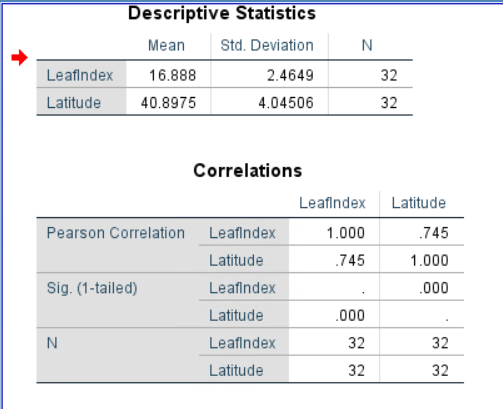
ASSIGNMENT – 1 – CSC 424

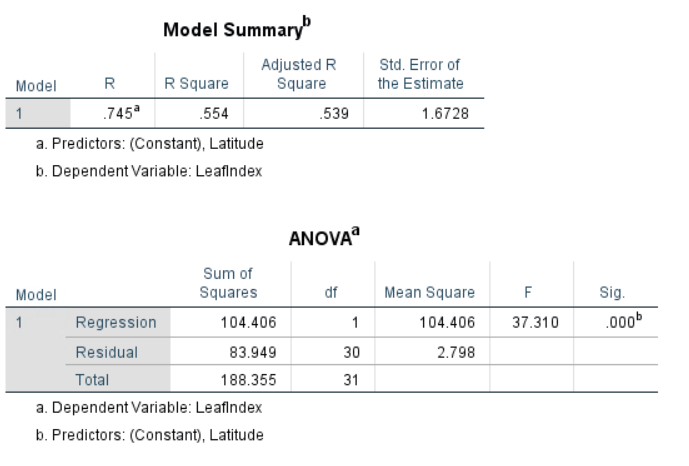
BY

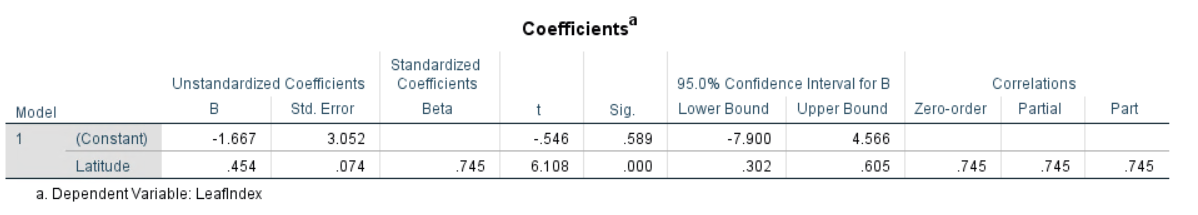
Sasidhar Mukthinuthalapati (ID: 1838169)

Problem – 1:

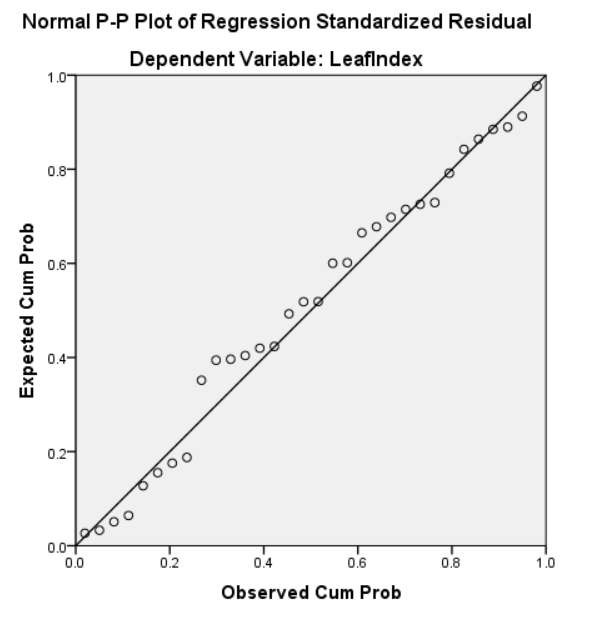
1. Regression of Leaf Index on Latitude

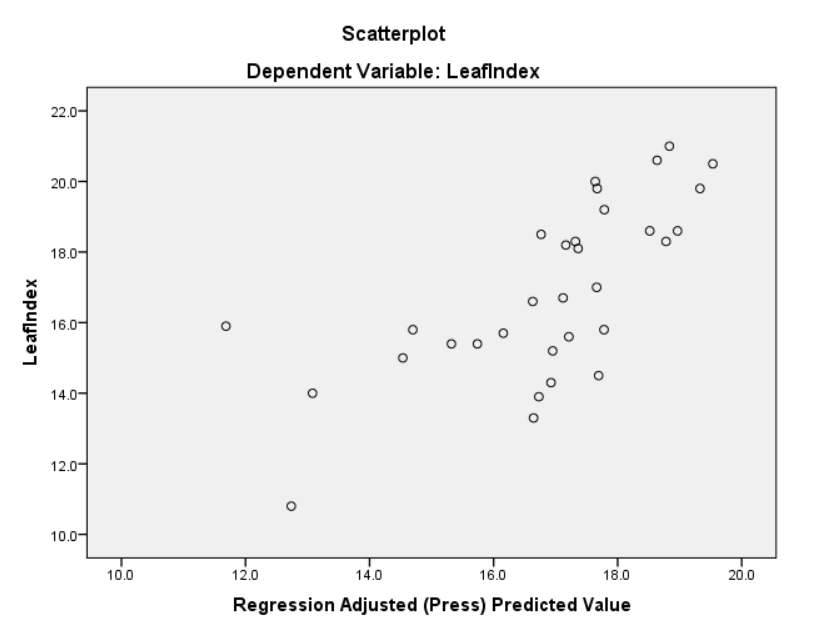




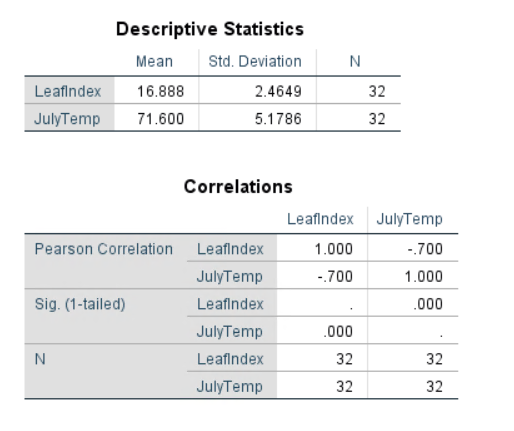


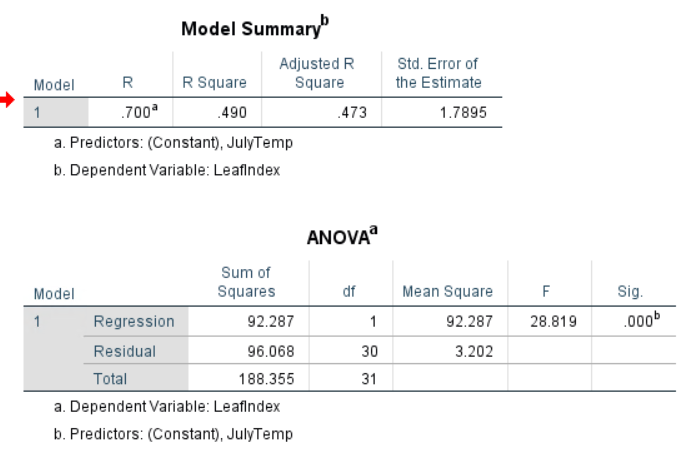
* As we can see from the above tables that Latitude alone can explain up to 53.9% of variance in LeafIndex.
* LeafIndex also has a significance less than 0.05 and hence we can say that it is a useful predictor for Leaf Index.
* The significance for the F-value related to this model is also less than 0.05 and hence we can say that this model is a decent model.
* We can see that Leaf Index and Latitude are positively correlated and that the magnitude of the Beta coefficient is also positive.

