CA1: Multiple Regression

Year and country not relevant for the regression model. So will be ignored.

Remaining columns:

-Life Expectancy(Dependent Variable)

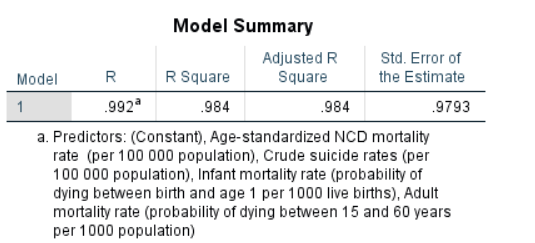
-Adult Mortality Rate

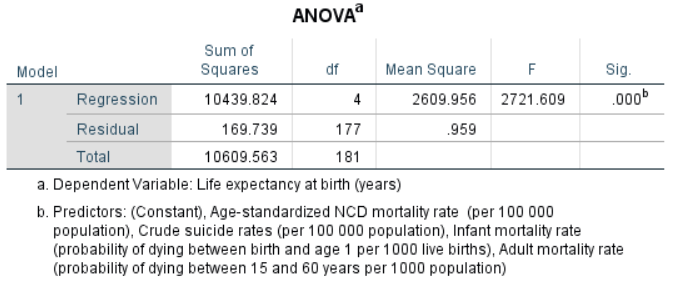
-Infant Mortality Rate

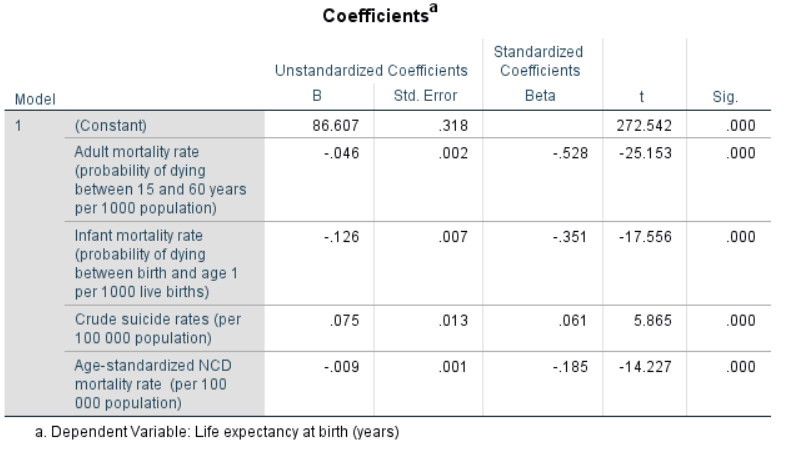
-Crude Suicide Rate

-NCD mortality rate

Initial stats of regression analysis:



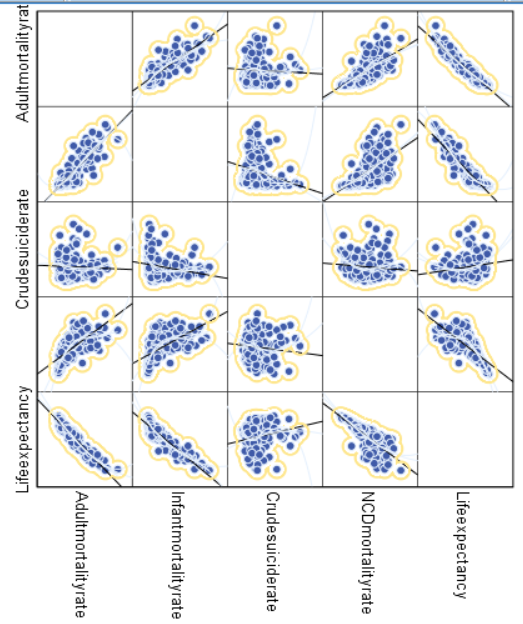




OLS and other Assumption, Diagnostic Testing:

