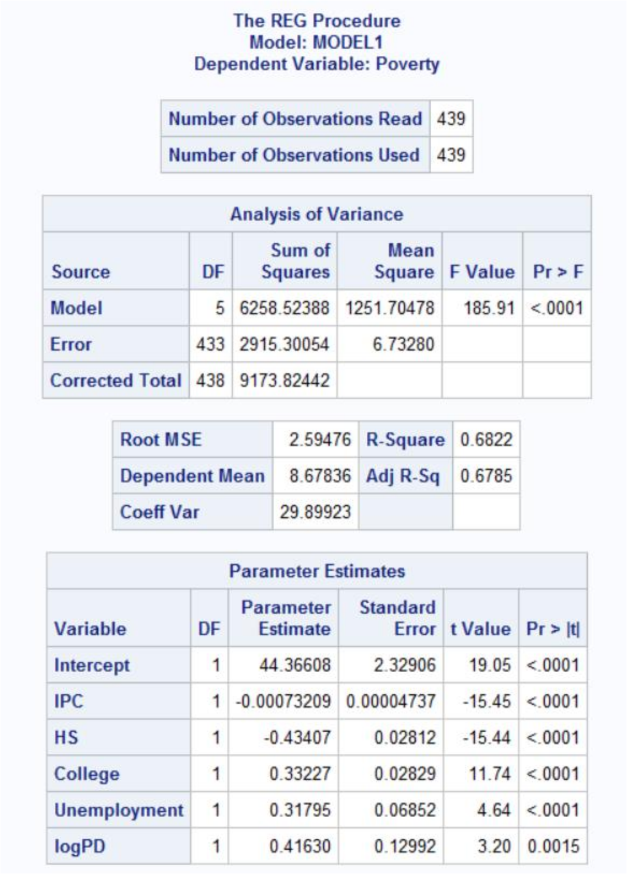


Pearson Correlation Coefficients, N = 440					
	Unemployment	IPC	HS	College	logPD
Unemployment	1.00000	-0.32214	-0.59360	-0.54091	-0.19019
IPC	-0.32214	1.00000	0.52300	0.69536	0.50980
HS	-0.59360	0.52300	1.00000	0.70779	0.11529
College	-0.54091	0.69536	0.70779	1.00000	0.35950
logPD	-0.19019	0.50980	0.11529	0.35950	1.00000



Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	45.02989	2.37683	18.95	<.0001	0
IPC	1	-0.00074528	0.00004835	-15.42	<.0001	2.40193
HS	1	-0.44414	0.02866	-15.50	<.0001	2.52135
College	1	0.34012	0.02887	11.78	<.0001	3.04637
Unemployment	1	0.30859	0.07003	4.41	<.0001	1.67188
logPD	1	0.46422	0.13240	3.51	0.0005	1.47315

Figure 13 (left): Pearson Correlation for Best Model
Figure 14 (right): Variance Inflation for Best Model



The REG Procedure
Model: MODEL1
Dependent Variable: Poverty

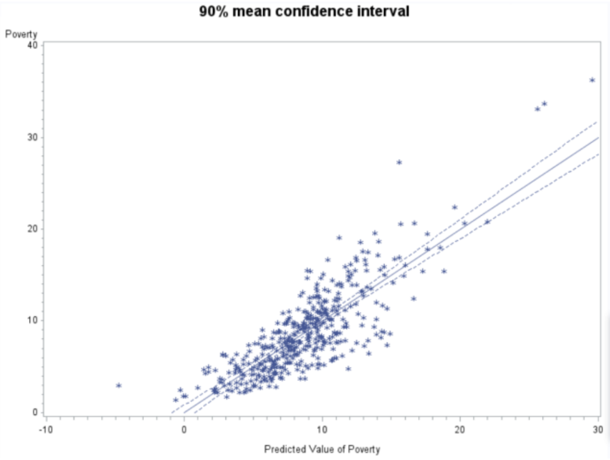
Number of Observations Read	439
Number of Observations Used	439

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	6258.52388	1251.70478	185.91	<.0001
Error	433	2915.30054	6.73280		
Corrected Total	438	9173.82442			

Root MSE	2.59476	R-Square	0.6822
Dependent Mean	8.67836	Adj R-Sq	0.6785
Coeff Var	29.89923		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	44.36608	2.32906	19.05	<.0001
IPC	1	-0.00073209	0.00004737	-15.45	<.0001
HS	1	-0.43407	0.02812	-15.44	<.0001
College	1	0.33227	0.02829	11.74	<.0001
Unemployment	1	0.31795	0.06852	4.64	<.0001
logPD	1	0.41630	0.12992	3.20	0.0015

Figure 15: Regression without Outlier.



