Population Estimates in the United States of America

Present Karmacharya
Justin Solomon

Introduction:

The dataset that was used in this project is that of the population estimates for the counties of each stat in the United States of America. Our 'Response Variable' will be the 'Population Estimates'. Our independent or predictor variables for this dataset are: Births, Deaths, Net migration. The Area_Name, State, nor FIPS are important for analyzing the data but are there for convenience in working with the data.

Fig: Snap shot of the data that will be used for the project

A	B	C	D	E	F	G
FIPS	**State**	**Area_Name**	POP_ESTIMATE_2018	Births_2018	Deaths_2018	NET_MIG_2018
1001	AL	Autauga County	55,601	655	532	35
1003	AL	Baldwin County	218,022	2,254	2,228	5,350
1005	AL	Barbour County	24,881	261	324	-215
1007	AL	Bibb County	22,400	250	256	-148
1009	AL	Blount County	57,840	681	662	-2
1011	AL	Bullock County	10,138	115	120	-32
1013	AL	Butler County	19,680	221	239	-200
1015	AL	Calhoun County	114,277	1,293	1,494	-182
1017	AL	Chambers Count	33,615	357	461	-38
1019	AL	Cherokee County	26,032	219	340	337
1021	AL	Chilton County	44,153	526	506	19
1023	AL	Choctaw County	12,841	134	176	-52
1025	AL	Clarke County	23,920	267	307	-135
1027	AL	Clay County	13,275	128	177	-43

Meaning of Variables:

In order to analyze the data it is required to know the meaning of the variables used. According to the dataset the Population estimate is made on July 1, 2018. The other variables (Births, Deaths, Net migration) are the amount of that variable over the course of the year prior to the population estimate (July 1, 2017 thru June 30, 2018).

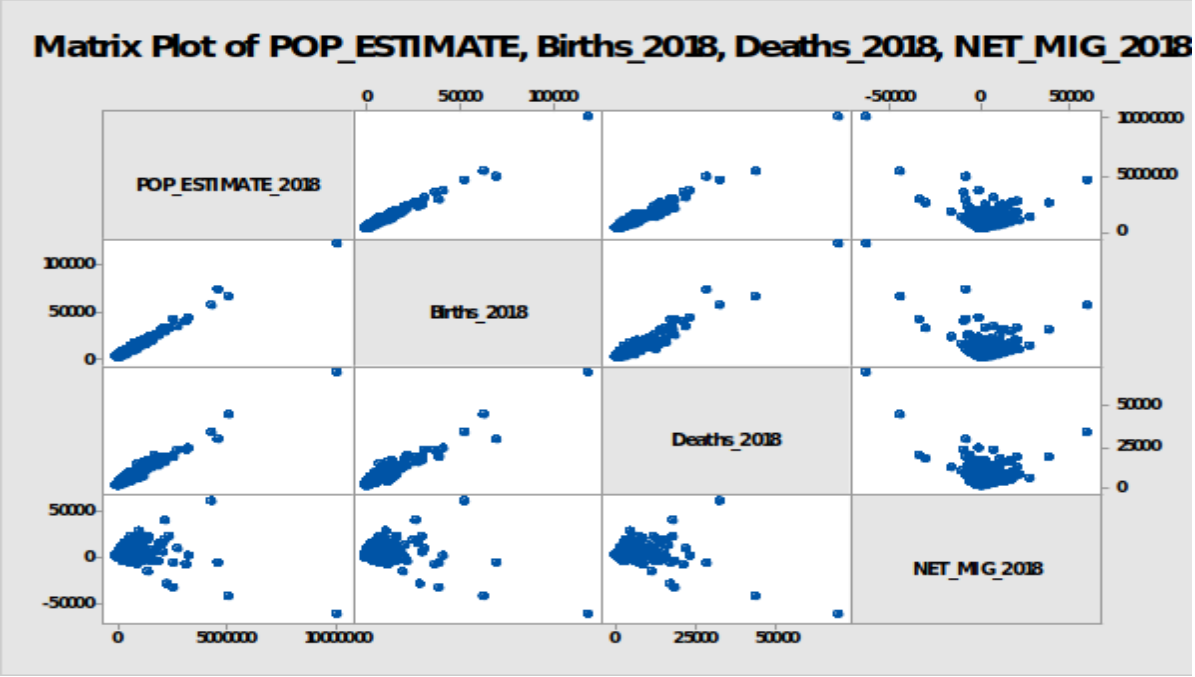
Questions:

The questions that we are trying to address from this dataset are as follows:

- How are the Births, Deaths, and Net migration related with the Population estimates?
- Which Variable from the independent variables seems to be more correlated with the response variable?
- Which county seem to have more response to the population estimates?
- Are the independent variables correlated with each other?

