Muuzaani Nkhoma (33.33%)

Shrijit Koirala (33.33%)

Steffie Szi Lin Phang (33.33%)

Regression Analysis of Minnesota  
Census Data 2015: MINNESOTA CHILD POVERTY LEVELS

A Project Report Submitted to Professor Zhang Shiju

of St. Cloud State University.

Class: Applied Regression Methods (STAT 421)

**Abstract**

This project analyzed the Minnesota portion of the US Census Data of 2015. Multiple regression analysis was conducted to predict the level of Child Poverty. Several computational techniques were used for variable selection. Box-Cox transformation was performed in the chosen linear regression model. Several model adequacy checking was undergone to ensure if the assumptions of LINE (Linearity, Independence, Normality, and Equal Variance) in the model hold.

The data had 33 independent variables and 1 chosen prediction variable. It also had

eighty- seven (87) observations representing counties in Minnesota. There was no need for imputation as there was no missing information in the Minnesota table.

**Expectations**

The data includes many demographic and non-demographic variables that would be shown in the table on the next page.

Purpose of the Project:

1. To determine the key factors that contribute most in explaining the level of Child Poverty in the state of Minnesota.

**Data Dictionary.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Description** | **Variable** | **Description** |
| County | County Census ID | ChildPoverty | % of children under poverty level |
| TotalPop | Total Population | Professional | % employed in management, business, science, and arts |
| Men | Number of Men | Service | % employed in service jobs |
| Women | Number of Women | Office | % employed in sales and office jobs |
| Hispanic | % of population that is Hispanic | Construction | % employed in natural resources, construction, and maintenance |
| White | % of population that is white | Production | % employed in production, transportation, and material movement |
| Black | % of population that is black | Drive | % commuting alone in a car, van, or truck |
| Native | % of population that is Native American / Native Alaskan | Carpool | % carpooling in a car, van, or truck |
| Asian | % of population that is Asian | Transit | % commuting on public transportation |
| Pacific | % of population that is Native Hawaiian or Pacific Islander | Walk | % walking to work |
| Citizen | Number of Citizens | OtherTransp | % commuting via other means |
| Income | Median Household Income ($) | WorkAtHome | % working at home |
| IncomeErr | Median household Income error ($) | MeanCommute | Mean commute time (minutes) |
| IncomePerCap | Income per capita ($) | Employed | % employed (16+) |
| IncomePerCapErr | Income per capita error ($) | PrivateWork | % employed in private industry |
| Poverty | % under poverty level | PublicWork | % employed in public jobs |
| SelfEmployed | % under poverty level | FamilyWork | % in unpaid family work |
| Unemployment | Unemployment rate (%) |  |  |

**Linear Regression Model.**

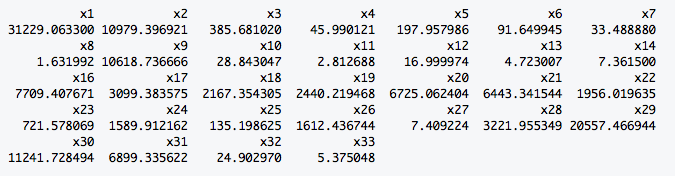
1. **Full Model**

A correlation matrix was produced to check for potential relationships between variables. There were some variables which had linear relationships which is an early indication of potential multicollinearity.

A full model including all regressors was fitted to the data. Even though overall the full model had an R-square of 87.25% and F-statistic of 11.55 on 32 and 54 degrees of freedom with a very small p-value, using the t-test only 4 out of the 33 regressors significantly contributed to the model and had a p-value of less than 0.05. Further model adequacy checking was performed.

|  |  |  |
| --- | --- | --- |
| AIC | BIC | PRESS |
| 440.4484 | 524.2893 | 1186.997 |

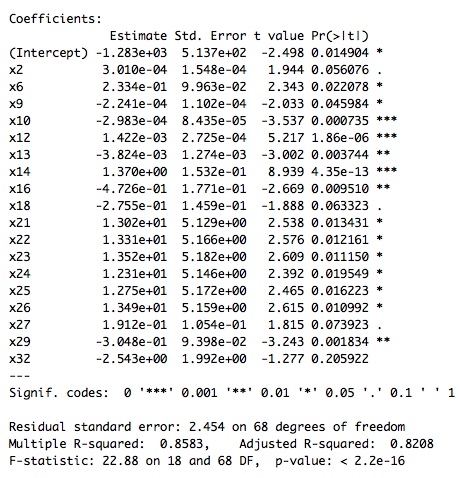
We then checked the model for multicollinearity using the VIF. Most of the variables had a VIF value of more than 10, this is an indication of multicollinearity. Thus, it was concluded that this model is not adequate.



Further model building for variable selection process was performed to obtain a better model.

