International Journal of Statistics and Applied Mathematics 2021; 6(6): 24-30

ISSN: 2456-1452
Maths 2021; 6(6): 24-30

www.mathsjournal.com
Received: 03-08-2021
Accepted: 11-12-2021

MV Narayana Murthy
Research Scholar, Department of Statistics, S.V University, Tirupati, Andhra Pradesh, India

Akasam Srinivasulu
Research Scholar, Department of Statistics, S.V University, Tirupati, Andhra Pradesh, India

M Bhupati Naidu
DDE, Professor in Department of Statistics, S.V University, Tirupati, Andhra Pradesh, India

Corresponding Author:
MV Narayana Murthy
Research Scholar, Department of Statistics, S.V University, Tirupati, Andhra Pradesh, India

Testing the assumptions of simple linear regression in R

MV Narayana Murthy, Akasam Srinivasulu and M Bhupati Naidu

Abstract
The simple linear regression technique is a statistical technique which allows only two variables one is independent variable and another is dependent variable among the relationship is linear. Linear regression is a statistical procedure for calculating the value of a dependent variable from an independent variable. Linear regression measures the association between two variables. It is a modeling technique where a dependent variable is predicted based on one or more independent variables. Linear regression analysis is the most widely used of all statistical techniques. The present study discuss the implementation of linear regression using the R language for imports and exports of the country India from 1991 to 2017. This data can be downloaded from the world bank website by using the library WDI.

Keywords: Simple linear, statistical technique, library WDI

Introduction
Linear regression will show the relationship between two variables by applying a linear equation to observed data. One variable is supposed to be an independent variable, and the other is to be a dependent variable. It is not necessary that here one variable is dependent on others, or one causes the other, but there is some critical relationship between the two variables. In such cases, we use a scatter plot to imply the strength of the relationship between the variables. If there is no relation or linking between the variables, the scatter plot does not indicate any increasing or decreasing pattern. For such cases, the linear regression design is not beneficial to the given data. The measure of the extent of the relationship between two variables is shown by the correlation coefficient. The range of this coefficient lies between -1 to +1. This coefficient shows the strength of the association of the observed data for two variables.
A linear regression line equation is written in the form of:

$$Y = a + bX$$

where X is the independent variable and plotted along the x-axis
Y is the dependent variable and plotted along the y-axis
The slope of the line is b, and a is the intercept.
Now, here to find the value of the slope of the line, b, plotted in scatter plot and the intercept, a.

$$a = \frac{[(\sum y)(\sum x^2) - (\sum x)(\sum xy)]}{[n(\sum x^2) - (\sum x)^2]}$$

$$b = \frac{[n(\sum xy) - (\sum x)(\sum y)]}{[n(\sum x^2) - (\sum x^2)]}$$

The expansion to multiple and vector-valued predictor variables is known as multiple linear regression, also known as multivariable linear regression. Almost all real-world regression patterns include multiple predictors, and basic explanations of linear regression are often explained in terms of the multiple regression form.

The most popular method to fit a regression line in the XY plot is the method of least-squares. This process determines the best-fitting line for the data by reducing the sum of the squares of the vertical deviations from each data point to the line. If a point rests on the fitted line accurately, then its perpendicular deviation is 0. Because the variations are first squared, then added, their positive and negative values will not be cancelled.

In regression line where the regression parameters a and b are defined, the properties are given as:

- The line reduces the sum of squared differences between observed values and predicted values.
- The regression line passes through the mean of X and Y variable values
- The regression constant (a) is equal to y-intercept the linear regression
- The regression coefficient (b) is the slope of the regression line which is equal to the average change in the dependent variable (Y) for a unit change in the independent variable (X).

In the linear regression line, the equation is given by;

$Y = a+bX$

Where
a is a constant
b is the regression coefficient
The value of the regression coefficient.

$b = \Sigma [(xi – x)(yi – y)] / \Sigma [(xi – x)2]$

Where xi and yi are the observed data sets.
And x and y are the mean value.
The regression model has been developed as a typical statistical model based on the idea by Francis Galton in 1886 . To establish the simplest typical regression model, we set following four assumptions for the regression model .

- **Linear:** The means of X subgroups are on the straight line representing the linear relationship between X and Y, points on the line represent subgroup means and they are connected as a straight line.
- **Independent:** The observations are independent to each other, which is a common assumption for general classical statistical models.
- **Normal:** The X subgroups have normal distribution. Based on this assumption, we could express the full nature of a subgroup only using the mean and variance without any further explanation.
- **Equal variance:** All the subgroup variances are equal. Based on this assumption, we can simplify the calculation procedure as obtaining a common variance estimate instead of calculating each subgroup variance separately.

