Diagnostic Analysis of Heteroscedasticity, Autocorrelation and Multicollinearity in a Regression Model

Problem Define:

Fit a model of multiple regression analysis (choose variables from the world bank data) and check the problem of "Heteroscedasticity" with various tests and interpret the results.

Methodology:

Heteroscedasticity refers to a condition in a statistical model in which the variability of the residuals or errors is not constant across all levels of the independent variables.

Let, the dependent variable, Y = population growth (annual) and

The independent variables are respectively,

X_1 = net migration

X_2 = number of infant deaths

X_3 = fertility rate

- Population growth refers to the rate at which the number of individuals in a population increase over a year.
  Annual population growth rate =

$$\frac{population\ at\ the\ end\ of\ the\ year - population\ at\ the\ beginning\ at\ the\ year}{population\ at\ the\ beginning\ at\ the\ year} \times 100\%$$

- Net migration refers to the difference between the number of people entering a geographic area and the number of people leaving the same area over a specific period.

- Fertility rate refers to the average number of children that would be born to a woman over her lifetime.

TFR $=\sum_1^n ASFR \times 5$; where ASFR = age specific fertility rate

So, the model is, $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \in_i$

After fit the model we run some test to find the problem of heteroscedasticity. Such as,

- BPG test
- White test
- Park test

Then test the hypothesis as,

H_0: Absence of heteroscedasticity.

H_1: Presence of heteroscedasticity.

Result:

Studentized Breusch-Pagan test		
data: model		
BP = 5.6289	df = 3	p-value = 0.1311

Comment:

From the result table we can see that the result of p-value is 0.1311 which is greater than 0.05. So, we can conclude that there is no significant evidence of heteroscedasticity in our regression model.

.....

Problem Define:

Fit a model of multiple regression analysis (choose variables from the world bank data) and check the problem of "Autocorrelation" with various tests and interpret the results.

Methodology:

Autocorrelation refers to the correlation between values of the same variable at different points in time or space.

Let, the dependent variable, Y = population growth (annual) and

The independent variables are respectively,

X_1 = net migration

X_2 = number of infant deaths

X_3 = fertility rate

- Population growth refers to the rate at which the number of individuals in a population increase over a year.
  Annual population growth rate =

$$\frac{population\ at\ the\ end\ of\ the\ year - population\ at\ the\ beginning\ at\ the\ year}{population\ at\ the\ beginning\ at\ the\ year} \times 100\%$$

- Net migration refers to the difference between the number of people entering a geographic area and the number of people leaving the same area over a specific period.

- Fertility rate refers to the average number of children that would be born to a woman over her lifetime.

TFR = $\sum_1^n ASFR \times 5$; where ASFR = age specific fertility rate

So, the model is, $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \in_i$

After fit the model we run some test to find the problem of autocorrelation. Such as,

- Durbin Watson d-test
- Run test

Then test the hypothesis as,

H_0: Absence of autocorrelation.

H_1: Presence of autocorrelation.

Result:

Runs Test - Two sided
data: residual (e)
Standardized Runs Statistic = -5.2083
p-value = 1.906e-07

Comment:

From the calculation we find that the value of "Run statistic" is -502083 and p-value is 1.906e-07. Here the p-value is significantly less than 0.05. So, we can accept the null hypothesis and strongly conclude that there is significant autocorrelation in the data.

.....

Problem Define:

Fit a model of multiple regression analysis (choose variables from the world bank data) and check the problem of "Multicollinearity" with various tests and interpret the results.

Methodology:

Multicollinearity occurs when two or more independent variables in regression model are highly correlated with each other.

Let, the dependent variable, Y = population growth (annual) and

The independent variables are respectively,

X_1 = net migration

X_2 = number of infant deaths

X_3 = fertility rate

- Population growth refers to the rate at which the number of individuals in a population increase over a year.
  Annual population growth rate =

$$\frac{population\ at\ the\ end\ of\ the\ year - population\ at\ the\ beginning\ at\ the\ year}{population\ at\ the\ beginning\ at\ the\ year} \times 100\%$$

- Net migration refers to the difference between the number of people entering a geographic area and the number of people leaving the same area over a specific period.

- Fertility rate refers to the average number of children that would be born to a woman over her lifetime.

$$TFR = \sum_1^n ASFR \times 5$$; where ASFR = age specific fertility rate

So, the model is, $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \in_i$

After fit the model we run some test to find the problem of multicollinearity. Such as,

- Examination of determinant of X^TX
- Examination of correlation matrix
- Variance inflation factors (VIF)
- Tolerance
- Eigen value decomposition

Result:

VIF results

x1	x2	x3
1.23333	11.00111	11.71567

Comment:

From the calculation we can see that VIF(X_1) =1.23333, VIF(X_2)=11.00111, and VIF(X_3)=11.71567.For X_1, this value is close to 1,indicating that X_1 has little to no multicollinearity with other independent variable. For both X_2 and X_3 the value is above 10 which indicate very high degree of multicollinearity.

.....