**[Test for the Assumptions of Linear Regression Using R](https://www.meetup.com/fedpofa-r-users-group/events/299459705/)**

BEING A SHORT TRAINING PRESENTED AT OFFA R USERS GROUP MEETING ON 26TH MARCH, 2024 AT STATISTICAL LABORATORY, STATISTICS DEPARTMENT, THE FEDERAL POLYTECHNIC OFFA, NIGERIA

BY

**UDOKANG, ANIETIE EDEM (OGANIZER, ORUG)**

CHIF LECTURER, STATISTICS DEPARTMENT, THE FEDERAL POLYTECHNIC OFFA, NIGERIA

**Simple Linear Regression**

****

Where

 - the dependent variable

- independent variable

- constant or the intercept on - axis

- coefficient or slope

- random error

**Multiple Linear Regression**

****

Where

 - the dependent variable

****- independent variables

****- constant or the intercept on - axis

****- coefficients

- random error

**Assumptions**

1. There must be a linear relationship between the variables
2. The error term must be homoscedastic (Equal Variance)
3. There must not be autocorrelation in the error term
4. There is no existence of multicollinearity in the data
5. The residuals must be normally distributed

In the multiple Linear Regression of **.**

Let the Y be Gross Domestic Product represented by GDP and X1 be Import represented by IMP and X2 be Export represented by EXP.

Therefore,

****

This the model that will be use to illustrate the different tests that will be carried out.

**1. Test for Linearity**

**Scatter Diagram**

The plot of GDP against IMP, EXP and GVTREV will be done using Scatter Diagram to determine linearity.

**R Code**

# Read Data EXIMGDP in Excel from Document

EXIMGDP=read.csv('EXIMGDP.csv',head=T)

plot(EXIMGDP$EXP, EXIMGDP$GDP, xlab="EXP,IMP and GVTREV", ylab="GDP",main= "Plot of GDP Against EXP,IMP and GVTREV",col="red",pch=16,xlim=c(20,50),ylim=c(60,110))

points(EXIMGDP$IMP, EXIMGDP$GDP,col="blue",pch=16)

points(EXIMGDP$GVTREV, EXIMGDP$GDP,col="green",pch=16)

legend(20,110,legend=c('EXP', 'IMP', 'GVTREV'), pch=c(16,16,16),col=c('red', 'blue', 'green'))

# There is linear relationship as indicated by the scatter diagram between the response variable and the explanatory variables.

**Action:** No action required. If any of explanatory variables did not show linearity, then it should be transformed using any of the appropriate method such as logarithm and differencing.

**Estimation of the Tentative Model for other Tests**

head(EXIMGDP,20)

lm1=lm(GDP~IMP+EXP+GVTREV,data= EXIMGDP)

summary(lm1)

#The p-value of 9.278e-14 indicates the model is suitable to the data (the model has passed the goodness of fit test).

#This notwithstanding, the test for some important assumptions of regression model can further improve the model.



**Test for the Assumptions**

**2. Homoscedasticity (Constant Variance)**

**Goldfield-Quandit Test**

H0: There is homoscedasticity (Constant variance)

Vs

H1: There is heteroscedasticity (Variance are not constant)

**R Code**

library(lmtest)

gqtest(lm1)

Since p-value = 0.9895 is not less than 0.05 level of significance, there is homoscedasticity.

**Action:** If there is heteroscedasticity, the data needs to be transformed using an appropriate transformation technique such as logarithm and reciprocal.

**3. Autocorrelation (The Error Terms are Independent)**

**Durbin-Watson Test**

H0: There is no autocorrelation

Vs

H1: There is autocorrelation

**R Code**

library(lmtest)

dwtest(lm1)

The DW = 1.316 and p-value = 0.01142<0.05, meaning that there is autocorrelation.

**Action:** The original data should be transformed using autocorrelation of the residuals (random term-U) between Ut and Ut-1.

**4. No Multicollinearity (No Existence of High Correlation between the Explanatory Variables)**

**Variance Inflation Factor (VIF)**

H0: There is no multicollinearity

Vs

H1: There is multicollinearity

Decision Rule:

VIF = 1, there is no multicollinearity.

1<VIF<=5, there is moderate multicollinearity

VIF>5, there is high correlation between a given explanatory variable and other explanatory variables, hence existence of multicollinearity

**R Code**

library(car)

vif(lm1)

**R Output**

IMP EXP GVTREV

5.692972 7.785403 4.282330

There is existence of multicollinearity in IMP and EXP at severe level. But GVTREV has a moderate existence of autocorrelation which may require any action.

**Action**-Remove EXP with the highest VIF or find way appropriate way of combining the two of EXP and IMP.

**5. Residuals must be Normally Distributed**

**Q-Q Plot**

**R Code**

library(forecast)

resid<- resid(lm1)

qqnorm(resid)

#This is near normality even though the points are not in a straight line but are close it, except one point which normality can be considered by these observation.

#Let this be sunjected to a test of hypothesis using Shapiro-Wilk Test.

**Shapiro-Wilk Test.**

H0: The residuals are normally distributed

Vs

H1: The residuals are not normally distributed

**R Code**

shapiro.test(residuals(lm1))

# The Shapiro test statistic W = 0.95848, p-value = 0.5139>0.05, hence the residuals have a normal distribution.

**Action:** No action. If the residuals are not normally distributed the original can be transformed or another model of non-linear form can be used.

This is end of today’s short training hosted by the Offa-R-Users-Group (ORUG) a place for learning and using R. I wish to you to be part of the training session either online or physical. If you are a guest, find time to register as a member to actualize your goal in using R.

The ORUG (<https://www.meetup.com/fedpofa-r-users-group/> ) is sponsored by R Consortium and AniKem\_Consult. For any Enquiry Contact the Organizer (WhatsApp: +2349030912602, email: anietieeu@yahoo.com)

**BY NEXT QUARTER WE SHALL CONSIDER ACTIONS TO BE TAKEN WHEN THE ASSUMPTIONS ARE VIOLATED.**

THANKS FOR BEING PART OF THIS SHORT TRAINING.