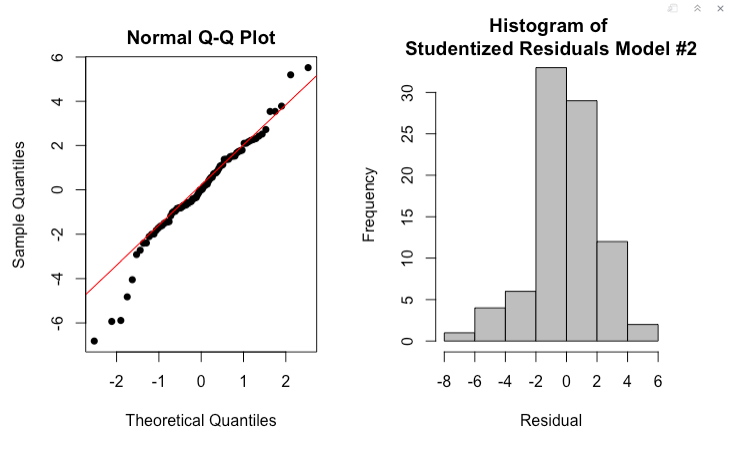
1. **Stepwise Selection.**

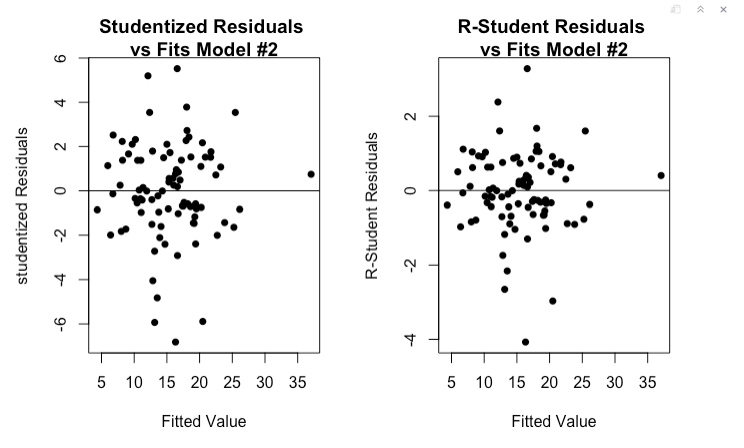
This model chose 18 regressors out of the 33. Using the t-test only 15 of the regressors were significant with a p-value of less than 0.05. This model has an Adjusted R-Squared of 82.08% which means 82% of the variability is explained by this model.



**Residual Analysis.**

A residual and normality analysis was performed. According to the normal Q-Q Plot and the Histogram, it does not indicate much of a problem except skewness to the left and a spike in the middle. The residual plots look appropriate with perhaps several outliers at the bottom.

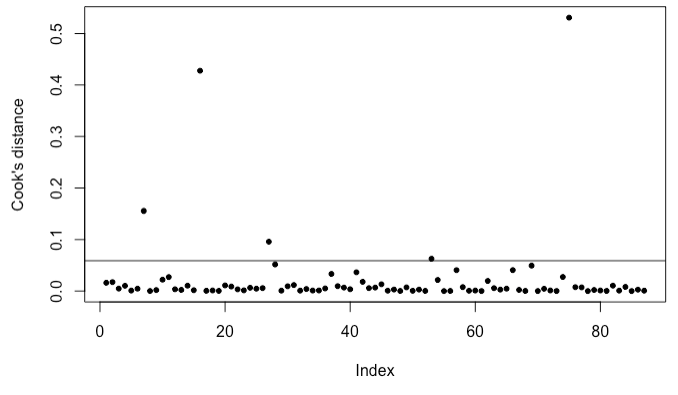
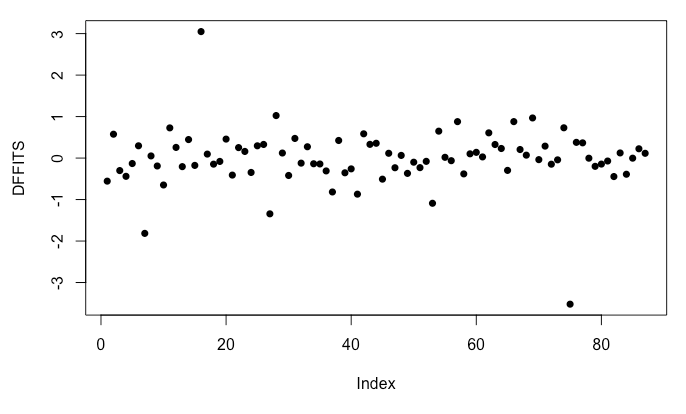