Data source, variables and Methodologies

Data source

The purpose of the study is to identify the best linear model by checking the assumptions of residuals in simple linear regression model. The data of imports and exports of the country INDIA is taken from 1991 to 2017. This data can be downloaded from the world bank website by using the library WDI.

Variables

In the present study Exports is the dependent variable and imports is the independent variable, Linear regression.

Methodology

In Simple linear regression, draw a straight line as ‘close’ as possible to the data points. In other words, the difference between the line and data points should be minimized. To accomplish the concept mathematically we use the ‘least squares method’. the vertical difference between the line and the corresponding data point represents a deviation from the estimated mean (on the line) and actually observed Y value for a specific X value. The deviation is called an ‘error’. The theoretical distribution of errors is assumed as normal distribution. The mean of errors in a subgroup is zero and the variance of subgroups is set to a value, $\sigma 2$.The errors are squared and summed to make a squared sum of errors. The least square method aims to minimize the squared sum of errors and to obtain the slope and intercept values. The estimated line should show best fit which lies closest to the observed data points. Specifically, we use the first derivatives of squared sum of errors and set the value into zero to get the slope and intercept which produce minimum values of sum of squared errors.
The main idea is that the line represents means of Y for subgroups of X, not individual data points . Subgroups of X have certain distributions around their means. Therefore, individual data points can have some distances from the straight line by various amounts of deviations from their means, this model is a ‘regression model’ and especially a ‘simple linear regression model’ when only one X variable is included. In the regression model the values on the line is considered as the mean of Y corresponding to each X value and we call the Y values on the line as $\hat{Y}$ (Y hat), the predicted Y values.

Least Squares Method

The method of regression analysis begins with a set of data points to be plotted on an x- and y-axis graph. An analyst using the least-squares method will generate a line of best fit that explains the potential relationship between independent and dependent variables.The least-squares method provides the overall rationale for the placement of the line of best fit among the data points

being studied. The most common application of this method, which is sometimes referred to as "linear" or "ordinary," aims to create a straight line that minimizes the sum of the squares of the errors that are generated by the results of the associated equations, such as the squared residuals resulting from differences in the observed value, and the value anticipated, based on that model.The line of best fit determined from the least squares method has an equation that tells about the relationship between the data points. Line of best-fit equations may be determined by computer software models, which include a summary of outputs for analysis, where the coefficients and summary outputs explain the dependence of the variables being tested.

Coefficient of Determination

The coefficient of determination or R squared method is the proportion of the variance in the dependent variable that is predicted from the independent variable. It indicates the level of variation in the given data set.

- The coefficient of determination is the square of the correlation(r), thus it ranges from 0 to 1.
- With liner regression, the coefficient of determination is equal to the square of the correlation between the x and y variables.
- If R2 is equal to 0, then the dependent variable cannot be predicted from the independent variable.
- If R2 is equal to 1, then the dependent variable can be predicted from the independent variable without any error.
- If R2 is between 0 and 1, then it indicates the extent that the dependent variable can be predictable. If R2 of 0.50 means, it is 50 percent of the variance in the y variable is predicted from the x variable. If 0.80 means, 80 percent of the variance in the y variable is predicted from the x variable, and so on.

The value of R2 shows whether the model would be a good fit for the given data set. In the context of analysis, for any given per cent of the variation, it(good fit) would be different. For instance, in a few fields like rocket science, R2 is expected to be nearer to 100 %. But R2 = 0(minimum theoretical value), which might not be true as R2 is always greater than 0.The value of R2 increases after adding a new variable predictor. Note that it might not be associated with the result or outcome. The R2 which was adjusted will include the same information as the original one. The number of predictor variables in the model gets penalized. When in a multiple linear regression model, new predictors are added, it would increase R2. Only an increase in R2 which is greater than the expected(chance alone), will increase the adjusted R2.R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable . Whereas correlation explains the strength of the relationship between an independent and dependent variable, R-squared explains to what extent the variance of one variable explains the variance of the second variable. So, if the R2 of a model is 0.50, then approximately half of the observed variation can be explained by the model's inputs.

- The coefficient of determination is a complex idea centered on the statistical analysis of models for data.
- The coefficient of determination is used to explain how much variability of one factor can be caused by its relationship to another factor.
- This coefficient is commonly known as R-squared (or R2), and is sometimes referred to as the "goodness of fit."
- This measure is represented as a value between 0.0 and 1.0, where a value of 1.0 indicates a perfect fit, and is thus a highly reliable model for future forecasts, while a value of 0.0 would indicate that the model fails to accurately model the data at all.

Review of results

- From the Box-plot graph there is no identification of outliers in the data
- The two variables Exports and imports are highly correlated, the correlation between these two variables is 0.99.
- From model1 the P value of intercept is lessthan 0.05, so intercept is statistically significant. The p value of variable imports is smaller than 0.05, so this variable is also statistically significant.