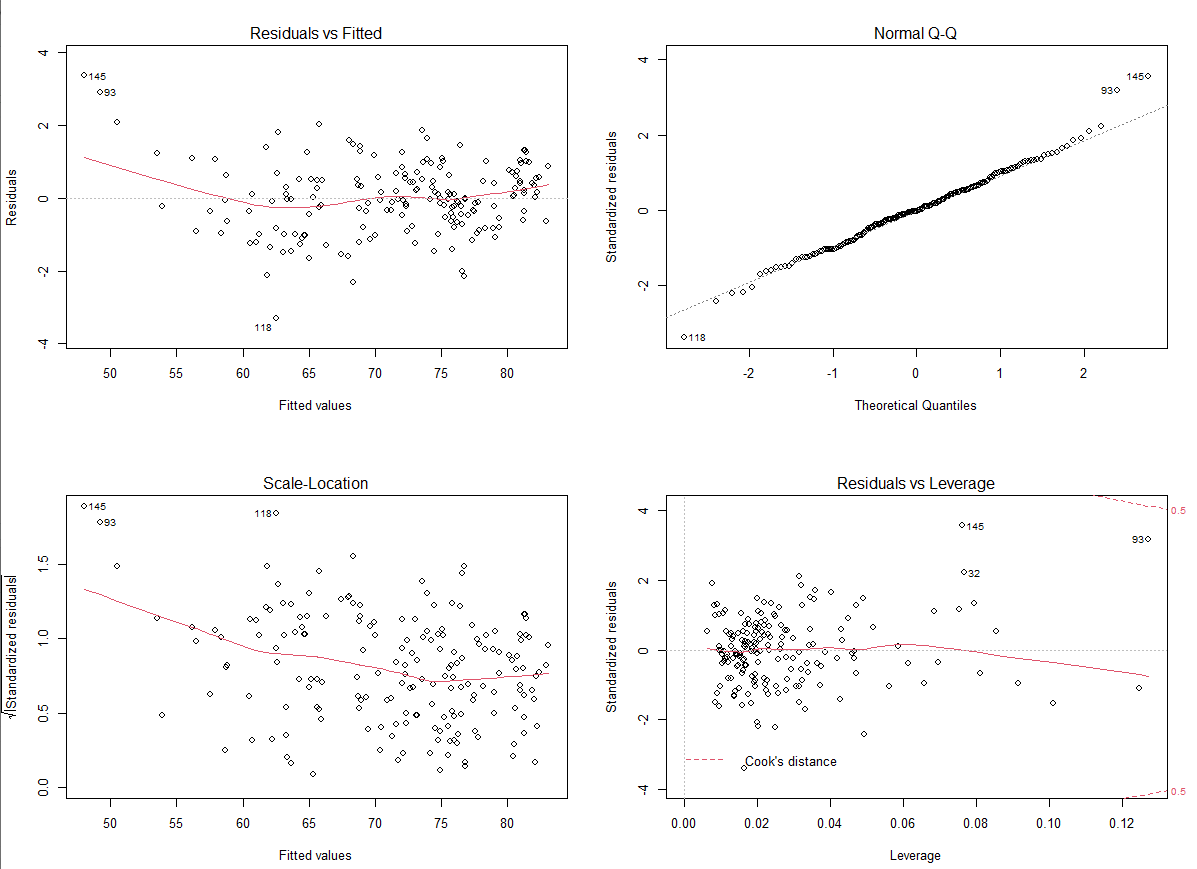
Gauss Markov Assupmtions:

i>Correct Functional Form: In this case Linearity



We can see that the DV has linear relationships with the IVs.

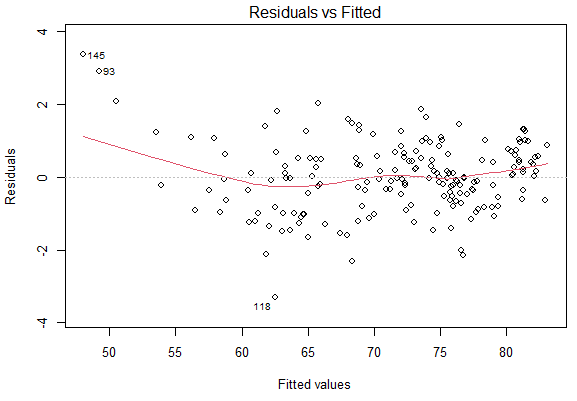
Same thing can be analyzed in R through Diagnostic plots.