The Original Process:

We started by graphing the data using a matrix plot to see how each variable affects the population estimate and also how the independent variables are correlated. The plot showed some interesting things.



Correlations

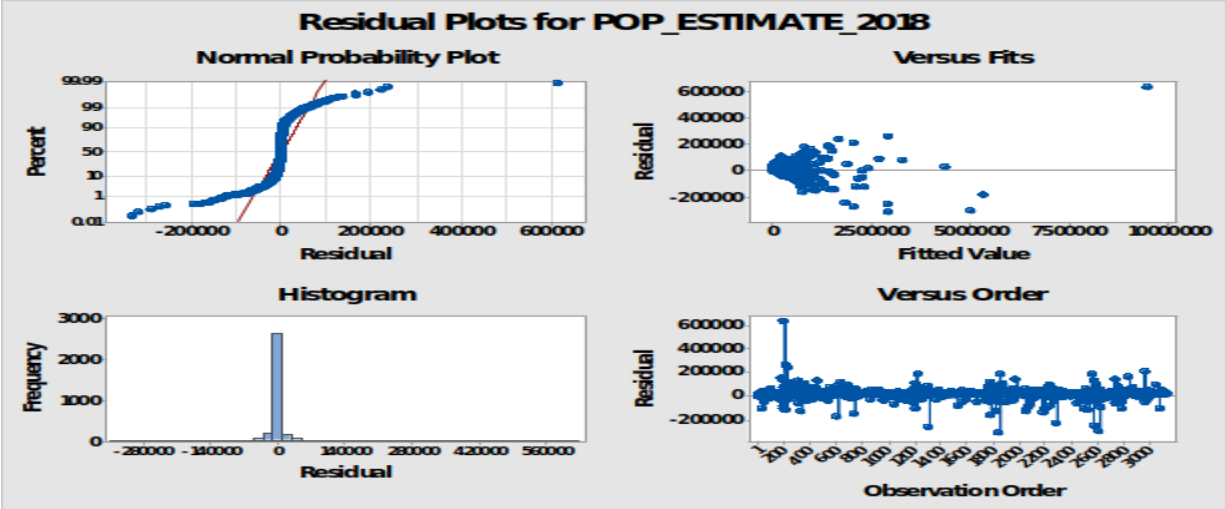
	POP_ESTIMATE_201	Births_2018	Deaths_2018
Births_2018	0.993		
	0.000		
Deaths_2018	0.983	0.968	
	0.000	0.000	
NET_MIG_2018	-0.099	-0.116	-0.089
	0.000	0.000	0.000

Cell Contents
Pearson correlation
P-Value

From the matrix plot we learn that there is obviously some extreme outliers that may cause issues in the analysis. We also see that there may be some multicollinearity issues as well. This is shown in the correlation table above. This caused some concern as we proceed forward with that analysis.

Next we decided to see what would happen if we looked at the correlation matrix and we found that all the p-values where basically zero. This meant that there was a correlation not only between all the independent variables and the response variable, but that there was a correlation between every independent variable as well. We were unable, because of this, to determine which independent variable was most correlated with the response variable. This raised some concern but we proceeded forward with the analysis.

We then tried to run the regression to see what would occur with the analysis. This gave us an interesting understanding of the residuals. From the plots obtained from minitab (shown below) you can see that there is an obvious issue with the assumptions. There is a huge lack of normality and an issue with equal variance.



Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
26506.5	99.37%	99.37%	99.15%

Coefficients

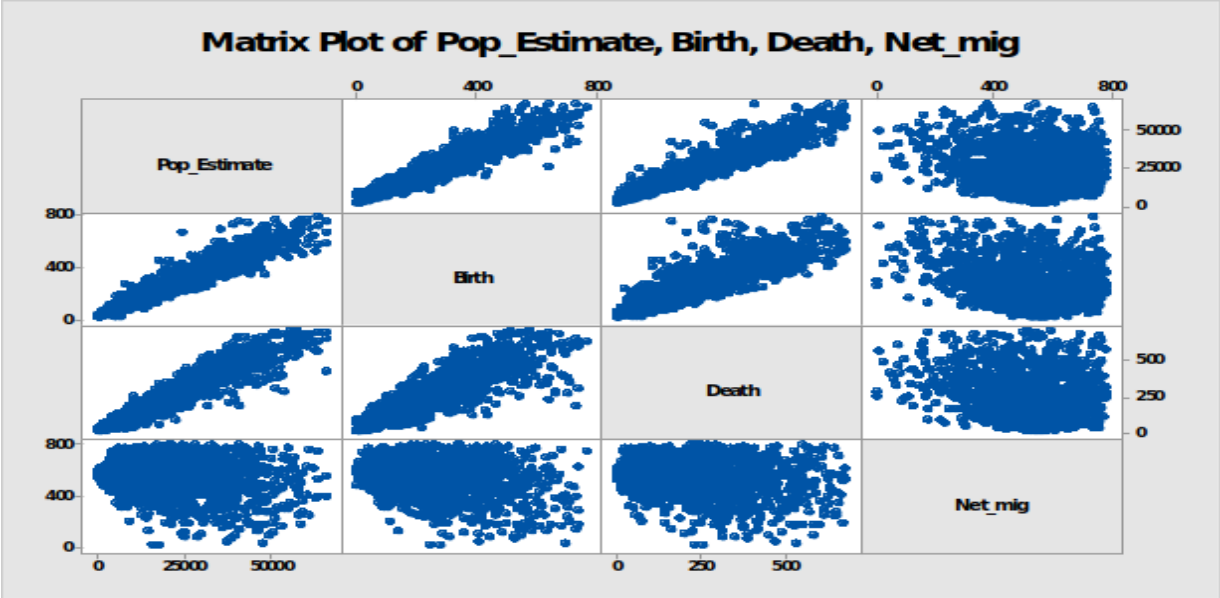
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-3128	516	-6.06	0.000	
Births_2018	53.062	0.461	115.15	0.000	16.24
Deaths_2018	46.731	0.768	60.85	0.000	16.15
NET_MIG_2018	0.934	0.171	5.47	0.000	1.02

Analysis of Variance

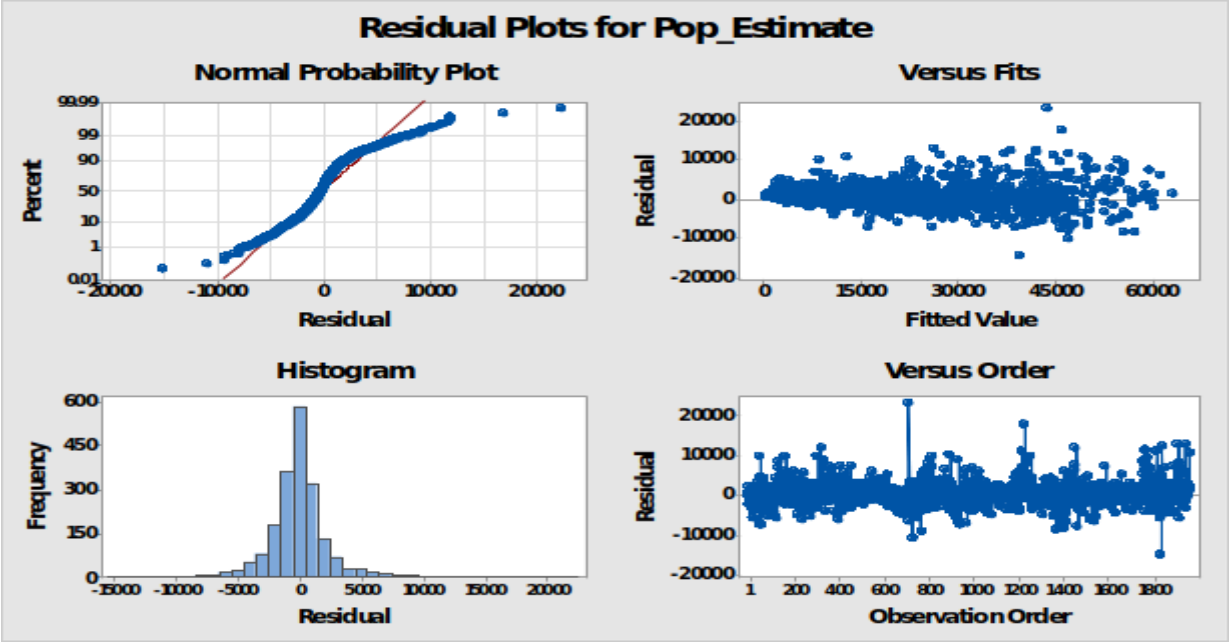
Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	3.47116E+14	1.15705E+14	164683.20	0.000
Births_2018	1	9.31617E+12	9.31617E+12	13259.71	0.000
Deaths_2018	1	2.60181E+12	2.60181E+12	3703.16	0.000
NET_MIG_2018	1	21005914549	21005914549	29.90	0.000
Error	3138	2.20474E+12	702592793		
Total	3141	3.49320E+14			