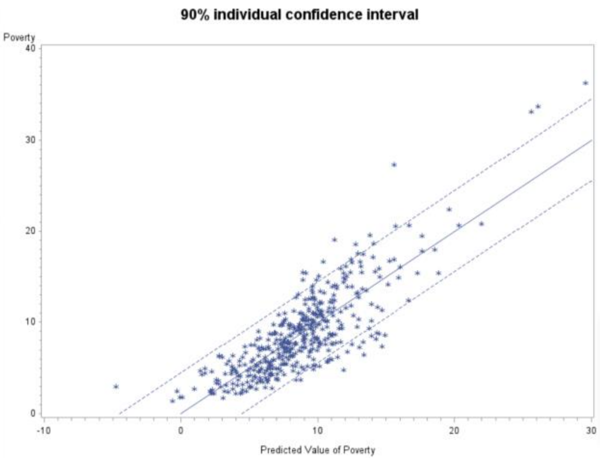
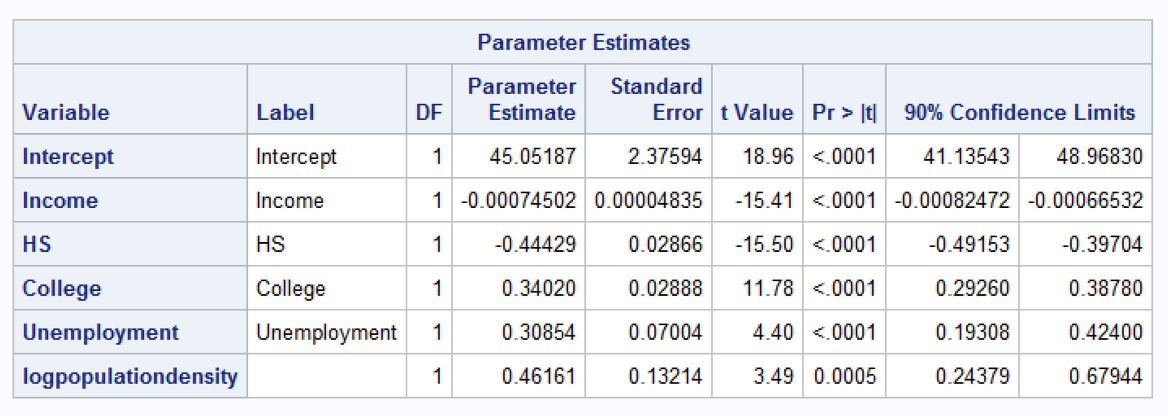
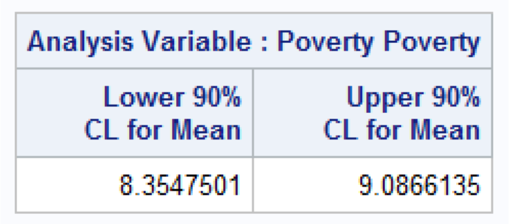


Figure 16 (left): The 90% confidence interval for the mean
Figure 18 (right): 90% CI for Individual points.



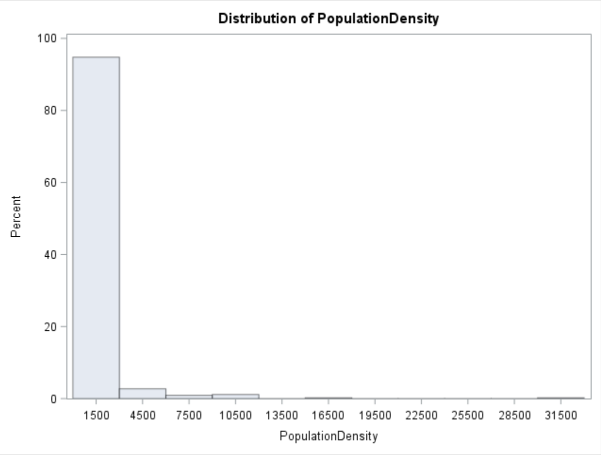
Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	90% Confidence Limits	
Intercept	Intercept	1	45.05187	2.37594	18.96	<.0001	41.13543	48.96830
Income	Income	1	-0.00074502	0.00004835	-15.41	<.0001	-0.00082472	-0.00066532
HS	HS	1	-0.44429	0.02866	-15.50	<.0001	-0.49153	-0.39704
College	College	1	0.34020	0.02888	11.78	<.0001	0.29260	0.38780
Unemployment	Unemployment	1	0.30854	0.07004	4.40	<.0001	0.19308	0.42400
logpopulationdensity		1	0.46161	0.13214	3.49	0.0005	0.24379	0.67944