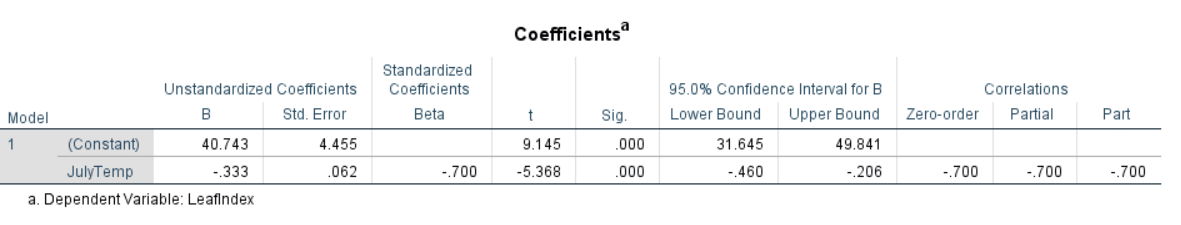




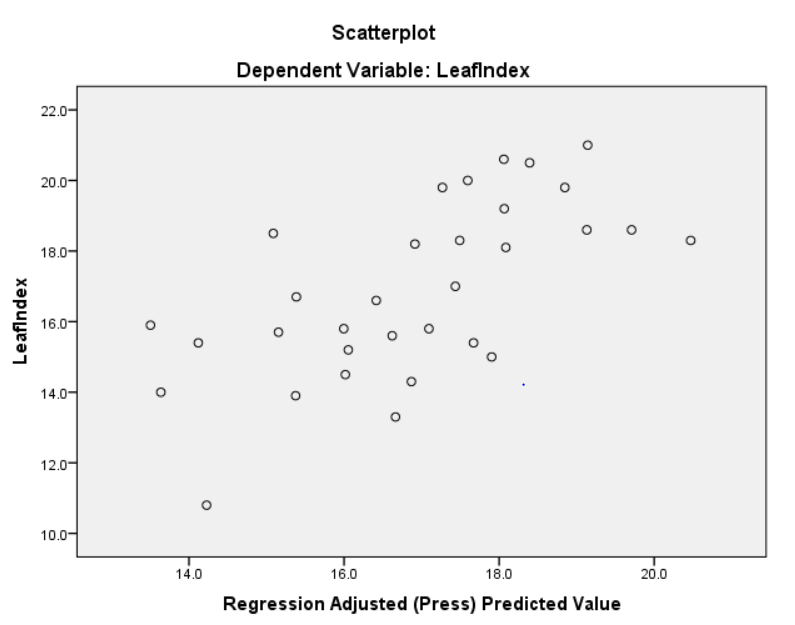
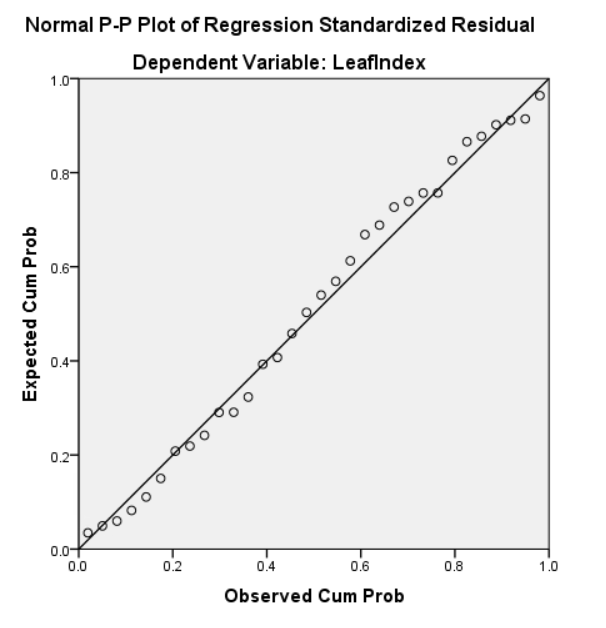
* As we can see from the normality plot that there aren’t any curves present in the top and bottom of the normality line and hence there aren’t many outliers in this data set but since there are points which deviate away from the normal line we can say that there are quite few influential points present in the dataset.





