Assumption checking of a Multiple Linear Regression Model

Tanvi Koyande[1]
[1]Assistant Professor
[1]Department of Mathematics and Statistics,
[1]Mulund College of Commerce, Mumbai, India

Abstract: A statistical assessment of the average relationship between two or more variables is called regression analysis. There are two kinds of variables used in regression analysis. One is the independent variable, represented by (X), while the other is the dependent variable, typically represented by (Y). Another name for the independent variables is predictors or regressors. Using regressors, the goal is to predict the value of the dependent variable. When one independent variable (X) is used to predict the dependent variable, the mathematical model is known as a Simple Linear Regression Model and is expressed as $Y = b_0 + b_1X$. where 'b_0' denotes the intercept and 'b_1' the regression coefficient, and b_0 and b_1 are the constants. The mathematical model known as a multiple linear regression model is given by $Y = b_0 + b_1X_1 + b_2X_2 + ...+b_nX_n$ and is employed when multiple independent variables are utilized to predict the dependent variable. Estimating the value of the regression coefficients is the goal of regression analysis. The least squares method is used to estimate these regression coefficients. However, these estimates will only be accurate and unbiased if the model meets all of the assumptions. This article addresses a linear regression model's basic assumptions and methods for resolving violations.

Index Terms: Multiple linear regression, assumptions

I. INTRODUCTION

Several assumptions form the foundation of the multiple linear regression model. It is crucial to meet each and every assumption. If the assumptions are not met, the resulting estimates may be biased and will not accurately forecast the value of the dependent variable. The five assumptions of multiple linear regression model are;

1) Linear relationship: The dependent variable and each of the independent variables possess a linear relationship, according to the first multiple linear regression assumption. For each independent variable (X) and dependent variable (Y), a scatter diagram is analysed to verify the linearity assumption. Plotting the residuals versus the fitted values is another technique. If the assumption of linearity is violated, the dependent and/or independent variables can be transformed nonlinearly. Transformations like square root or log can be used. Another option is to include another regressor that is a nonlinear function of one of the independent variables. For instance, X^2 is introduced. Eliminating the independent variable, which has no linear relationship to the dependent variable, is an additional strategy.

2) Independence of errors: According to the conventional linear regression model, the residual term associated with any observation is independent of the residual term associated with any other observation. Autocorrelation is the term used to describe the correlation between the residual terms. The presumption that there is no autocorrelation in the error terms fails if autocorrelation is present. A Durbin-Watson test is the most straightforward method to ascertain whether this assumption is true. The test will provide values between 0 and 4, with values between 0 and 2 indicating positive autocorrelation and values between 2 and 4 indicating negative autocorrelation. The midpoint, or a value of 2, indicates the absence of autocorrelation. Ordinary least square estimators are linear, unbiased, consistent, and normal in the presence of autocorrelation, but they are inefficient. The confidence intervals are also broader. Therefore, generalized least square estimators are used in place of ordinary least square estimators when autocorrelation is present.

3) Homoscedasticity: The variance of every residual term is assumed to be constant in the linear regression model. This is a homoscedasticity assumption. Heteroscedasticity is the absence of homoscedasticity. When heteroscedasticity exists in a regression study, the model's results become untrustworthy. To ascertain heteroscedasticity, a scatterplot of residuals versus fitted values might be utilized. It suggests heteroscedasticity if there is any pattern, such a cone-shaped distribution. Heteroscedasticity can also be determined with the Breusch-Pagan Godfrey test. By taking the log, square root, or cube root of the dependent variable, heteroscedasticity can be eliminated. The ordinary least square estimators are unbiased and consistent but inefficient when heteroscedasticity is present. While weighted least square estimators allocate weight to individual data points based on the variance of their fitted values, ordinary least square estimators

apply equal weight to each observation. Therefore, heteroscedasticity can also be fixed using weighted least square estimators.

4) Normality: The residuals are presumed to be normally distributed in the multiple linear regression model. A QQ plot of the residuals or statistical tests like the Shapiro-Wilk or Kolmogorov-Smironov can also be used to test it. Applying a nonlinear transformation to the dependent variable, such as taking the square root, log, or cube root of all of the dependent variable's values, might fix the situation if the normality assumption is violated.

5) No multicollinearity: The independent variables in a linear regression model are assumed to be uncorrelated. The independent variables shouldn't be multicollinear. The Variance Inflation Factor (VIF) approach is the most effective way to verify the assumption. Severe multicollinearity is indicated if the VIF value is higher than 5. Remedial measures to address multicollinearity include eliminating an independent variable that can be explained by others or, in cases of severe multicollinearity, using principal component regression or ridge regression.