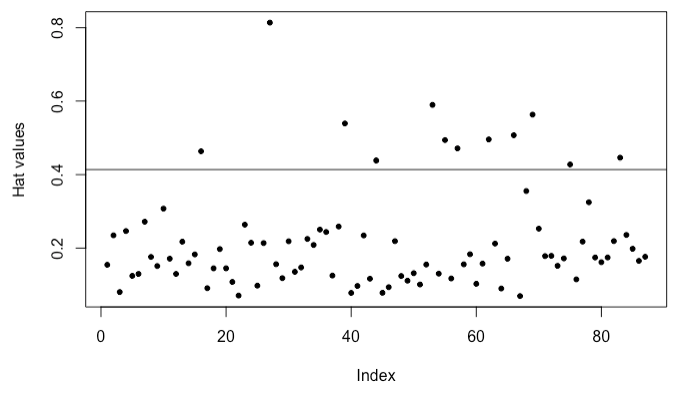
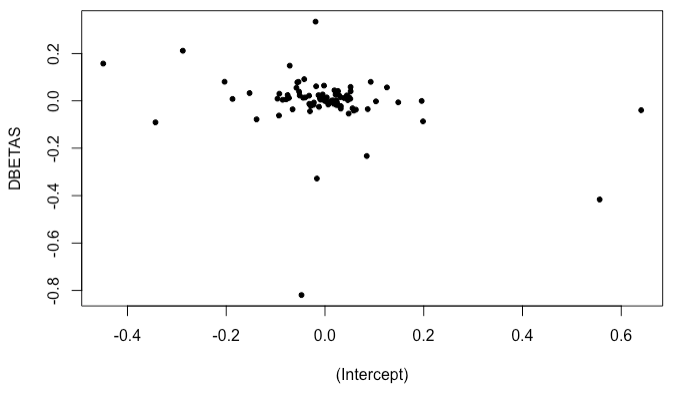
**Influential Analysis.**

An influential analysis was performed by looking at Cook’s D, DFFITS, DFBETA and leverage. According to DFFITS and DFBETAS there were no points that exceeded the cutoff value. The Cook’s D showed that there were 5 influential points.

DFFITS Cook’s D

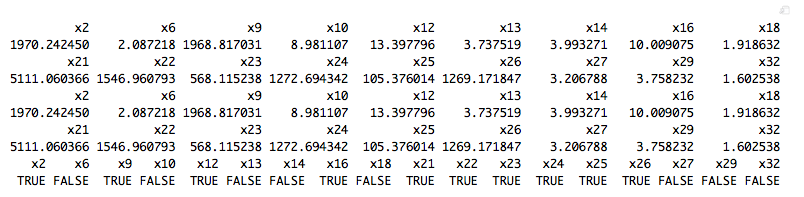


DFBETAS Hat Value



**Multicollinearity Check.**

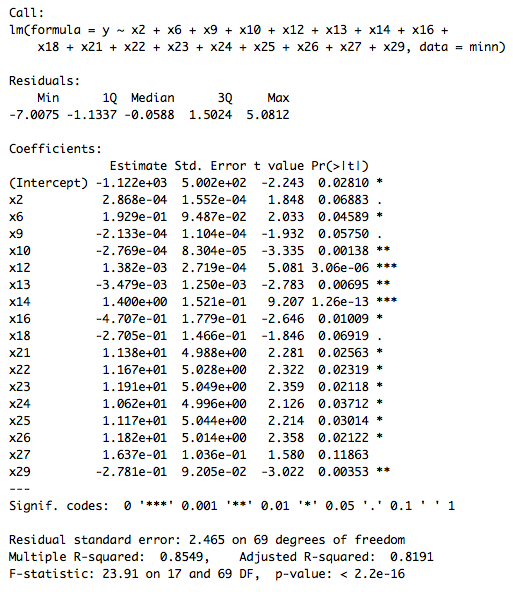
According to the output obtained from R, we see most of the variables have a VIF of more than 10. This is an indication of multicollinearity. Thus, it is concluded that this model is not adequate.



1. **Backward Selection Model.**

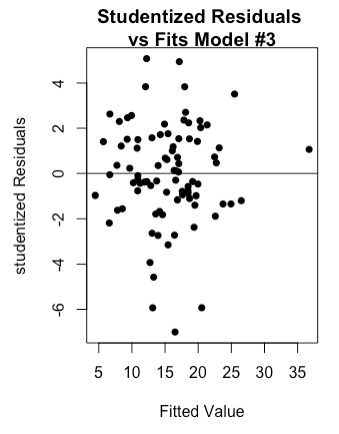
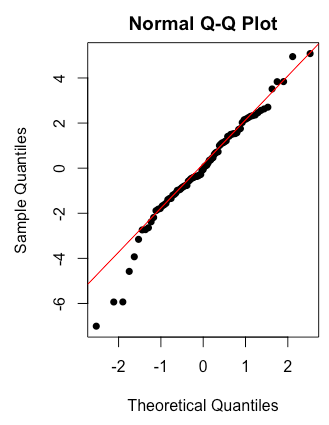
This model chose 17 regressors out of the 33. Using the t-test only 14 of the regressors

were significant with a p-value of less than 0.05. This model has an Adjusted R-Squared of 81.91% which means 82% of the variability is explained by this model, almost similar to the stepwise model.