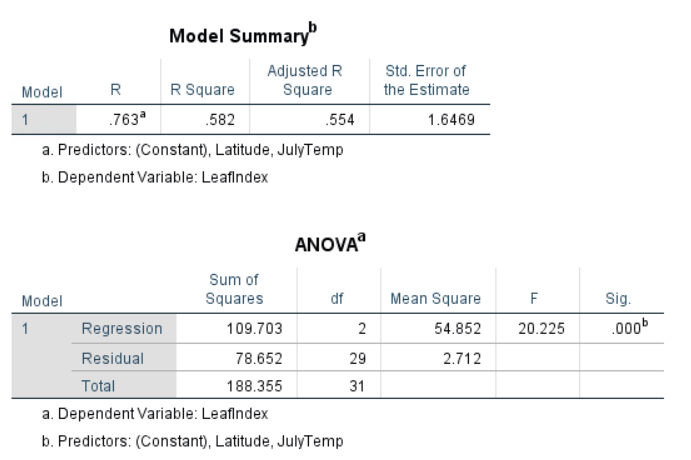
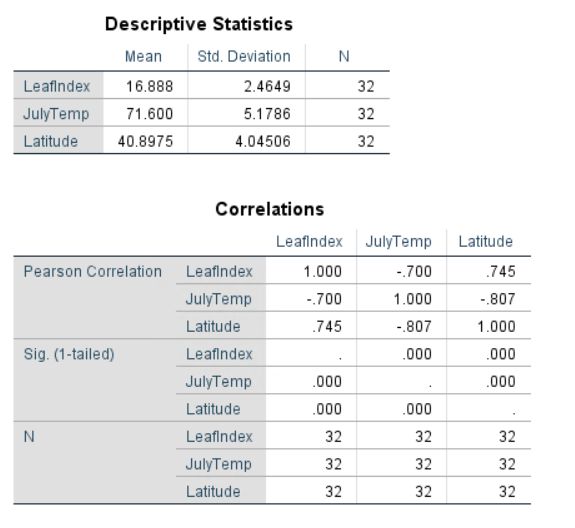


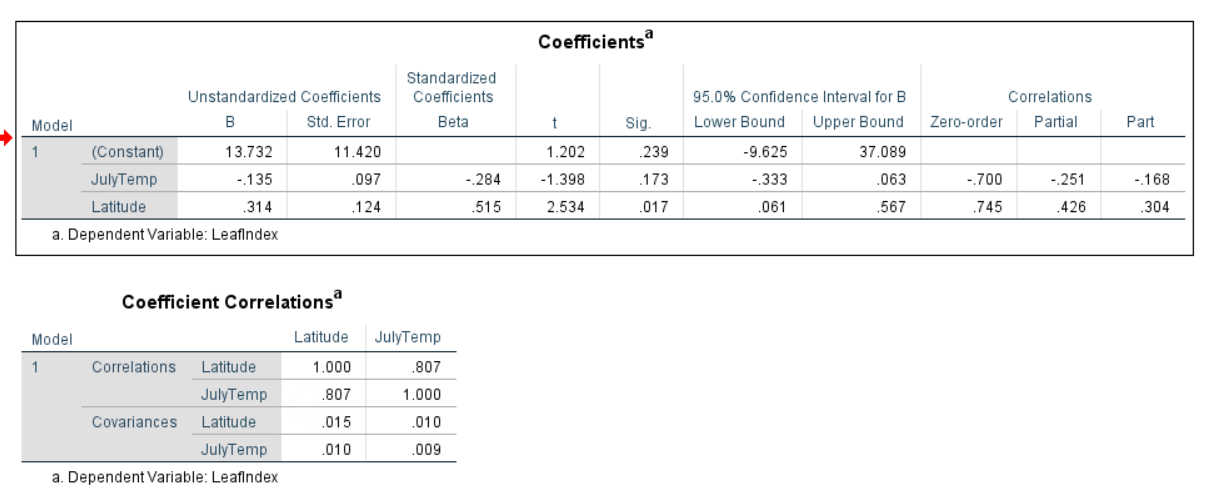
* As from the above tables we can see that LeafIndex is negatively correlated to July Temp and the magnitude of the Beta coefficient for July Temp is also negative.
* July Temp alone can explain up to 47.3% of variance in Leaf Index.
* Significance value associated with July Temp is also less than 0.05 suggesting that this variable is a useful predictor.
* The significance value associated with the F-value tells us that this is a decent model but whereas the Adjusted R-Squared value tells us that only 47.3% variance is being captured.
* On the whole, we can say that the July Temp is also a useful predictor.