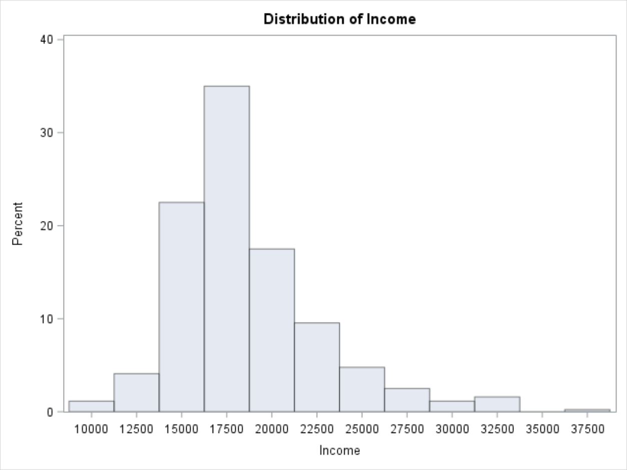
Figure 17: 90% confidence limits for the variables

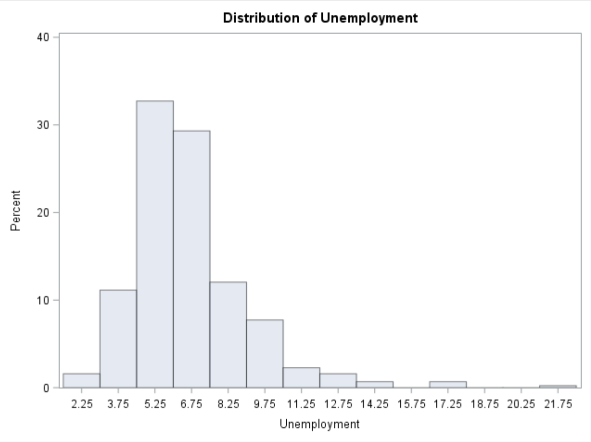


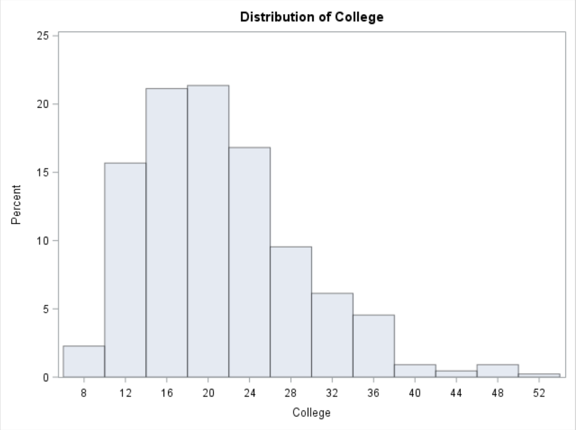
Analysis Variable : Poverty Poverty	
Lower 90% CL for Mean	Upper 90% CL for Mean
8.3547501	9.0866135