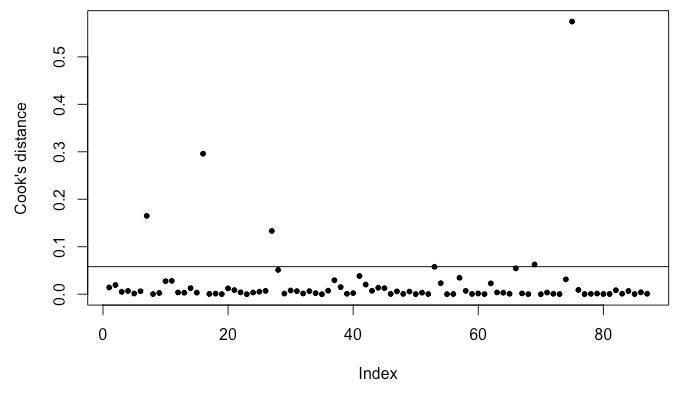
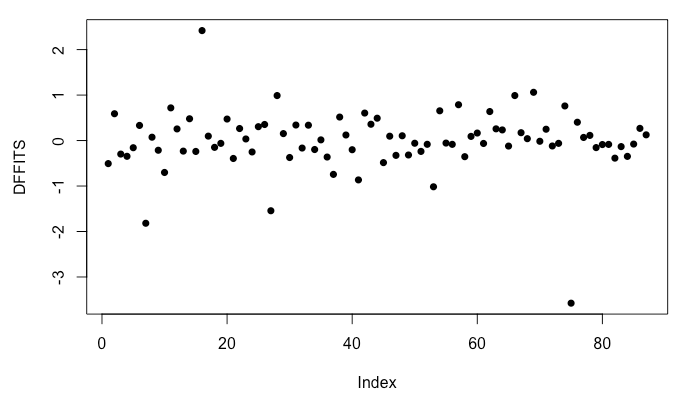
**Residual Analysis.**

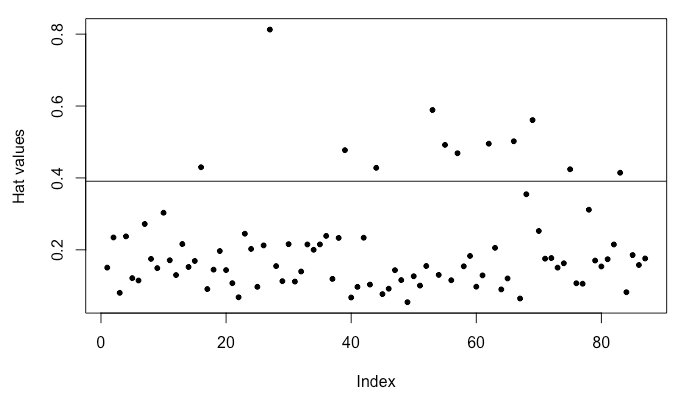
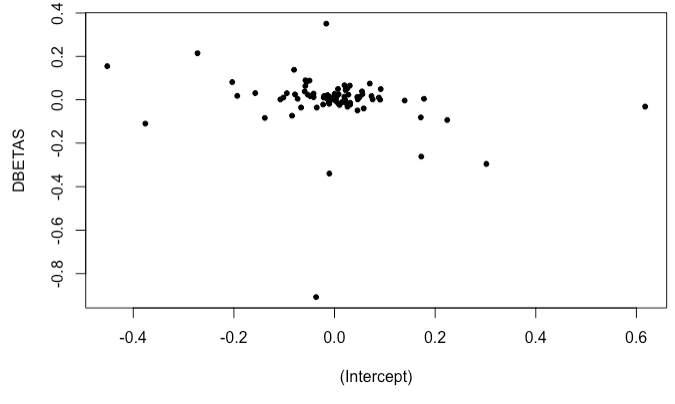
A residual and normality analysis was performed. According to the normal Q-Q Plot, it does not indicate much of a problem in normality except skewness to the left. The residual plots look appropriate with perhaps several outliers at the bottom. Like the stepwise selection model.



**Influential Analysis.**

An influential analysis was performed by looking at Cook’s D, DFFITS, DFBETA and leverage. According to to DFFITS and DFBETAS there were no points that exceeded the cutoff value. The Cook’s D showed that there were 4 influential points.



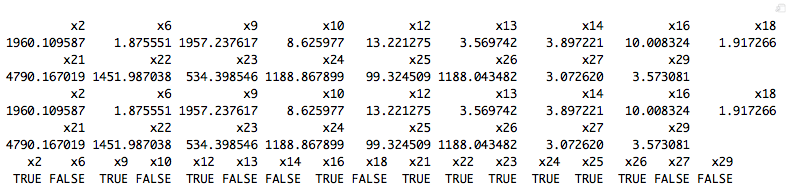


**Multicollinearity Check.**

According to the output obtained from R, we see most of the variables have a VIF of more

than 10. This is an indication of multicollinearity. Thus, it is concluded that this model is

not adequate.