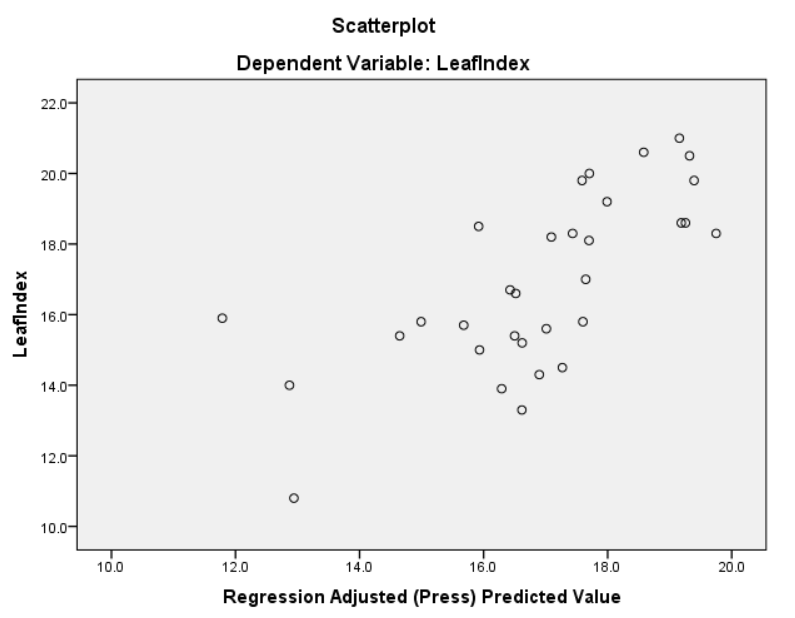
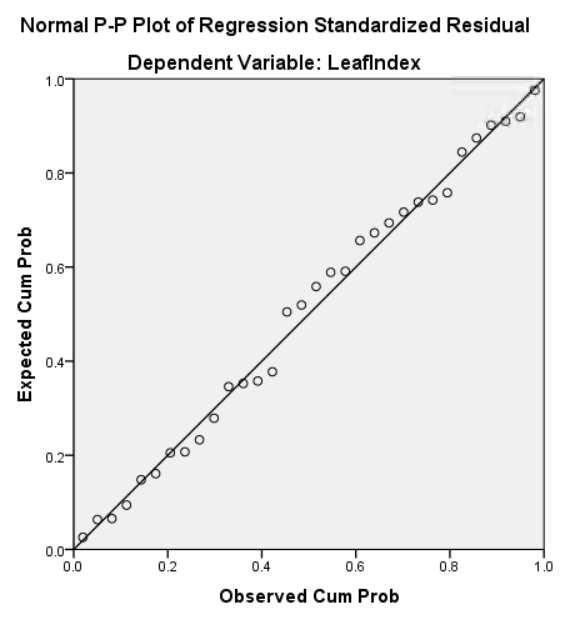
* From the normality plot we can see that there aren’t any curves at the start and the bottom of the normality plot which tells us that there aren’t any outliers present. But the deviations from the normality plot suggest the presence of influential points.
* From the actual vs predicted graph we can see that most the values are being predicted with slight error.