Notice that our adjacent R square is really good, and the VIF suggests multicollinearity.

The, we began to try to transform our response to normalize the residuals. This was tedious as the data was still showing a lot of outliers and we had to remove outliers just to normalize the data. Thus, we used the process of IQR formula (Q1-1.5*IQR & Q3+1.5IQR). This resulted in losing almost half of the data. Before normalizing the data we saw what had occurred with the removal of the outliers. The following plots are the matrix plot:



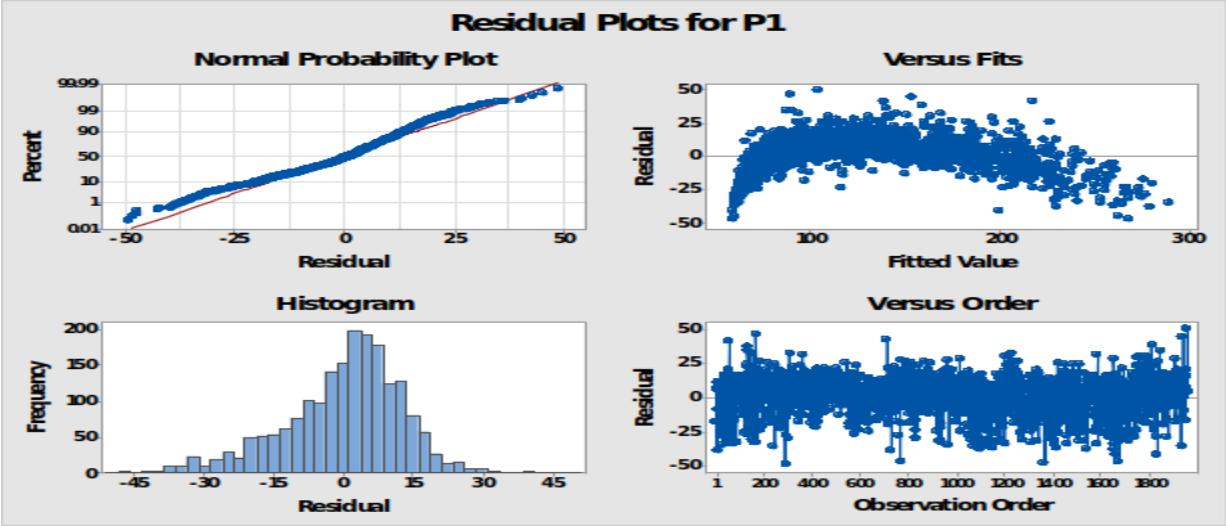
After this we can see the residual plot in the next picture.



These plots show that the data is much closer and seems to be fairly close to equal variance; however, there is still need for more equal variance and still it does not satisfy the condition of normality. The p-value for the normal probability is still less than .005 this is not good enough for the analysis. We realized that we were going to have to do some transformation to fix the data.

After trying many different transformations like box-cox, root y it still wasn't satisfying the normality or equal variance condition, so we used a program in minitab called the Johnson Transformation to transform our response variable to be normal. After the transformation the p-value for the normality of the response variable was .43 which allowed us to assume that the reduced and transformed data was finally normal.