Figure 19: 90% CL of the mean for the response variable









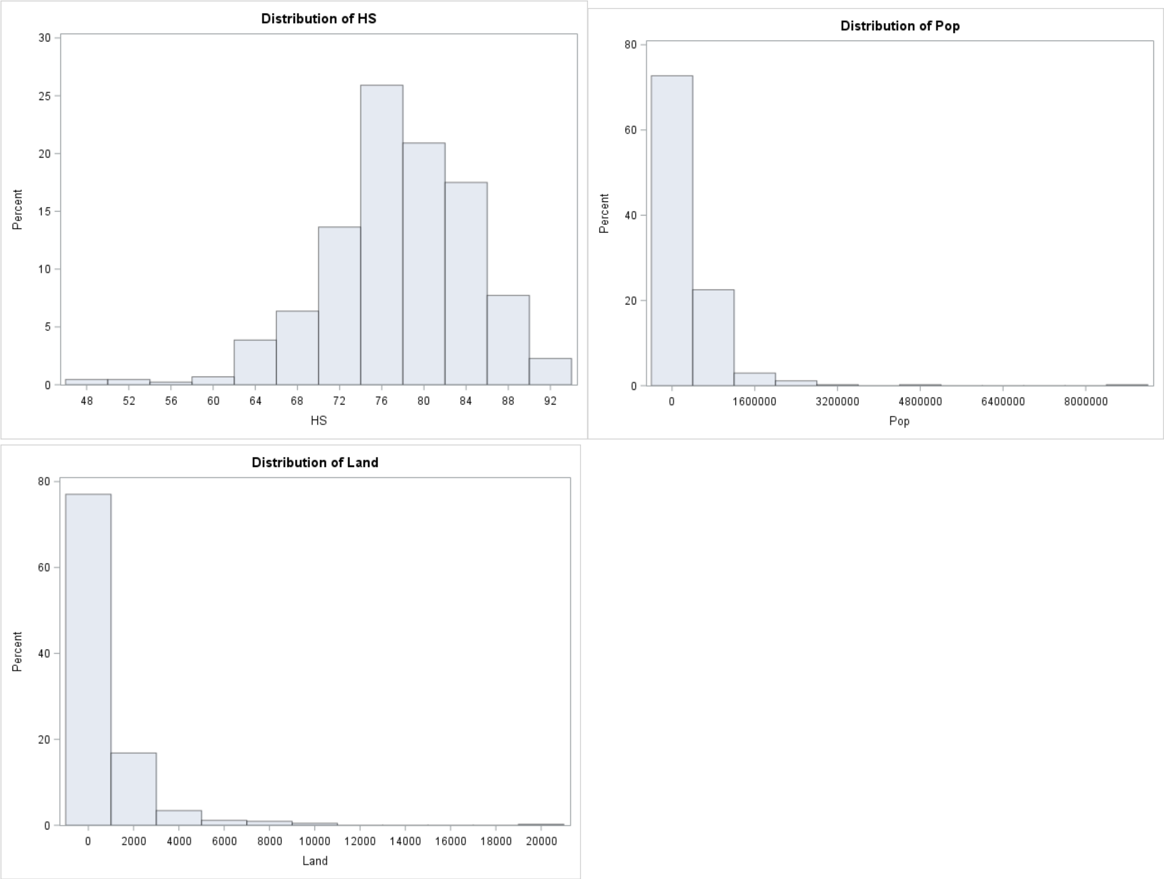


Figure 20: Histograms of each variable to check distribution

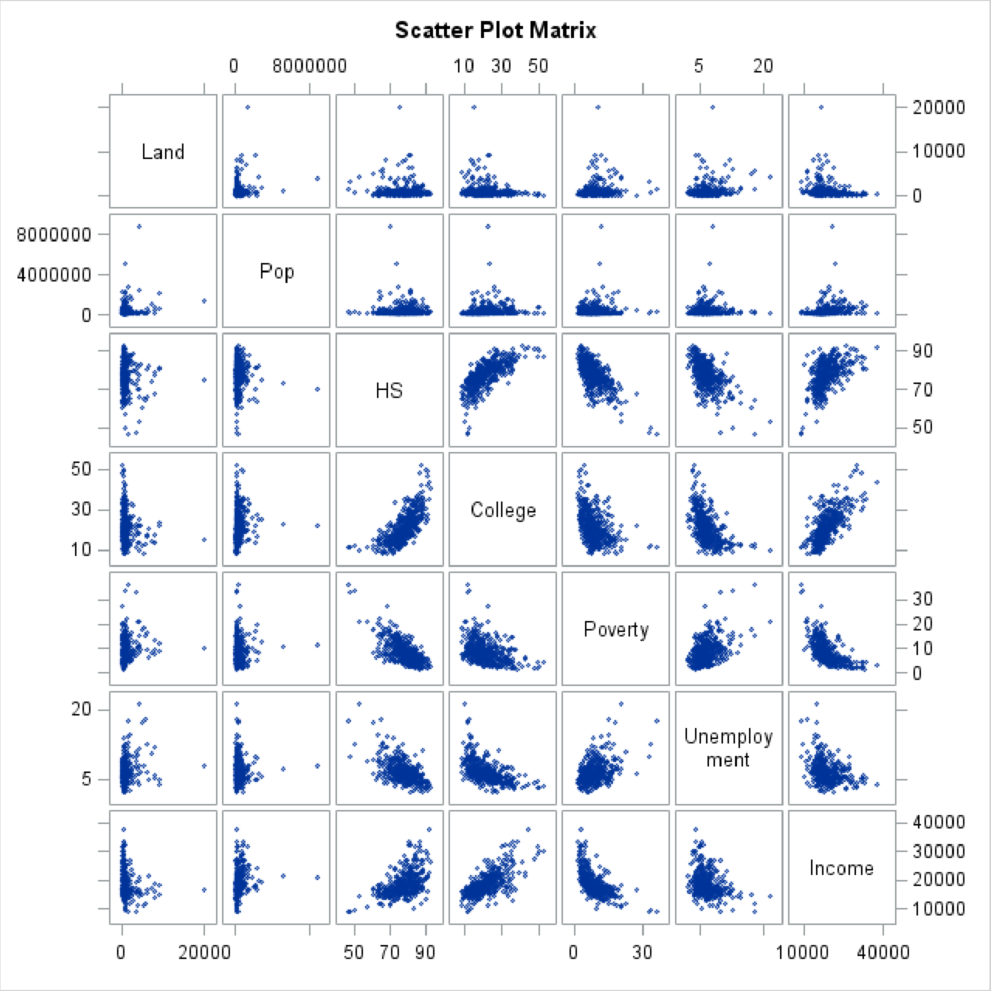