1. Complete model



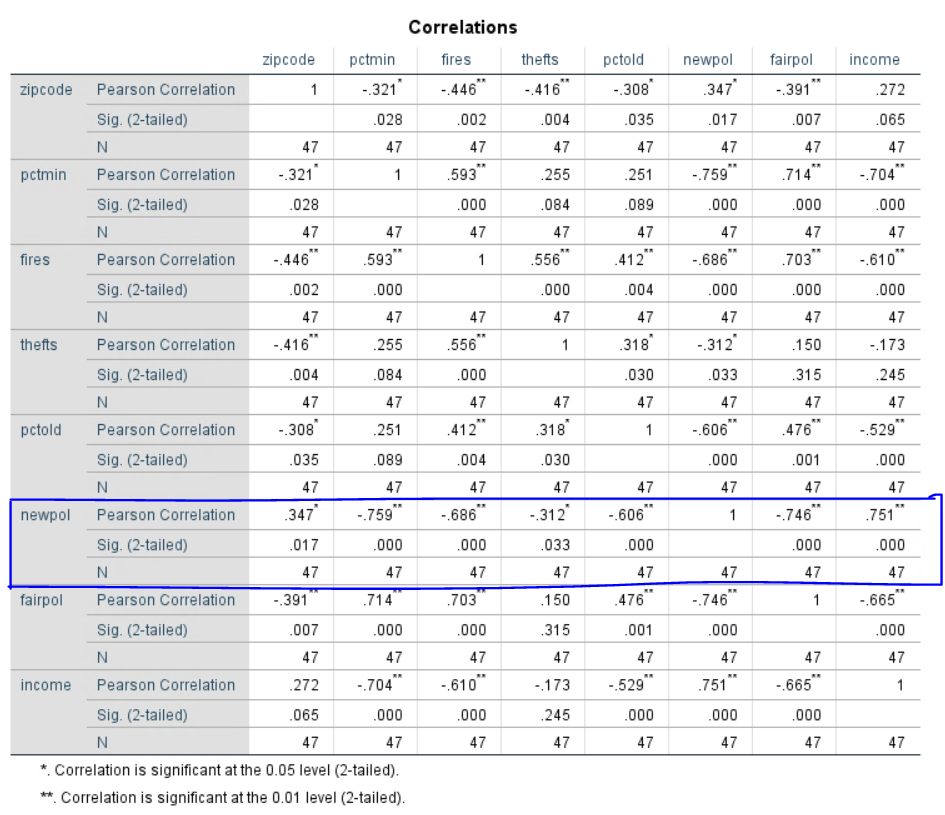


* As we can see that by having both July Temp and Latitude in the same model we can explain up to 55.4% variance in Leaf Index.
* In this both the predictors are significant.
* The slope coefficients in each case are different (example: Latitude in (a) had slope which was equal to 0.454 an in this case it is .314, the direction is the same but the magnitude has changed and this is due to the fact that we have added another variable which is also correlated to the dependent variable). Hence by adding variables which are correlated to the dependent variable the slope coefficients will change.
* The same can be said about the JulyTemp variable which in (b) had a slope coefficient of -0.333 and in this case, it is -0.135.
* But in this model, we can see that the July temp has significance value greater than 0.05.
* From the normality plot we can see that few points are deviating from the normal line which may suggest the presence of influential points.



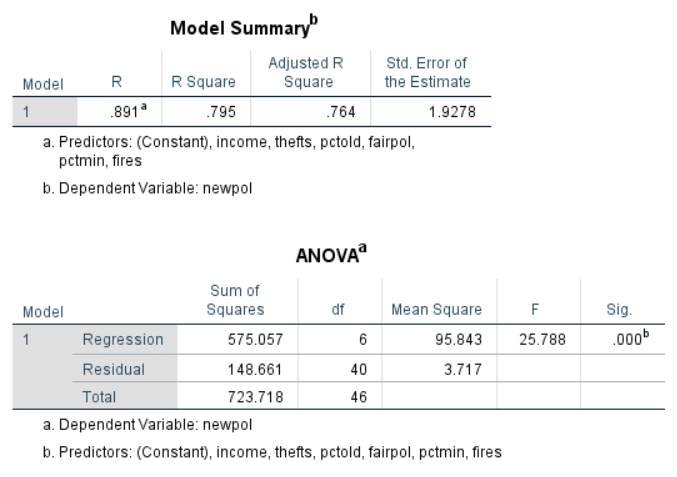
Problem – 2)

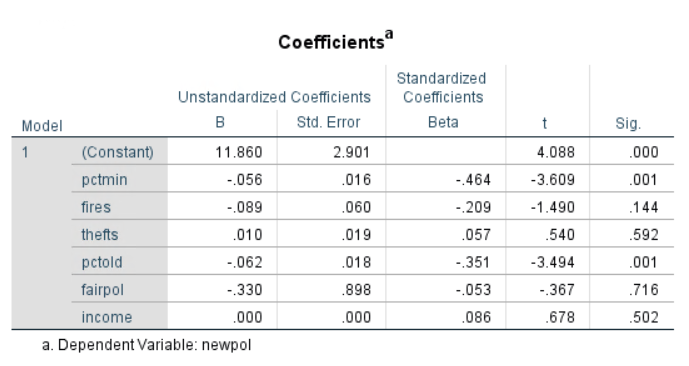
1. Based on the correlation values the signs are as follows:



|  |  |  |
| --- | --- | --- |
| VARIABLE NAME | SIGN FROM PEARSON’S CORRELATION | PREDICTED/EXPECTED SIGN IN THE FINAL REGRESSION EQUATION |
| ZIPCODE | 0.347 | POSITIVE |
| PCTMIN | -0.759 | NEGATIVE |
| FIRES | -0.686 | NEGATIVE |
| THEFTS | -0.312 | NEGATIVE |
| PCTOLD | -0.606 | NEGATIVE |
| FAIRPOL | -0.746 | NEGATIVE |
| INCOME | 0.751 | POSITIVE |