Finally, we ran the regression against this normalized response variable and we got the following residual plots.



We Learn from these plots that the residuals are finally more normalized. As we looked at the variances we say that there was an obvious relation as it was a curve of fairly equal width, suggesting constant variance, but it still had a patter.

We concluded at this point that the data could not be analyzed using the methods learned in our applied regression class or the datasets was not fairly random. It was time to reevaluate our data and move to a new approach.

The Revision Process:

We looked again at our raw data and decided to see what would happen if we tried to analyze the data by state (including the District of Columbia) instead of county. And we also decided to use Births, Deaths, Natural Increase, International Migration, Domestic Migration, and Net Migration. We hoped that this would allow for a better model.

Meaning of Additional Variables:

Just as the original independent variables (Births, Deaths, Net migration), the additional independent variables (Natural Increase, International Migration, Domestic Migration) are the amount of that variable over the course of the year prior to the population estimate (July 1, 2017 thru June 30, 2018).

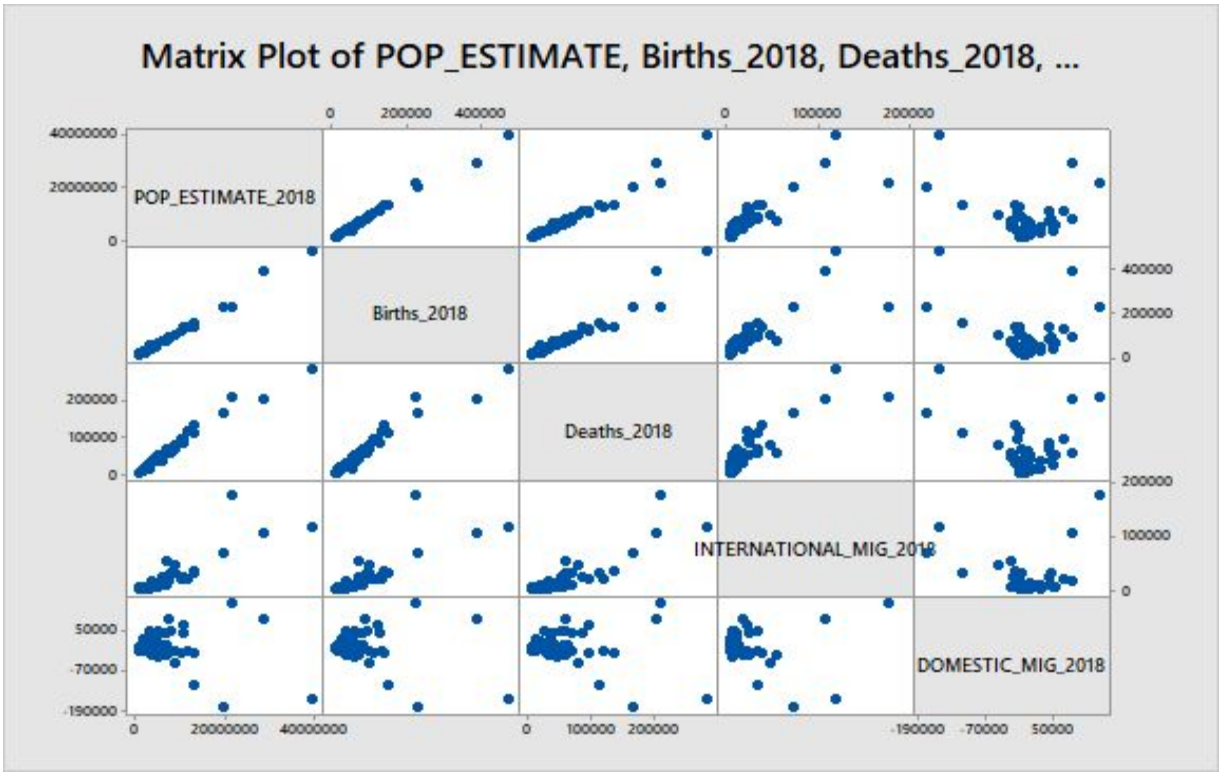
The Revised Questions:

The questions that we are trying to address from this revised dataset are:

- What independent variables relate to the population estimates?
- Is the result different that the county wise data?
- What is the best regression equation for the data?