Figure 21: Scatterplot matrix

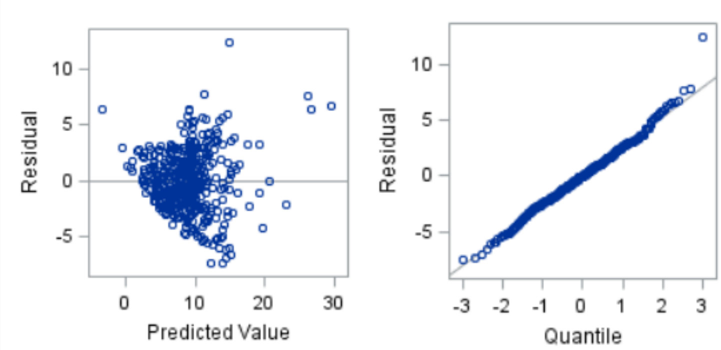
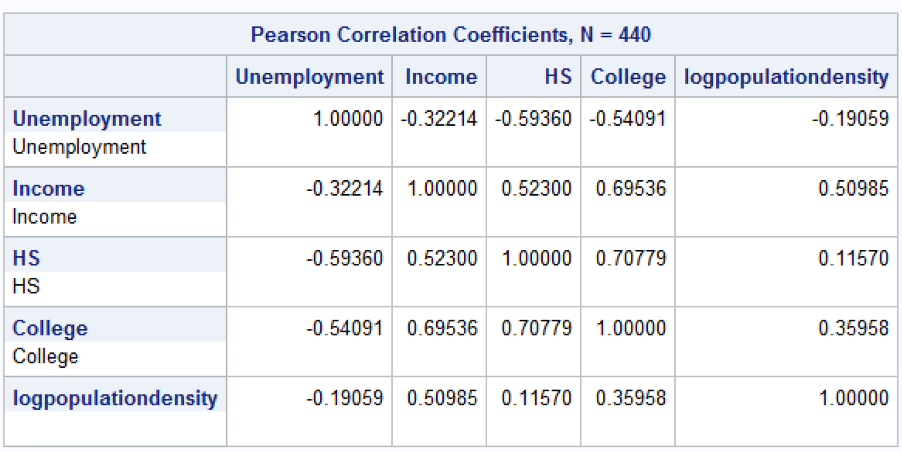


Figure 23(left): Residual Plot for the Best Model.
Figure 24(right): Quantile Plot for the Best Model



