1. **Find outliers in X, outliers in Y and influential points. You do not need to remove them just list a few of the ones found and indicate how you determine the points were outliers and/or influential.**

Y = Number of Crimes

X1= AREA is the area of the SMSA in 1000's of square miles

X2 = CITYPOP is the city population in thousands

X3 = GRAD is the percent of high school graduates

X4 = INCOME is the total income in 1976 in millions

X5 = LABOR is the number in the labor force in thousands

Logarithmic function to the base 10 has been applied to all variables to minimize outliers

Outliers in X :

Hat matrix values above 2\*(k+1)/n are considered outliers.

Here k = 5 and n = 100

LogCRIME points having Hat matrix values above 0.12 are considered outliers in X.

|  |  |
| --- | --- |
| **LOGCrime** | **h\_LOGCrime** |
| 4.464742837 | 0.231617 |
| 5.884213303 | 0.188872 |
| 4.709376486 | 0.141275 |
| 5.074221291 | 0.135845 |
| 4.44678298 | 0.135652 |
| 4.992553827 | 0.134996 |
| 4.419529832 | 0.133925 |
| 3.967117212 | 0.132860 |
| 5.724113413 | 0.121562 |

Outliers in Y:

Any *studentized deleted residuals(R Student)* value larger than a t-table value with n-k-2 degrees of freedom.

T-Table value for n=100 and k = 5 is 1.9858

There is only one point in LogCRIME that can be considered as an outlier since its R-Student value is above the t-table value

|  |  |
| --- | --- |
| **LOGCrime** | **rstudent\_LOGCrime** |
| 4.281783072 | 2.220210 |

Influential observations:

Cook’s Distance values larger than the fiftieth (50) percentile of an F distribution k+1 and n-k-1 degrees of freedom are considered influential

With k = 5 and n = 100, F-Value = **0.8977754**

There are no points with corresponding Cook’s Distance value higher than the F-Table value.

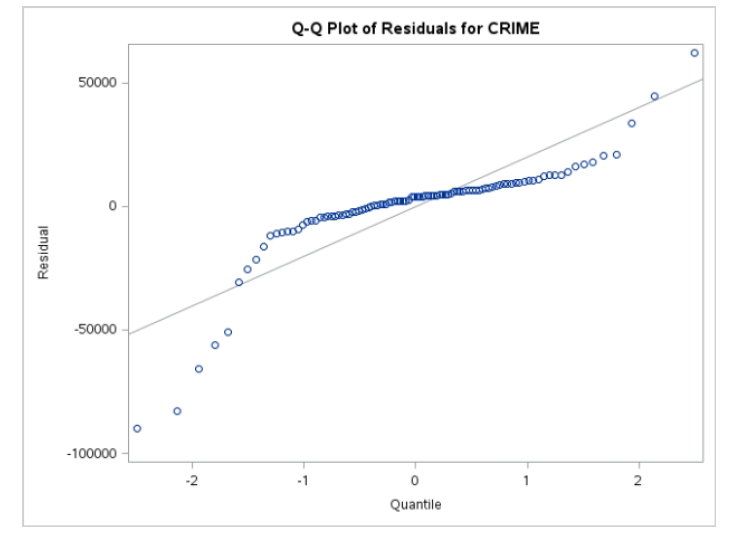
**2. Check for assumption violations**

**a**. if necessary use transformation to fix any assumptions that you note. Explain why you did or not find any assumption violations.