The Revised Process:

After running a few tests we found that we only needed Births, Deaths, Domestic Migration, and International Migration. The net-migration and natural increase columns was automatically removed by the outliers. We noticed that this could be due to the fact that birth death signifies the natural increase independent variable and domestic migration and internal migration results to net migration. Minitab might have noticed this, and automatically remove them from the data as they won't be independent. We wanted to see how these were related so we used the matrix plot to visualize the data.



We learned from this plot that there was again relationships between the independent variables. So we got a correlation matrix. Only this time not all the values were basically zero.

Correlations

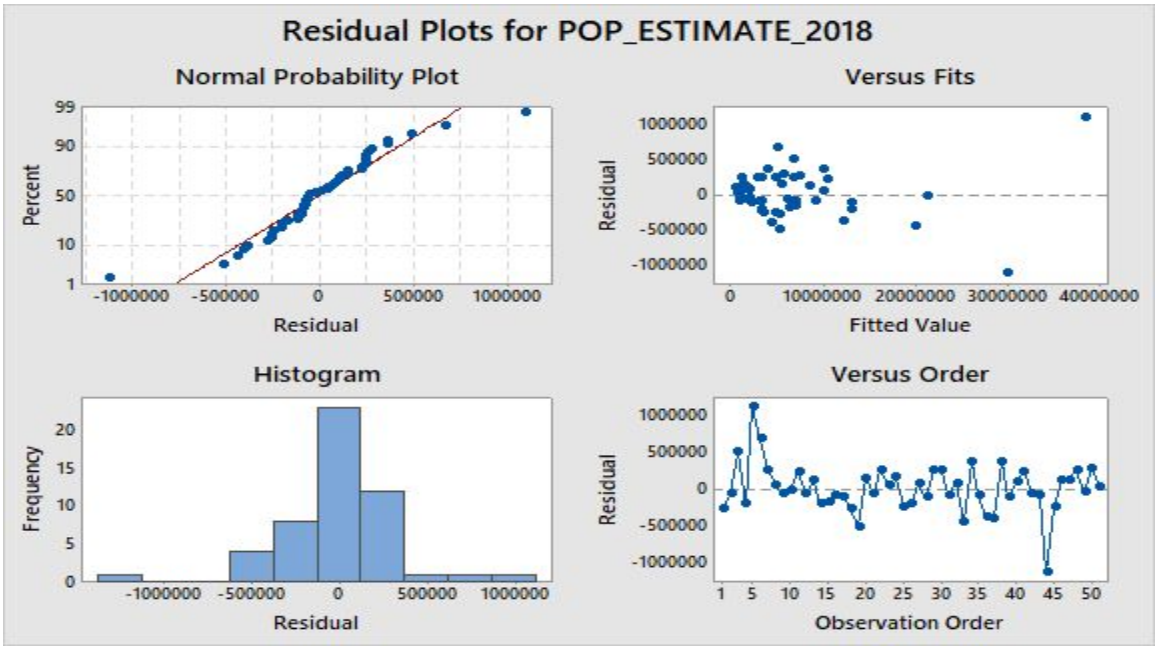
	POP_ESTIMATE_201	Births_2018	Deaths_2018
Births_2018	0.994		
	0.000		
Deaths_2018	0.982	0.962	
	0.000	0.000	
INTERNATIONAL_MI	0.857	0.829	0.874
	0.000	0.000	0.000
DOMESTIC_MIG_201	-0.241	-0.232	-0.213
	0.089	0.102	0.134

From this we learned that domestic migration was not correlated heavily with any of the other independent variables this was an interesting realization.