The First fig: Residual vs fitted

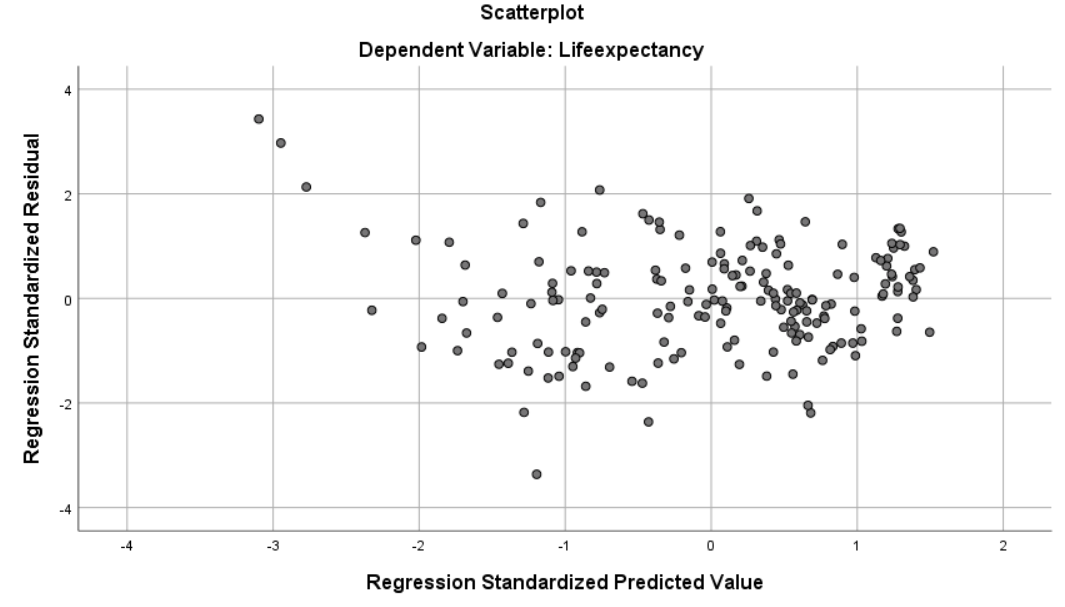
Check Linearity: Check the In the Residuals vs Fitted graph in R. You don’t want to see clear evidence of a curved relationship.

In my dataset, I do not see any Linear relation (curve) between residual and fitted. So we are good here.



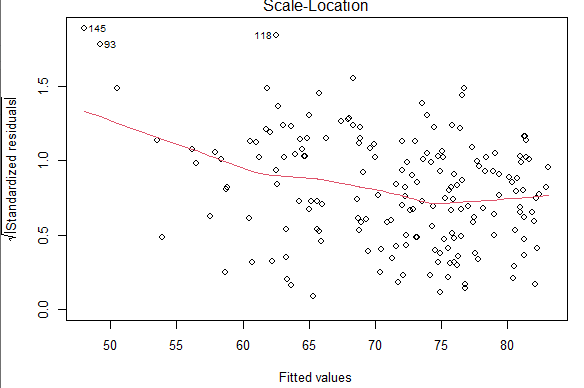
ii>Homoscadasticity: Errors should have constant Variance

ZRESID Vs ZPRED plot: It should be random without showing any pattern



In my dataset, it is random. So we are good here.

We can analyse the same in R:

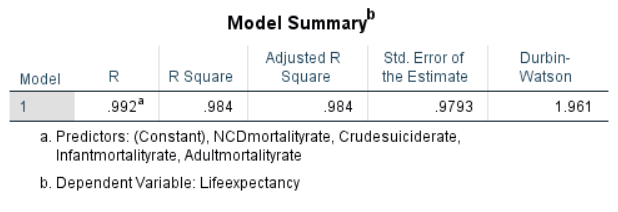


If you’ve met the constant variance assumption,the points in the Scale-Location graph of the R Diagnostic plots should be a random band around a horizontal line.

My dataset is satisfying this assumption. So we are good here.

iii> No Autocorrelation between errors

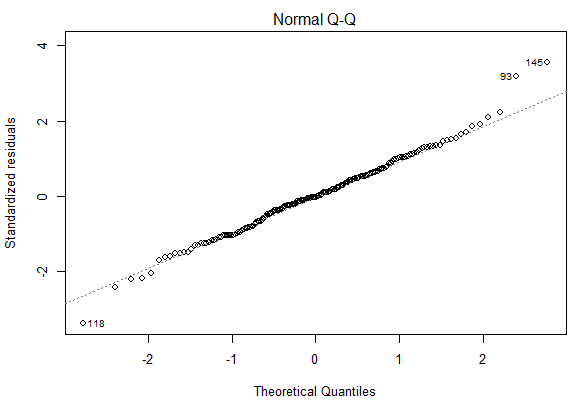
Durbin-Watson Statistic:



Here the value is showing 1.961 which is close to 2, i.e., no autocorrelation between errors.

iv> Errors should be normally distributed:

It can be seen in the R diagnostic plots (Normal Q-Q )



->If the dependent variable is normally distributed for a fixed set of predictor values, then the residual values should be normally distributed with a mean of 0.