It is concluded that both Stepwise Model and Backwards Model is not adequate due to its

severe multicollinearity. The all possible model method was then performed to find a better model.

**All Possible Regression Model.**

The all possible regression model method was performed in R.

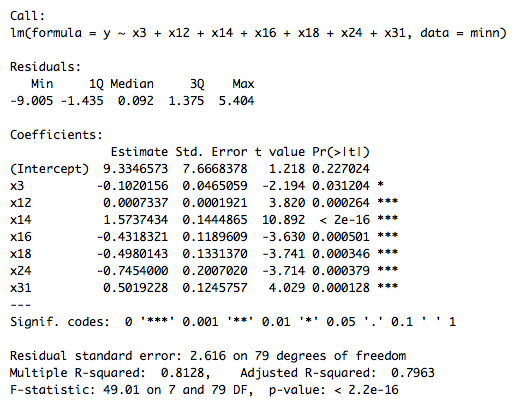
The top five performing models were selected from each of the selection criterion:

* Adjusted R-Squared,
* Residual Sum of Squares (RSS),
* Bayesian Information Criteria (BIC),
* R-Squared, and
* Mallow’s Cp statistic.

The 7 regressor model was chosen because it is a simpler model and it performed better in most of the model selection criteria, especially as it was also the only model that had its Cp statistic closer to the number of regressors plus intercept.

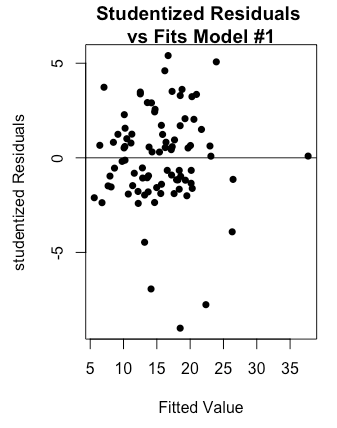
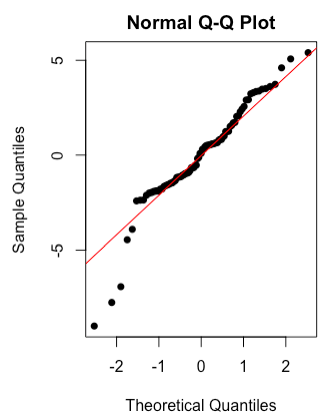
`According to the output, using the t-test all 7 regressors are significant as they have a p-value of less than 0.05. This model has an Adjusted R-Squared of 0.7963, which means that 80% of the variability is explained by the model. It is a 2% difference from the Stepwise and Backward Models, but this model is simpler and performs better on PRESS and VIF’s.

Further analysis was then done on this model to check for its adequacy.



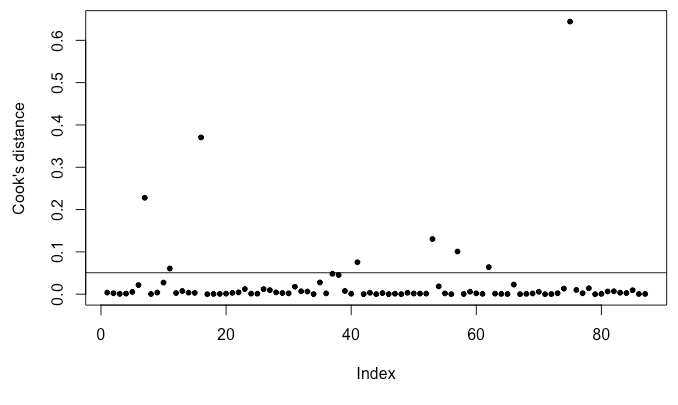
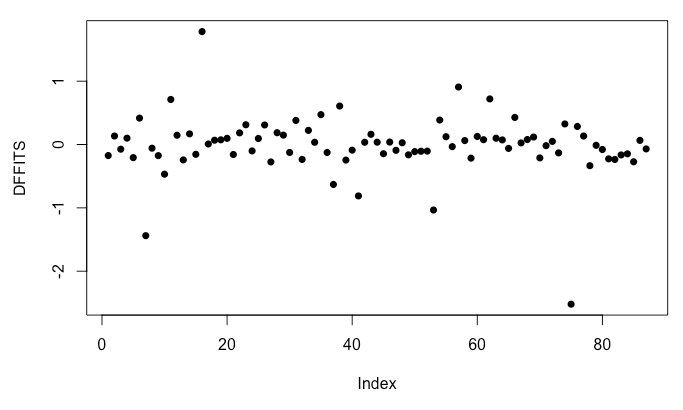
**Residual Analysis.**