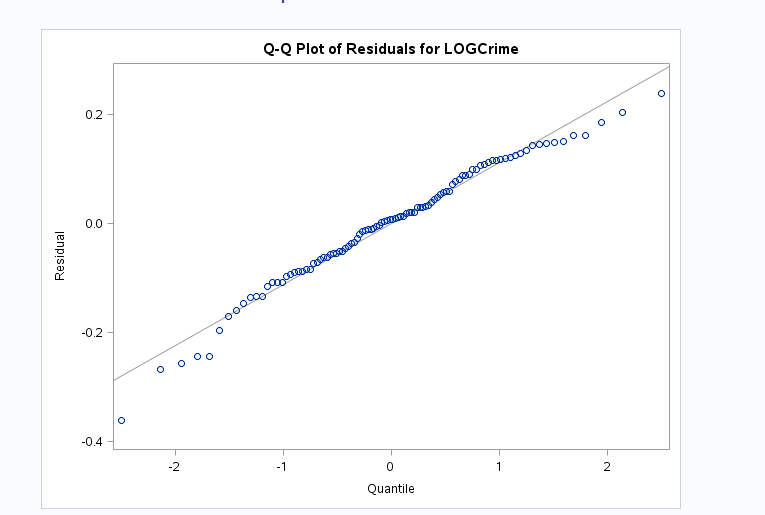
**b**. check assumptions again after your transformation.

**c**. comment on whether the assumption transformations helped or not. If they do not help, then revert back to the original variables.

Normality plot obtained using SAS (Before Transformation)

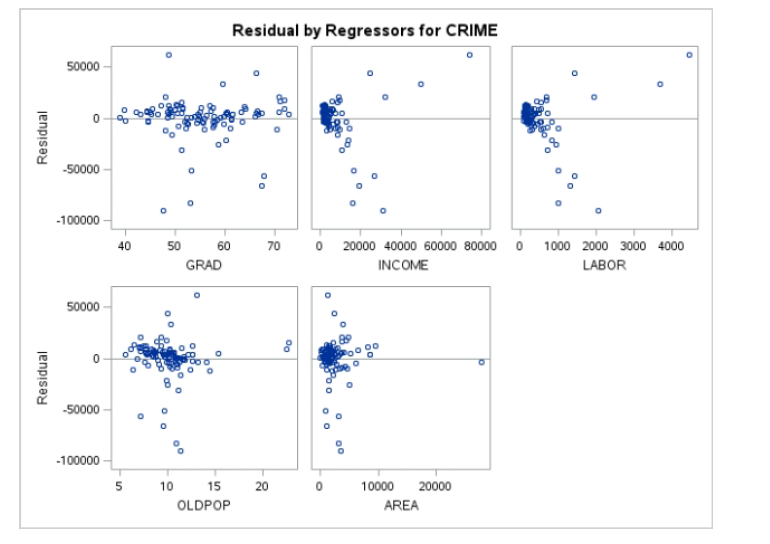


By applying logarithmic function to the Base 10 to all the dependent and independent variable, a significant change in normality graph can be observed where most points are on the line

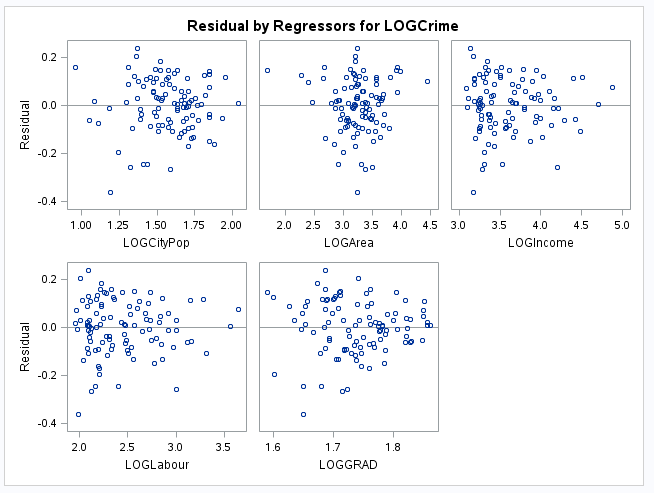


Residual plot having all assumptions met will have random scattered pattern without having any particular pattern or points being concentrated at one particular region.

The assumptions seem to be violated when the original values before transformation have been taken. Four out of the five residual plots below obtained using SAS (leaving GRAD) have displayed clear violations of assumptions.



However after transforming the values we now see that all assumptions being met for all the dependent variables with each residual plot being scattered randomly as expected.



Hence we are retaining the transformations to produce our models.