The overall P value is also very smaller than 0.05, so simple linear regression model is valid for this data

Testing Heteroscedasticity

H0: Hetroscedasticity is not present
H1: Hetroscedasticity is present
In model1 $p<0.05$ we reject H0, so heteroscedasticity present in the error terms in model1

Testing Autocorrelation

H0: There is no autocorrelation
#H1: there is autocorrelation
In model1 $P<0.05$, so we reject our H0, so there is autocorrelation present in the error terms

Testing of Normality

H0: Residuals are normally distributed
H1:Residuals are not normally distributed

In model1 If P>0.05 then we do not reject our H0, so the residuals are normally distributed
In this model the assumptions of simple linear regression Autocorrelation, Heteroscedasticity is not satisfied so this model is not suitable for the data. So we construct another model by taking the lag of the variable imports.
*Model2

Testing of Heteroscedasticity

In mode l2, Heteroscedasticity test the value of p>0.05 we accept H0, so there is no heteroscedasticity in the error terms

Testing of autocorrelation
In model 2, Autocorrelation test the value of P>0.05, so there is no evidence to reject our H0, so there is no autocorrelation in the error terms

Tesing of Normality
In model2 Normality test the value of P is greater than 0.05, so the error terms are normally distributed
Model2 satistying the assumptions of residuals, so this model is the best model.
*Model3

Testing of Heteroscedasticity
In model 3, Heteroscedasticity test the value of p>0.05 we accept H0, so there is no heteroscedasticity in the error terms

Testing of autocorrelation:
In model 3, Autocorrelation test the value of *P*<0.05, so presence of autocorrelation in the error terms

Tesing of Normality
In model3 Normality test the value of P is greater than 0.05, so the error terms are normally distributed
Model3 is not satisfying the residual assumption of Autocorrelation, model 3 is not a best fit.

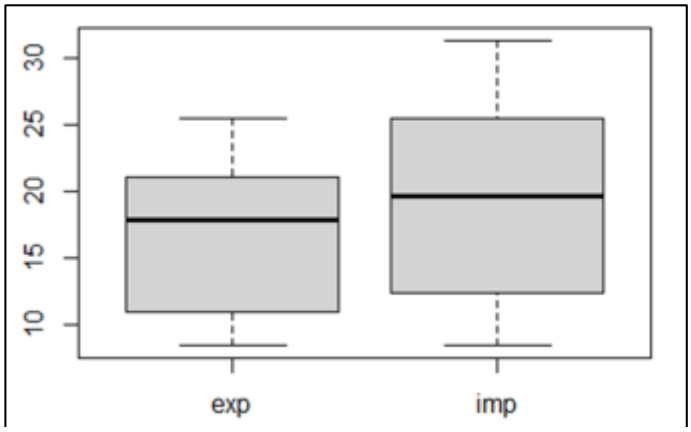

Fig 1: *Box plot

*correlation among variables
cor(exp,imp)
[1] 0.9901924
cor(ie)
exp imp
exp 1.0000000 0.9901924
imp 0.9901924 1.0000000

*Model1
model1<-lm(exp~imp,data=ie)
summary(model1)
##
Call:
lm(formula = exp ~ imp, data = ie)
##
Residuals:
Min 1Q Median 3Q Max
-1.42936 -0.51426 -0.04813 0.52763 1.66369
##
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.10388 0.43741 4.81 6.1e-05 ***
imp 0.76244 0.02152 35.44 < 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
Residual standard error: 0.8094 on 25 degrees of freedom
Multiple R-squared: 0.9805, Adjusted R-squared: 0.9797
F-statistic: 1256 on 1 and 25 DF, p-value: < 2.2e-16
Fitted values by using the model:-

```
fitted(model1)
## 1 2 3 4 5 6 7 8
## 8.579613 9.415766 9.588954 9.873142 11.271008 10.905677 11.198720 11.772325
## 9 10 11 12 13 14 15 16
## 12.292712 12.704544 12.347105 13.726659 14.031819 17.081689 19.179709 20.750416
## 17 18 19 20 21 22 23 24
## 21.078286 24.421028 21.829881 22.578534 25.803023 25.937077 23.767169 21.892303
## 25 26 27
## 18.961120 18.057273 18.839899
```

Testing the assumptions of Error terms:-

*Testing of heteroscedasticity

```
bptest (model1)
##
## studentized Breusch-Pagan test
##
## data: model1
## BP = 8.5687, df = 1, p-value = 0.00342
```

Testing of Autocorrelation

```
lmtest::dwtest(model1)
##
## Durbin-Watson test
##
## data: model1
## DW = 1.1434, p-value = 0.00481
## alternative hypothesis: true autocorrelation is greater than 0
```

.Normality test

```
shapiro. test(residuals(model1))
## Shapiro-Wilk normality test
##
## data: residuals(model1)
## W = 0.98327, p-value = 0.9276
```