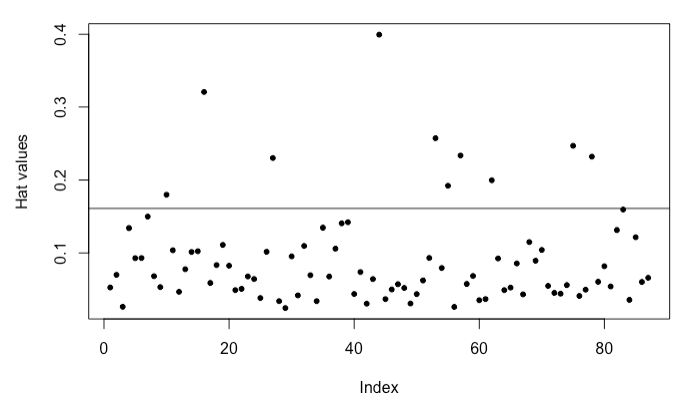
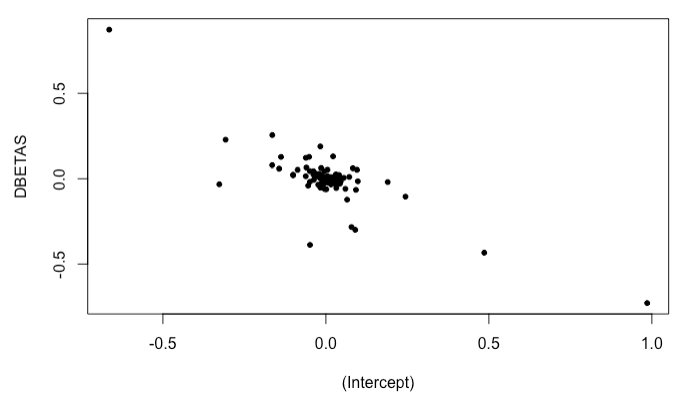
A residual and normality analysis was performed. According to the normal Q-Q Plot, it does not indicate much of a problem in normality except skewness to the left. The residual plots look appropriate with perhaps several outliers at the bottom similar to the stepwise & backward selection model.



**Influential Analysis.**

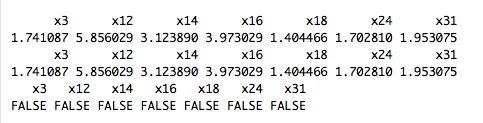
An influential analysis was performed by looking at Cook’s D, DFFITS, DFBETA and leverage. According to DFFITS and DFBETAS there were no points that exceeded the cutoff value. The Cook’s D showed that there were 7 influential points.





**Multicollinearity Check.**

According to the output obtained from R, we see all of the variables have a VIF of less than 10. This does not indicate any signs of multicollinearity. Thus, it is concluded that this model is adequate.

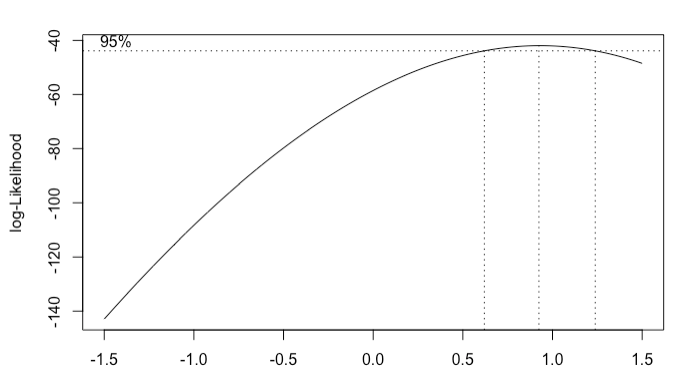
****

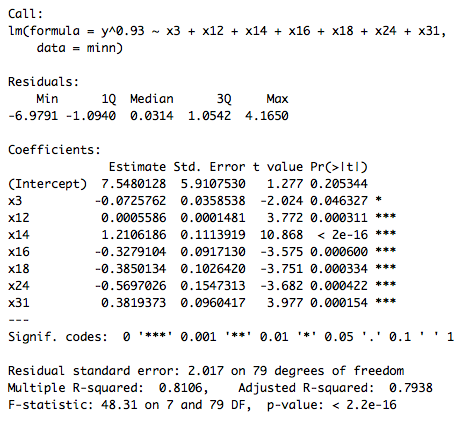
This model is chosen since it is both simpler and concluded to be adequate using the VIF test.

However, there were still skewness in the normality plot and several outliers in the residual plot. A transformation was considered to fine tune the model.