Pearson Correlation Coefficients, N = 440					
	Unemployment	Income	HS	College	logpopulationdensity
Unemployment Unemployment	1.00000	-0.32214	-0.59360	-0.54091	-0.19059
Income Income	-0.32214	1.00000	0.52300	0.69536	0.50985
HS HS	-0.59360	0.52300	1.00000	0.70779	0.11570
College College	-0.54091	0.69536	0.70779	1.00000	0.35958
logpopulationdensity	-0.19059	0.50985	0.11570	0.35958	1.00000

Table 5: Pearson Correlation output

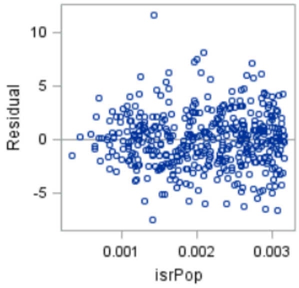
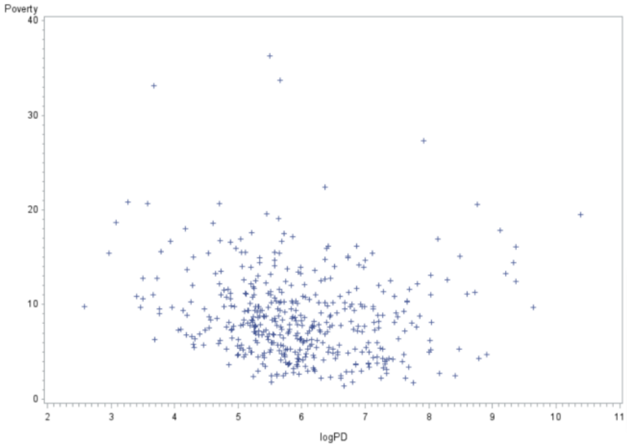
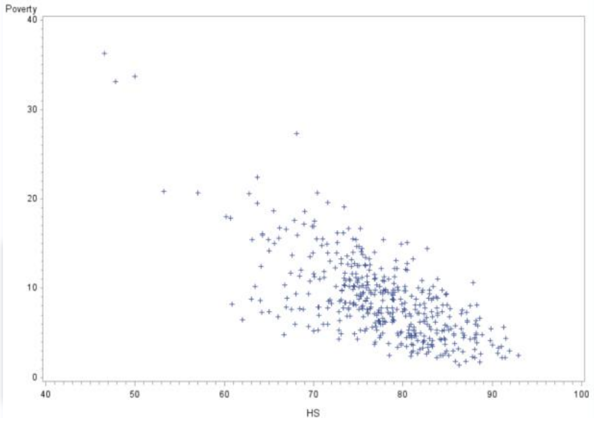
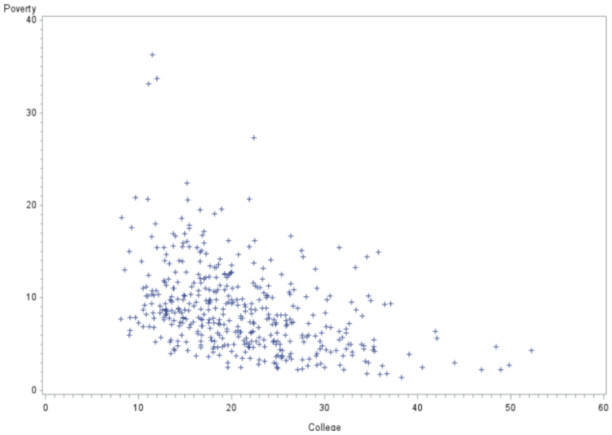
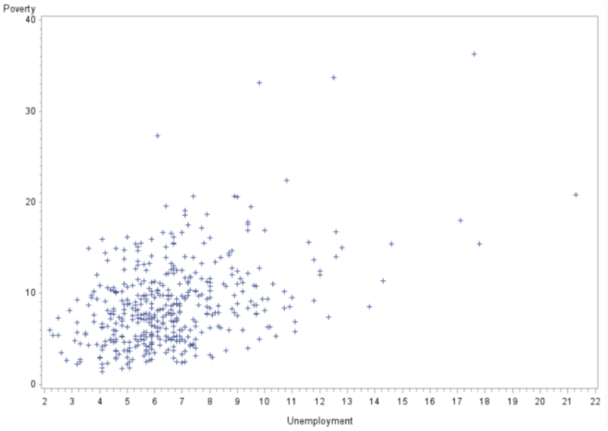