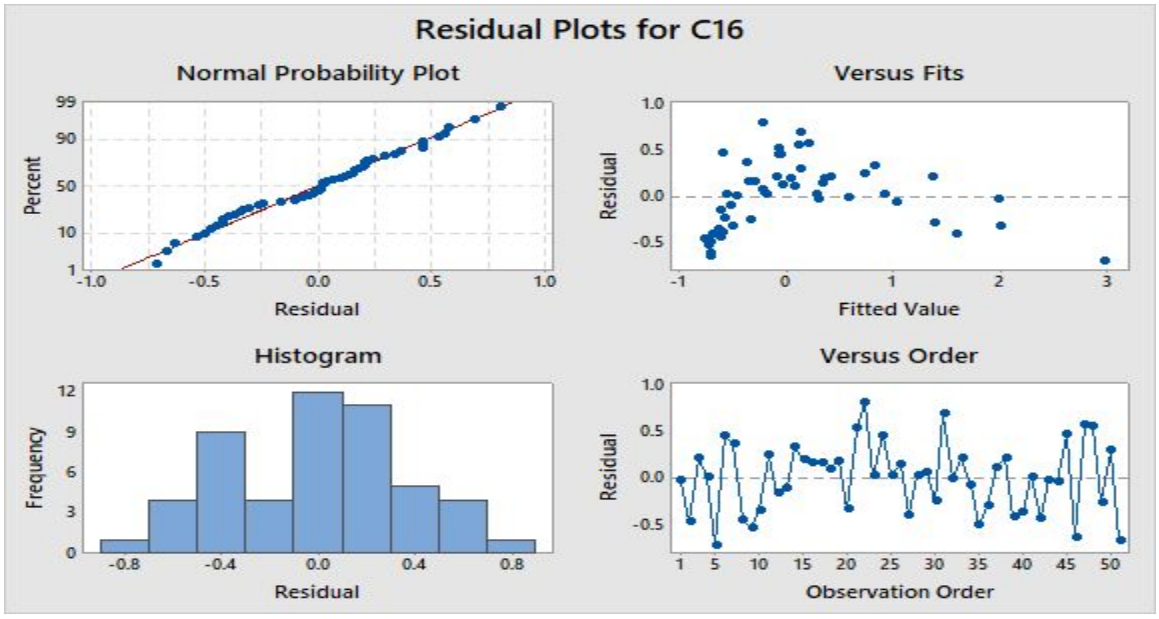
Now that the data was smaller we were able to proceed with the analysis and our knowledge of multicollinearity could possible help us analyze the data. We were hoping

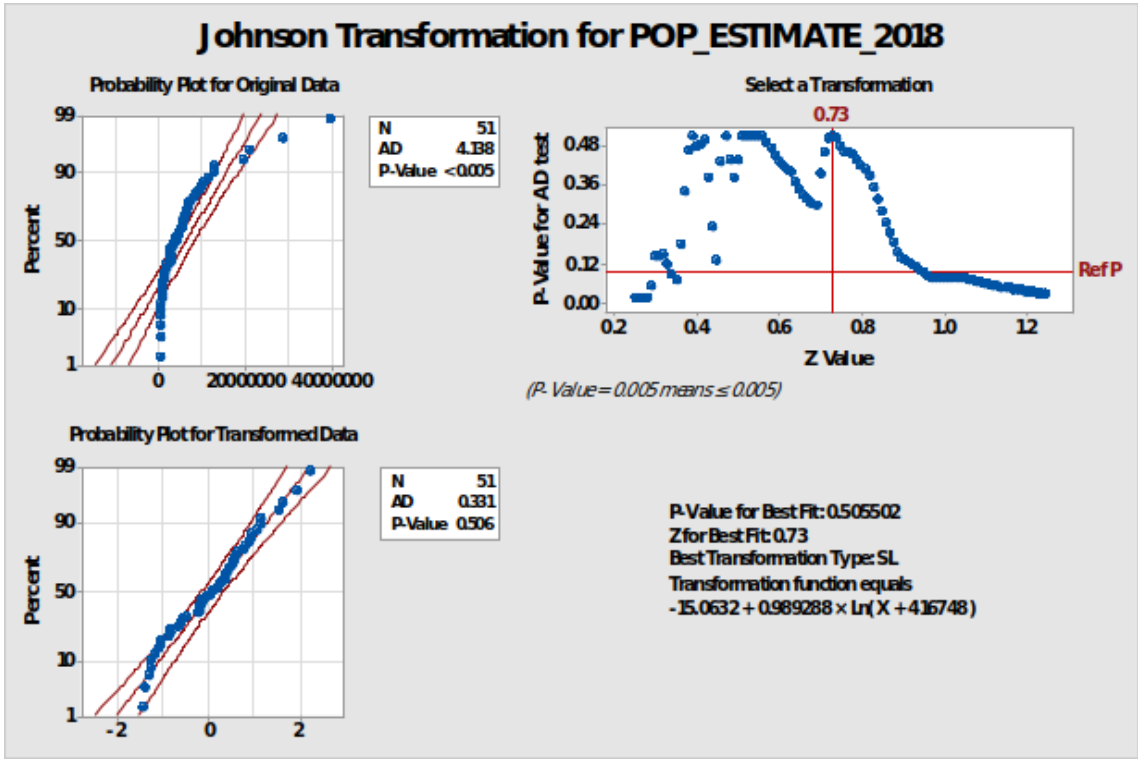
to be able to find a different result then the time before, but knew that it was still possible to come to the same conclusion that we already had.