*Model2

```
model2<-lm(exp~lag(imp),ie)
summary(model2)
##
## Call:
## lm(formula = exp ~ lag(imp), data = ie)
##
## Residuals:
## Min 1Q Median 3Q Max
## -3.8738 -0.9085 -0.1558 0.6093 3.2595
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.49202 0.86989 4.014 0.000508 ***
## lag(imp) 0.71000 0.04292 16.541 1.27e-14 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.61 on 24 degrees of freedom
## (1 observation deleted due to missingness)
## Multiple R-squared: 0.9194, Adjusted R-squared: 0.916
## F-statistic: 273.6 on 1 and 24 DF, p-value: 1.268e-14
```

Here multiple R2 and adjusted R2 is not so far away so the variable imports explains 91.6% variabity of exports by using the linear regression model

Fitted values by using the model:-

```
fitted(model2)
## 2 3 4 5 6 7 8 9
## 9.52238 10.30103 10.46230 10.72694 12.02867 11.68847 11.96135 12.49551
## 10 11 12 13 14 15 16 17
```

```
## 12.98010 13.36361 13.03076 14.31543 14.59960 17.43971 19.39344 20.85612
## 18 19 20 21 22 23 24 25
## 21.16144 24.27428 21.86134 22.55850 25.56123 25.68606 23.66539 21.91947
## 26 27
## 19.18988 18.34820
```

*Testing of Heteroscedasticity

```
bptest(model2)
##
## studentized Breusch-Pagan test
##
## data: model2
## BP = 2.3613, df = 1, p-value = 0.1244
```

Autocorrelation

```
lmtest::dwtest(model2)

## Durbin-Watson test
##
## data: model2
## DW = 1.858, p-value = 0.2809
## alternative hypothesis: true autocorrelation is greater than 0
```

Normality test

```
shapiro.test(residuals(model2))
##
## Shapiro-Wilk normality test

data: residuals(model2)
W = 0.95949, p-value = 0.3815
```

*Model3

```
Model3<-lm(exp~log(imp),ie)
summary(model3)
##
## Call:
## lm(formula = exp ~ log(imp), data = ie)
##
## Residuals:
## Min 1Q Median 3Q Max
## -1.4658 -0.5095 -0.0617 0.3444 2.2877
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) -22.4748 1.2379 -18.16 6.54e-16 ***
## log(imp) 13.6301 0.4277 31.87 < 2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.898 on 25 degrees of freedom
## Multiple R-squared: 0.976, Adjusted R-squared: 0.975
## F-statistic: 1016 on 1 and 25 DF, p-value: < 2.2e-16
```

Testing of Heteroscedasticity

```
bptest(model5)

## studentized Breusch-Pagan test
##
## data: model3
## BP = 0.10106, df = 1, p-value = 0.7506
```

Autocorrelation

```
lmtest::dwtest(model3)
##
## Durbin-Watson test
##
```

```
## data: model3
## DW = 1.0968, p-value = 0.003083
## alternative hypothesis: true autocorrelation is greater than 0
```

Normality test

```
shapiro.test(residuals(model5))
##
## Shapiro-Wilk normality test
##
## data: residuals(model5)
## W = 0.95217, p-value = 0.2419
```

Conclusion

In first model the overall P value is significant, so the simple linear regression model is valid but in the first model the assumption of simple linear regression is not valid, so we go for another form of simple linear regression, in the model3 the assumptions autocorrelation is not satisfied, but in the model2 the assumptions of autocorrelation, heteroscedasticity, normality are satisfied .so model2 is the best model.

References

1. Kothari CR, Garg Gaurav. Research Methodology Methods and Techniques, New Age International Publishers, New Delhi 2014.
2. David Letcher W, Joao Neves S. Determinants of undergraduate business student satisfaction 2010.
3. Gupta SP, Gupta MP. Business Statistics, Sultan Chand & Sons, New Delhi 2009.
4. James Hamilton D. Time Series Analysis, Princeton University Press Princeton, New jersey.
5. Spyros Makridakis, Steven C. Wheelwright, Rob Hyndman J. Forecasting Methods and Applications Third edition.
6. Schneider A, Hommel G, Blettner M. Linear regression analysis: Part 14 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2010;107:776-82.
7. Freedman DA. Statistical Models: Theory and Practice. Cambridge, USA: Cambridge University Press 2009.
8. Chan YH. Biostatistics 201: Linear regression analysis. Age (years). Singapore Med J 2004;45:55-61.
9. Chan YH. Biostatistics 103: Qualitative data – Tests of independence. Singapore Med J 2003;44:498-503.
10. Gaddis ML, Gaddis GM. Introduction to biostatistics: Part 6, correlation and regression. Ann merg Med 1990;19:1462-8.