* + - 1. The complete regression model is as follows:



s

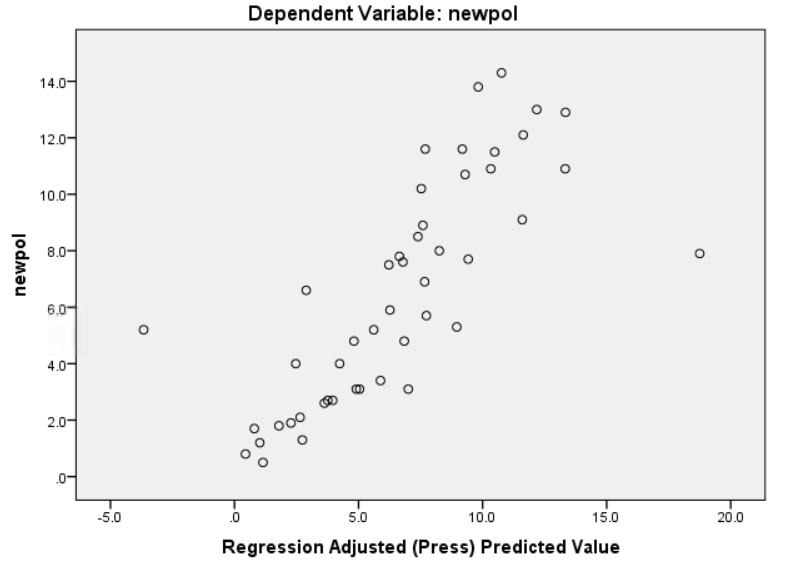
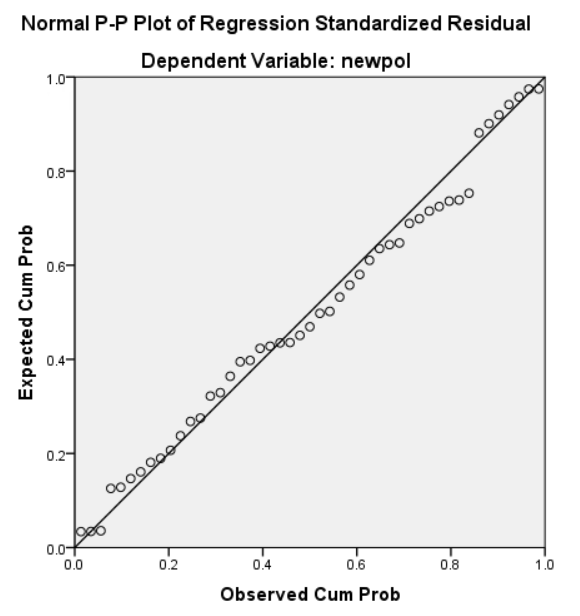
The P value of the overall F-test is significant, and hence our regression model predicts the response variable better than the mean of the response.

The complete model which includes all the variables can explain up to 76.4% variance of newpol variable.

* + 1. The predictors which are **significantly different from zero at the 0.05** level are **fires, pctmin, theft and fairpol. Keep in mind that income is not included in the list.**

The predictors which have significance value greater than 0.05 are fires, theft, fairpol and income.

* + 1. Only thefts variable has sign which is different than suggested by their simple correlations and this may be because I had removed zipcode and since while performing linear regression we are learning from the dataset and the changes are made to the beta coefficients of the variables and hence may have resulted in the change of sign of the theft variables.
    2. The plot of residuals vs predicted values are as follows:



The normality plot shows no curves at the top and bottom which means that there may not be any kind of outliers present in the dataset.

The actual vs predicted plot is shown right next to the normality plot, in this plot we can visualize that if the x=y line is present then there are many data points which are not lying on this line but are above or below the line which gives rise to the error in prediction.

There are two points one to the extreme left and the other to the extreme right which have a huge error. These points may be treated as outliers or influential points but for that we need to compute the dfbeta and tolerance values.

We can also see that the error is not completely normally distributed which can also be seen in the normality plot and the actual vs predicted value plot.

Here the predicted values are larger than the actual values over the range of 0-14. This means that we are over-estimating. Therefore, the model does not seem to provide an adequate fit and should be revised.

Problem – 3)

Multiple Regression analysis can be used to predict the temperature of a location given we have historic data from which we can learn from. The dependent variable in this case is Temperature and the independent variables will be Apparent temperature, Humidity, Wind Speed, Wind Bearing Direction, Air Pressure, Time and Date and Precipitation(like rain or snow etc).