We obtained the regression analysis of this data. We learned from the residual plots that a transformation was needed yet again. The residuals were far from normal and there was a need to equalize the variances of the residuals.



We tried many different ways to transform our data and finally had to use the Johnson Transformation tool again to normalize our response variable. After the transformation we obtained the following residual plots.


Residual Plots for C16
Normal Probability Plot
Percent
Residual
Versus Fits
Residual
Fitted Value
Histogram
Frequency
Residual
Versus Order
Residual
Observation Order


Johnson Transformation for POP_ESTIMATE_2018
Probability Plot for Original Data
N 51
AD 4.138
P-Value <0.005
Percent
Select a Transformation
0.73
P-Value for AD test
Ref P
Z Value
(P-Value = 0.005 means ≤ 0.005)
Probability Plot for Transformed Data
N 51
AD 0.331
P-Value 0.506
Percent
P-Value for Best Fit: 0.505502
Z for Best Fit: 0.73
Best Transformation Type: SL
Transformation function equals
-15.0632 + 0.989288 × Ln(X + 416748)

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.386492	83.13%	81.66%	56.29%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-0.8122	0.0857	-9.48	0.000	
Births_2018	-0.000003	0.000002	-1.21	0.233	13.71
Deaths_2018	0.000024	0.000004	5.85	0.000	18.37
INTERNATIONAL_MIG_2018	-0.000011	0.000004	-2.96	0.005	4.75
DOMESTIC_MIG_2018	0.000002	0.000001	1.97	0.055	1.18

Regression Equation

C23 = -0.8122 - 0.000003 Births_2018 + 0.000024 Deaths_2018 - 0.000011 INTERNATIONAL_MIG_2018 + 0.000002 DOMESTIC_MIG_2018

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	0.39918	0.09979	5.57	0.001
Births_2018	1	0.03014	0.03014	1.68	0.201
Deaths_2018	1	0.14470	0.14470	8.08	0.007
INTERNATIONAL_MIG_2018	1	0.27959	0.27959	15.60	0.000
DOMESTIC_MIG_2018	1	0.19201	0.19201	10.72	0.002
Error	46	0.82424	0.01792		
Total	50	1.22342			

The residuals were finally normalized however there seemed to be a curve in the varences. Also, notice that the R Square adjacent is significantly lower, but is still good. However, when we ran the Breusch-Pagan Test, it told us that the residuals had constant variance. This lead us to the same conclusion that a different model would

need to be used to interpret this data. Also there is a chance that there is a lack of randomness in the experiment.

The Conclusion:

After working with the data we have come to the conclusion that population estimates cannot be analyzed using these methods learned from our applied regression class. An alternate analysis effort is required to properly analyse the data set.Our normality condition was satisfied, but the constant variance in residual plot had pattern. Further investigation from breusch-pagan test for constant variance suggested that there was a constant variance whereas the residual plot noticeably suggested otherwise.

We realised that this could have happened due to one of the following reasons:

1. The data that was selected required different forms of regression technique that was not taught in the course.
2. The data that was selected could not be fully random or independent
3. The data require more years of information from early 2010 year to till date.

Recommendation:

Although, the regression equation was obtained, it is not very dependent as the constant variance assumption is not satisfied. This might lead to incorrect interpretation of the population estimates. Furthermore, a new approach of regression or more input of data can aid in properly interpreting the population estimates

Work Cited:

https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/