**Transformed Model.**

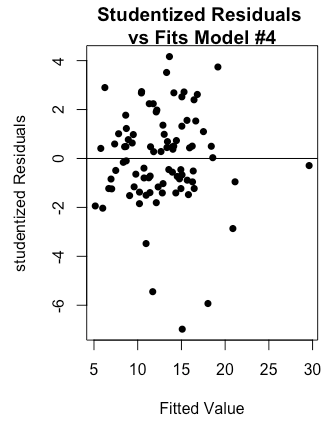
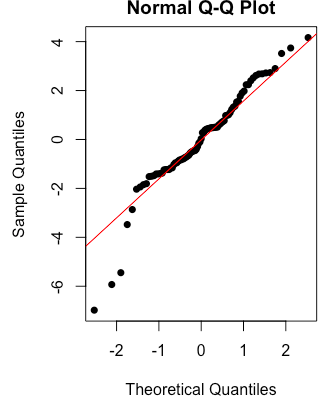
The boxcox transformation was used to make the model better. We obtained a lambda value of 0.93 which is then applied to the 7 regressor model chosen by the all possible method.



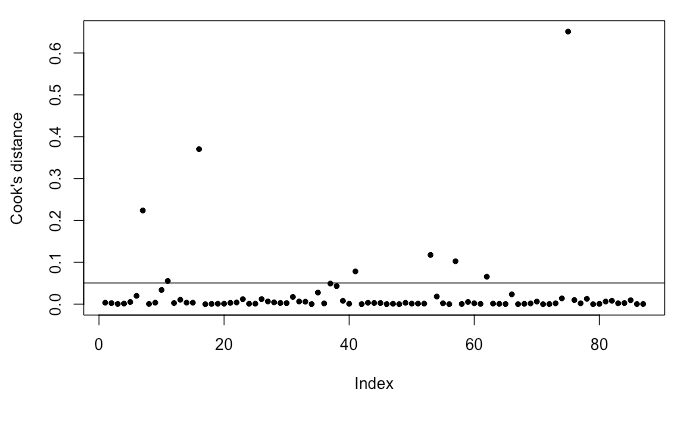


**Residual Analysis.**

A residual and normality analysis was performed. According to the normal Q-Q Plot, it does not indicate much of a problem in normality except skewness to the left. The residual plots look appropriate with perhaps several outliers at the bottom like the stepwise & backward selection model. Even after transforming the model, the problem with normality was not resolved, we suspect that it may be due to the influential points.



**Influential Analysis.**

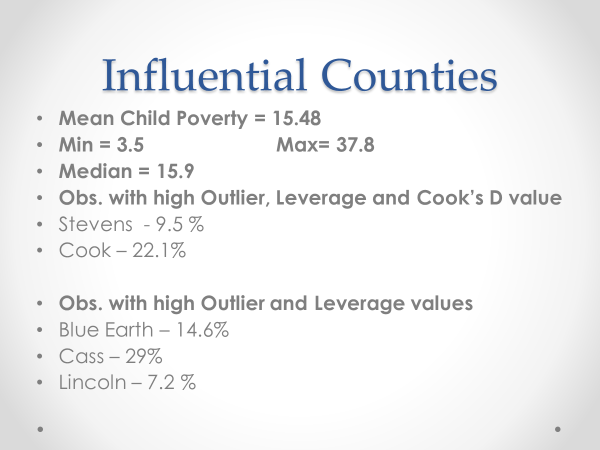
Using both Cook’s D graph above, there are 3 obvious influential points and was determined to be observations 7,16 and 75. These points were then excluded to test if our suspicions were right.

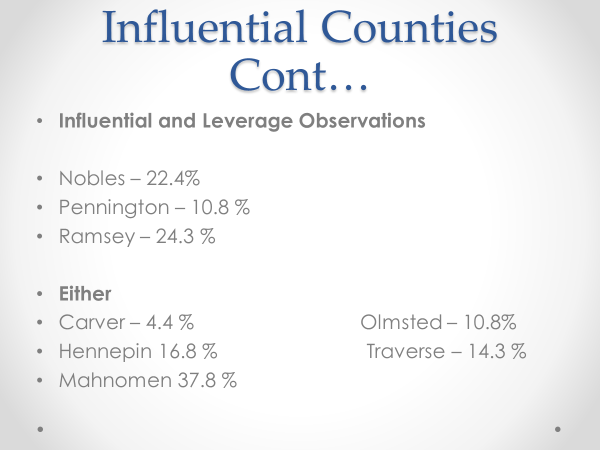
**Normal Quantile Plot and Residual Histogram after removing infl. Points.**

The residual histogram is now normal without the spike in the middle and the points are within boundaries of the normal quantile plot. However, these influential points cannot be removed because there were no errors in the observations.

Further investigation was done on the influential points to try to determine their nature. The influential points are counties which of mixed characteristics. But one important point was that the total population of these counties is 39% of the total population of Minnesota and that three (3) of the counties namely Hennepin (1,197,776), Ramsey (527,411), and Olmsted (148,736) are among the top ten highly populated counties in Minnesota.





Therefore, these influential points were still included in our final model.