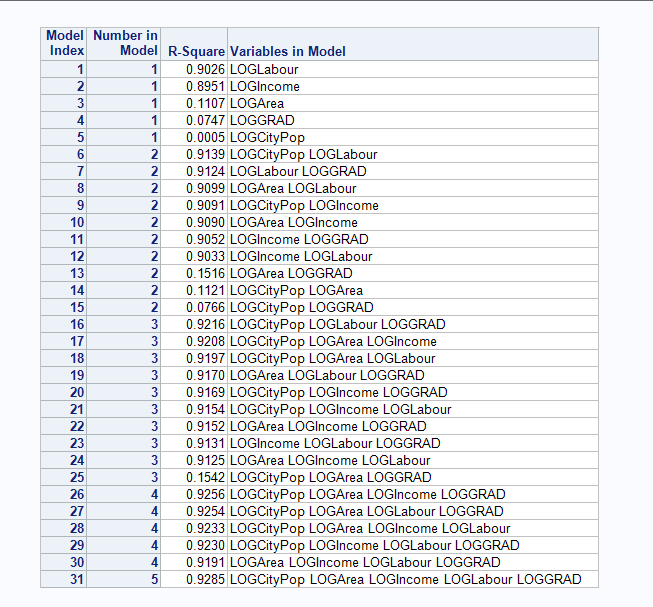
1. **Using all possible regression, reduce the list of possible models down to three, explaining your reasoning for choosing those three models.**

After running all possible regression we chose the 3rd ,4th and 5th Model because the R-square values look to be unstable in the 1st and 2nd model. If we observe the R-square values for 3rd ,4th and 5th models there are minor variations –

3rd Model |0.9125 to 0.9216 with the exception of 1542

4th Model |0.9191 to 0.9256

5th Model |0.9285



1. **Examine the three models in detail and choose one model. Explain why you choose that model and write down the least squares line for that model**

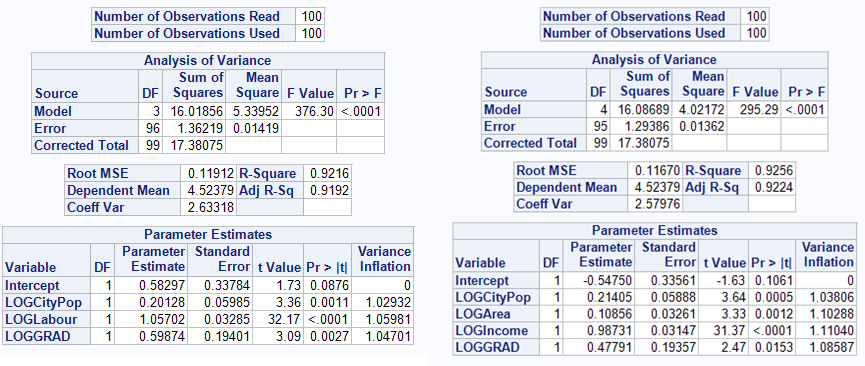
After comaring the three models on residual plots , R2, outliers in the residual plots , normal probability plots, signs of coefficients, normal probability plots , t-tests, Varince Inflation values we have chosen Model 4 mainly because of these factors:

* High Adjusted R square – 92.24%
* Significantly high F Value – 295.29
* Variance Inflation values are low for each variable
* Mean Square error is on the lower side 0.01362

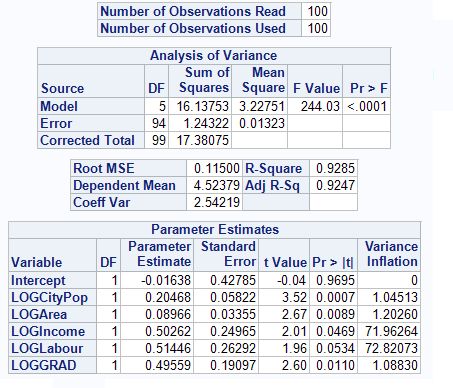
Least Square line

-0.5475 + 0.21405(**LOGCityPop) +** 0.10856(**LOGArea**) + 0.98731(**LOGIncome**) + 0.47791(**LOGGRAD**)