->The Normal Q-Q plot (upper right) is a probability plot of the standardized residuals against the values that would be expected under normality. If you have met the normality assumption,

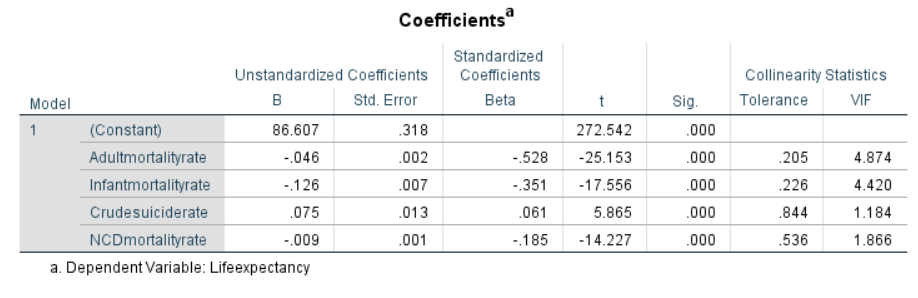
the points on this graph should fall on the straight 45-degree line.

->Alternatively, look at a histogram of the residuals to see if they appear to approximate a normal distribution.

In my dataset, it is satisfying the assumption. So we are good here.

v> Absence of Multicollinearity:

VIF and Collinearity Tolerance:

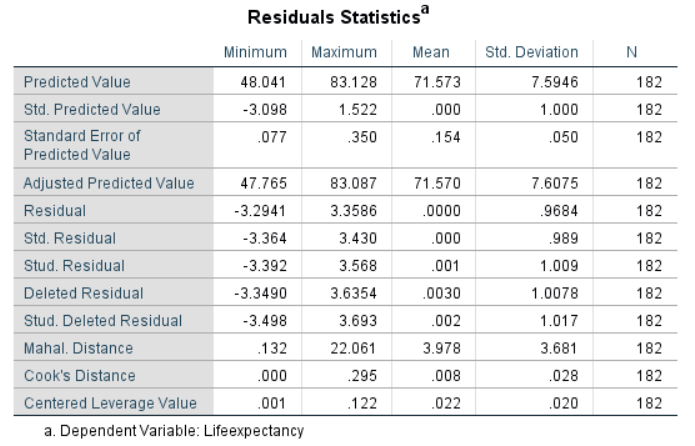


Here, the VIF values for all the features are less than 5 which is satisfying the assumption.

But the Collinearity tolerance for adult mortality, infant mortality and crude suicide rate is not close to 1. This is not satisfying the assumption. Doubt: need to ask Tony

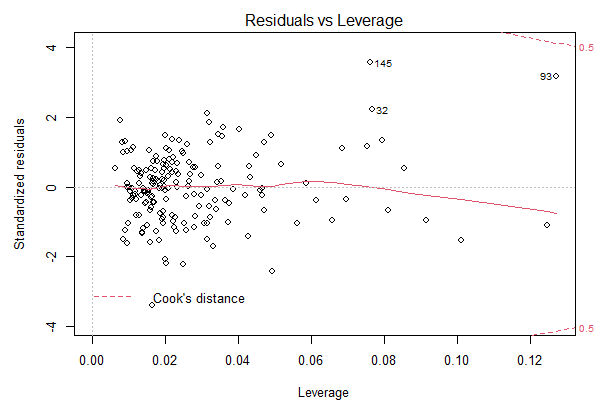
v> No influential datapoints(No outliers)

Cook’s distance and leverage values



Here we can see that maximum cook’s distance is 0.295 which is less than 1. So, we are good here.

This can also be analyzed in R diagnostic plot.



Here, we don’t see any effect of outliers. So, we are good here.