Figure 25: Model Residuals for isrPop









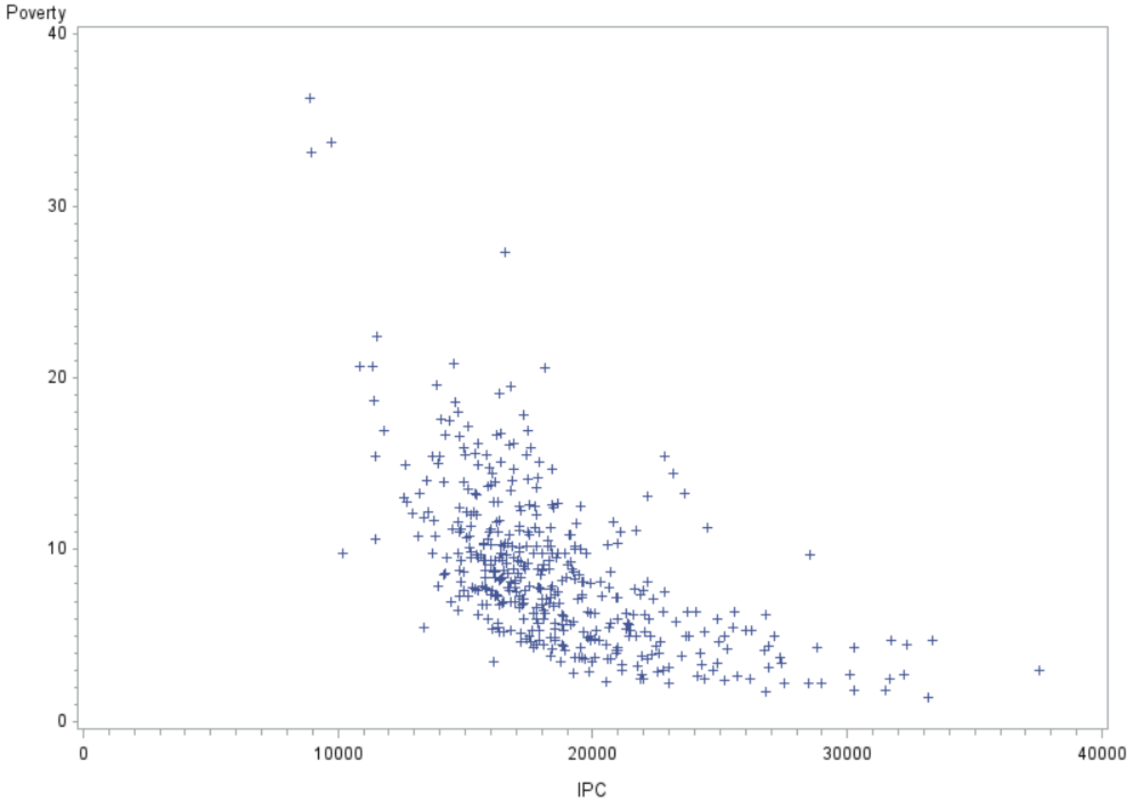
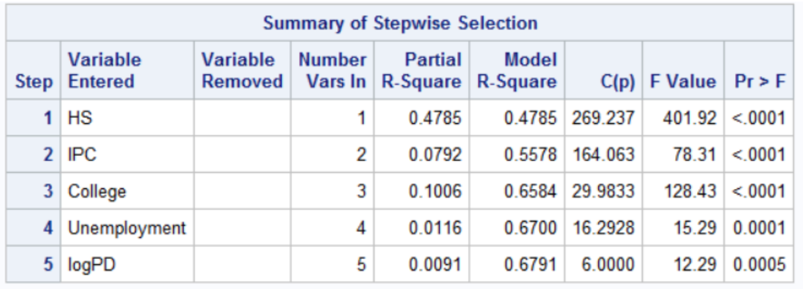


Figure 26: Scatterplots for variables in Final Model. Left to Right, Top to Bottom: logPD, HS, College, Unemployment, IPC



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	HS		1	0.4785	0.4785	269.237	401.92	<.0001
2	IPC		2	0.0792	0.5578	164.063	78.31	<.0001
3	College		3	0.1006	0.6584	29.9833	128.43	<.0001
4	Unemployment		4	0.0116	0.6700	16.2928	15.29	0.0001
5	logPD		5	0.0091	0.6791	6.0000	12.29	0.0005

Table 4: Stepwise Summary