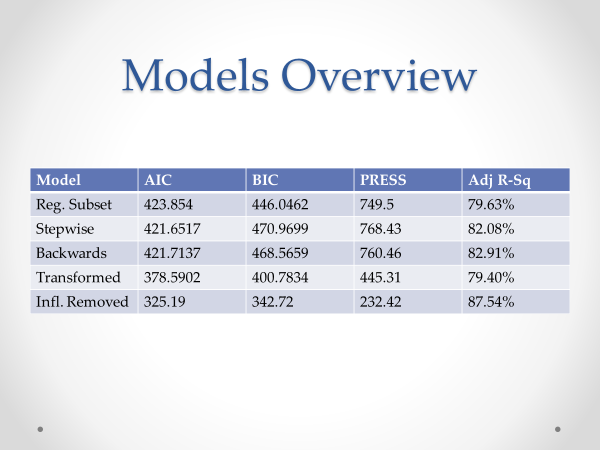
**Final Regression Model:**

Hence after using different techniques of regression analysis, the best model to predict child poverty was:

Child Poverty = 9.33 – 0.102\*White + 0.00073\*IncomePerCapita + 1.573\*Poverty – 0.43\*Professional -0.5\*Office -0.75\*Walk + 0.50\*SelfEmployed

We can see the relationship between Child Poverty and different explanatory variables. According to the equation, Child Poverty has a negative relationship with the variable White, provided that all other explanatory variables in the model are held constant. Likewise, it has a positive relationship with IncomePerCapita, Poverty and SelfEmployed in the equation with Office and Walk having a negative relationship.

Comparing this relationship to the correlation matrix in page four, two variables have a contradictory relationship compared to the regression equation. The two variables are IncomePerCap and Walk.



**Exploratory Analysis of Child Poverty Vs. Explanatory variables**

***Child Poverty Vs. White***

In Minnesota, Child Poverty seems to be low in areas that have high White population. Out of the six demographic variables in the dataset, only White was significant. The coefficient for White in the regression equation was -0.102, which shows a negative relationship. The correlation was -0.4560, which shows a negative relationship as well.

****

***Child Poverty Vs. Poverty***

As expected the relationship between Child Poverty and Poverty was positive in both regression equation and correlation matrix. It has the largest coefficient (1.573), with a correlation of 0.8. It is undoubtedly one of the most important variables in determining Child Poverty. The graph below looks fairly linear. It was also the variable with the lowest p-value.



***Child Poverty Vs. IncomePerCap***

The correlation between Child Poverty and IncomePerCap is -0.4881, whereas the relationship between the two is positive in the regression equation. Looking further into this it seems that the Child Poverty is less in counties where the income per capita (IncPerCap) is higher, the graph of which is shown in the next page.



As the relationship between the two variables is positive in the linear regression model, it gives an indication that Child Poverty is higher in areas with higher income per capita, which contradicts logic. However, in order to explain this situation, we need to understand that this is quite possible in multiple regression. In the regression equation Child Poverty is positively related to IncomePerCap given that the other variables (White, Poverty, Professional, Office, Walk and Self Employed) are held constant. These variables will have an effect on the relationship between Child Poverty and IncomePerCap.

One explanation could be that the other variables are pulling the expected value of Child Poverty down too far, so the regression explains more variation overall with a positive value for IncomePerCap to pull the expected value back up.

Another possibility could be because of confounding variables. Correlation does not mean causation, and the correlation value could have been influenced by confounding or extraneous factors. However, in this case the previous explanation seems to make more sense.

***Child Poverty Vs. Professional***

Child Poverty and Professional (% employed in management, business, science, and arts) have a negative correlation (-0.3877). In the regression equation Child Poverty has a negative slope (-0.43) as well. This shows that Child Poverty is low in areas with a higher population people that work Professional jobs.



***Child Poverty Vs. Office***

The correlation between Child Poverty and Office (% employed in sales and office jobs)

is -0.3443. In the regression equation Office has a negative coefficient as well. These both cases indicate that Child Poverty is low in areas with high population of people that work Office jobs.



***Child Poverty Vs. Walk***

This was one of the most unusual significant variables in the model. Logically it is hard to comprehend on how Walk (% walking to work) could be related to Child Poverty. Not only that, but it doesn’t behave the same way when compared between the correlation matrix and the regression equation. The correlation coefficient for Walk with Child Poverty is 0.2186, which is a weak positive relationship, however the regression coefficient for Walk is -0.75, which is pretty high when compared to other slopes.

As explained in Child Poverty Vs. IncomePerCap, similar factors might be the reason to why the regression coefficient and the correlation coefficient don’t share the same sign.



***Child Poverty Vs. Self Employed***

The correlation coefficient between Child Poverty and Self Employed was 0.1747, which albeit is not a strong correlation, it is positive nevertheless. In the regression equation Self Employed has a positive slope (0.5). It indicates that Child Poverty is higher in areas where more people are Self Employed.

