Test for the Assumptions of Linear Regression Using R

Test for the Assumptions of Linear Regression Using R

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 variable - constant or the intercept on - axis - coefficients - random error

#Assumptions *i. There must be a linear relationship between the variables* ii. The error term must be homoscedastic (Equal Variance) *iii. There must not be autocorrelation in the error term* iv. There is no existence of multicollinearity in the data \*v. The residuals must be normally distributed

In the multiple Linear Regression of

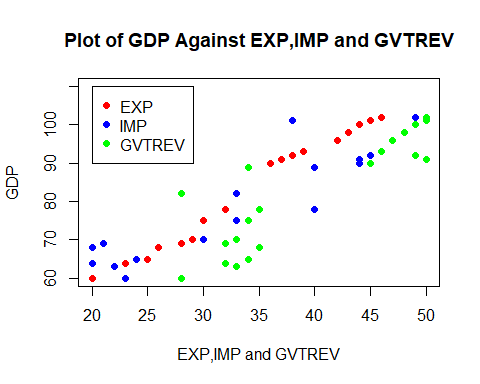
Let the be Gross Domestic Product represented by GDP and be Import represented by IMP and be Export represented by EXP.

Therefore,

This model will be used 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.

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)

## YEAR Time IMP EXP GDP GVTREV  
## 1 2000 1 23 20 60 28  
## 2 2001 2 22 22 63 33  
## 3 2002 3 20 23 64 32  
## 4 2003 4 24 25 65 34  
## 5 2004 5 20 26 68 35  
## 6 2005 6 21 28 69 32  
## 7 2006 7 30 29 70 33  
## 8 2007 8 33 30 75 34  
## 9 2008 9 40 32 78 35  
## 10 2009 10 33 33 82 28  
## 11 2010 11 40 34 89 34  
## 12 2011 12 44 36 90 45  
## 13 2012 13 44 37 91 50  
## 14 2013 14 45 38 92 49  
## 15 2014 15 46 39 93 46  
## 16 2015 16 47 42 96 47  
## 17 2016 17 48 43 98 48  
## 18 2017 18 49 44 100 49  
## 19 2018 19 38 45 101 50  
## 20 2019 20 49 46 102 50

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

##   
## Call:  
## lm(formula = GDP ~ IMP + EXP + GVTREV, data = EXIMGDP)  
##   
## Coefficients:  
## (Intercept) IMP EXP GVTREV   
## 24.13308 0.27234 1.39374 0.04009

summary(lm1)

##   
## Call:  
## lm(formula = GDP ~ IMP + EXP + GVTREV, data = EXIMGDP)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.0446 -1.3102 0.3726 1.1387 5.2232   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 24.13308 2.62016 9.211 8.52e-08 \*\*\*  
## IMP 0.27234 0.11294 2.411 0.0283 \*   
## EXP 1.39374 0.17717 7.867 6.89e-07 \*\*\*  
## GVTREV 0.04009 0.12855 0.312 0.7592   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.244 on 16 degrees of freedom  
## Multiple R-squared: 0.9798, Adjusted R-squared: 0.976   
## F-statistic: 258.3 on 3 and 16 DF, p-value: 9.278e-14

#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.

#2. Homoscedasticity (Constant Variance) ##Goldfield-Quandit Test : There is homoscedasticity (Constant variance) Vs : There is heteroscedasticity (Variance are not constant)

library(lmtest)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

gqtest(lm1)

##   
## Goldfeld-Quandt test  
##   
## data: lm1  
## GQ = 0.12049, df1 = 6, df2 = 6, p-value = 0.9895  
## alternative hypothesis: variance increases from segment 1 to 2

#Since p-value = 0.9895 is not less than 0.05 level of significance, there is homoscedasticity. *Action:*Nil. 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 : There is no autocorrelation Vs : There is autocorrelation

library(lmtest)  
dwtest(lm1)

##   
## Durbin-Watson test  
##   
## data: lm1  
## DW = 1.316, p-value = 0.01142  
## alternative hypothesis: true autocorrelation is greater than 0

#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 and .

#4. No Multicollinearity (No Existence of High Correlation between the Explanatory Variables) ##Variance Inflation Factor (VIF) : There is no multicollinearity Vs : 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

library(car)

## Loading required package: carData

vif(lm1)

## IMP EXP GVTREV   
## 5.692972 7.785403 4.282330

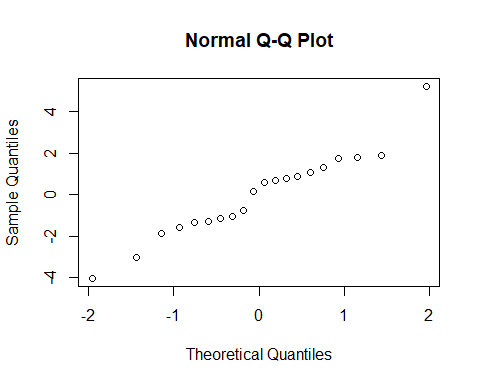
#There is existence of multicollinearity in IMP and EXP at severe level. But GVTREV has a moderate existence of autocorrelation which may require any action. *Action:* Remove EXP with the highest VIF or find an appropriate way of combining the two of EXP and IMP.

#5. Residuals must be Normally Distributed ##Q-Q Plot

library(forecast)

## Registered S3 method overwritten by 'quantmod':  
## method from  
## as.zoo.data.frame zoo

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.

: The residuals are normally distributed Vs : The residuals are not normally distributed

shapiro.test(residuals(lm1))

##   
## Shapiro-Wilk normality test  
##   
## data: residuals(lm1)  
## W = 0.95848, p-value = 0.5139

# 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](mailto: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.