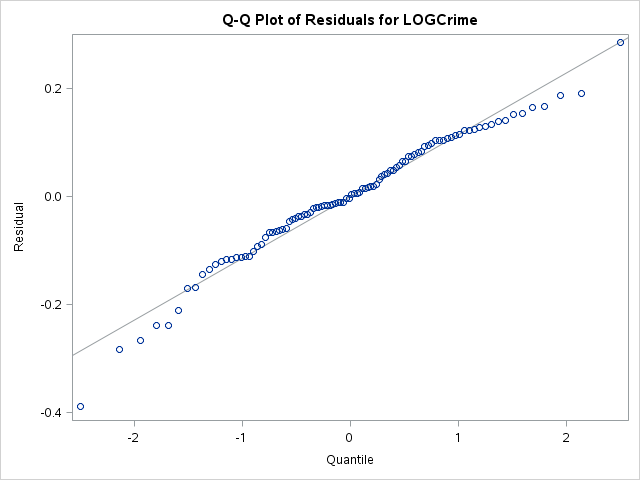
Model 3 and 4 (below)

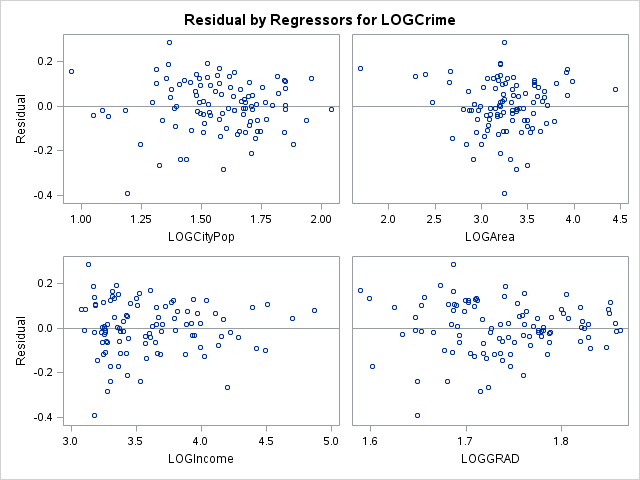


Model 5 (Below)



Other factors for the 4th Model





**Validation Set of Data:**

**5. Using the least squares equation found in 4 from your training set, predict the values of the dependent variable in your validation data set. Average the absolute differences between the actual value and the predicted value. Discuss the average prediction you would expect in the future using this model.**

Least Square Line:

-0.5475 + 0.21405(**LOGCityPop) +** 0.10856(**LOGArea**) + 0.98731(**LOGIncome**) + 0.47791(**LOGGRAD**)

Using this equation the least square line in the 100 samples given in our data set we found that the average percentage error is 21.29 % which is considerably low. However if we chose to use this model in the future we cannot certainly say that this model can predict the crimes in a region when another data set is used.

Here is the link to excel file –



**6. Using the validation data set, see if the model found in 4 is useful.**

|  |  |
| --- | --- |
| Y = | Number of Crimes |
| X1= | Log of City Population |
| X2= | Log Area |
| X3= | Log Income |
| X4= | Log Number of graduates |
| MSR= | 4.02172 |
| MSE= | 0.01362 |
| n= | 100 |
|  | 0.05 |
| k = | 4 |
| n-k-1= | 95 |

H₀: β₁ = β₂ = β3 = β4

H₁: not all β's= 0

Test Statistic:

F = MSR/MSE= 4.02172/0.01362=295.2805

Rejection Region:

Reject H₀ if F > F(k,n-k-1,α)=F(4, 95,0.05) = **2.46749362**

Decision:

Reject Ho

Conclusion:

Using the specified model, we can conclude that population, graduates, area and income are useful for estimating the average value of Number of Crimes

In addition to this

· The percentage error is low = 21.2984%

· R-Square is significantly high = 92.24%

· Variance inflation factors are close to 1