II. DATA

The 'mtcars' dataset from R-studio is the one used in this article. R is the software used to run the multiple regression model. There are eleven variables and thirty-two observations in the dataset. The dependent variable (Y) that we are trying to predict is mileage per gallon (mpg). Weight of engine (wt), horsepower (hp), and displacement (disp) are the independent variables being examined. Out of the ten variables that may be used, this article has only taken into account the three independent variables mentioned above for the sake of simplicity. The snapshot of the data is shown below.

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

III. ASSUMPTIONS CHECKING

The multiple linear regression for one dependent variable and three independent variables is given by

$$Y = b_o + b_1X_1 + b_2X_2 + b_3X_3 + e$$

Where, Y: mileage per gallon (mpg)

X_1: displacement (disp)

X_2: horsepower (hp)

X_3: weight of engine (wt)

b_o, b_1, b_2, b_3 are the regression coefficients and e is the random error.

The regression coefficients are estimated by the method of least square and the model obtained is,

$$Y = 37.105505 - 0.000937X_1 - 0.031157X_2 - 3.800891X_3$$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.105505   2.110815  17.579  < 2e-16 ***
disp        -0.000937   0.010350  -0.091  0.92851
hp          -0.031157   0.011436  -2.724  0.01097 *
wt          -3.800891   1.066191  -3.565  0.00133 **
```

The regression coefficient (b3) implies that one unit change in weight will change (decrease) the mileage of the car by 3.800 units. Similar interpretation can be made about effect of other independent variables.

The significance of the regression parameters is tested using Global testing and individual t-test

```
Residual standard error: 2.639 on 28 degrees of freedom
Multiple R-squared:  0.8268,    Adjusted R-squared:  0.8083
F-statistic: 44.57 on 3 and 28 DF,  p-value: 8.65e-11
```

Global Testing

$H_0: b_1 = b_2 = b_3 = 0$ i.e., The independent variables do not contribute to the dependent variable

$H_1:$ not H_0 i.e., At least one of the regression coefficients is not zero.

Based on the above result, we reject the null hypothesis because the p-value is less than 0.05.

As a result, the dependent variable is significantly impacted by the independent variables.

Individual t-test

We determine the significance of the independent variables by rejecting the Global test's null hypothesis. We are now interested in determining which independent variables have a substantial effect on the dependent variable.
The weight of the engine (wt) and horsepower (hp) t-test has a pvalue of less than 0.05. They therefore have a big impact on the model.

R-square
Approximately 80% of the variation in the car's mileage is explained by the displacement, horsepower, and engine's weight.

Multicollinearity
The VIF method is used to check for multicollinearity
VIF $= \frac{1}{1-r^2}$ where r^2 is the multiple r^2 for the regression of X_j on the other independent variables.

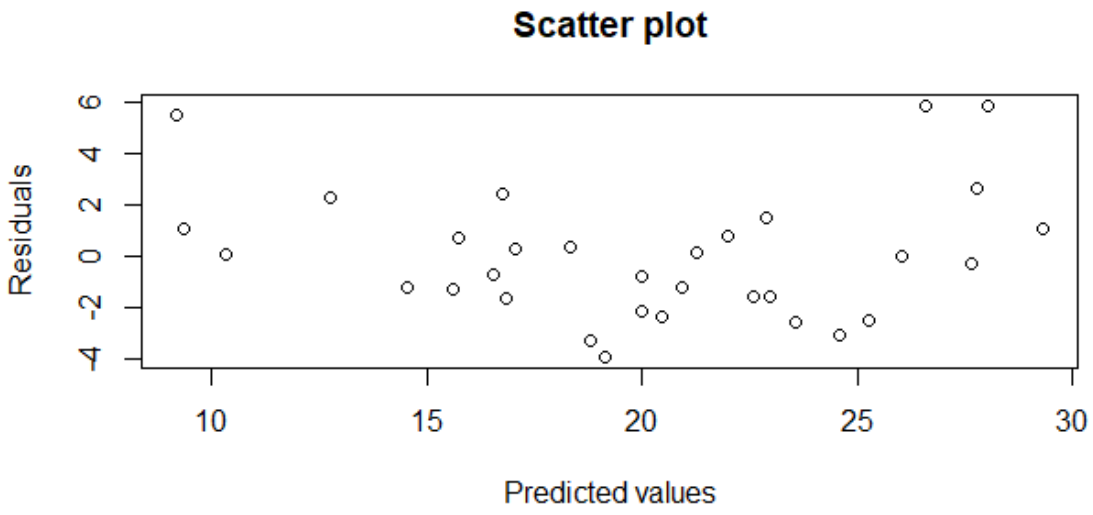
```
    disp       hp       wt
7.324517 2.736633 4.844618
```

The displacement variable's VIF is larger than 5, it indicates severe multicollinearity. Consequently, it suggests that the displacement variable should be removed from the model.

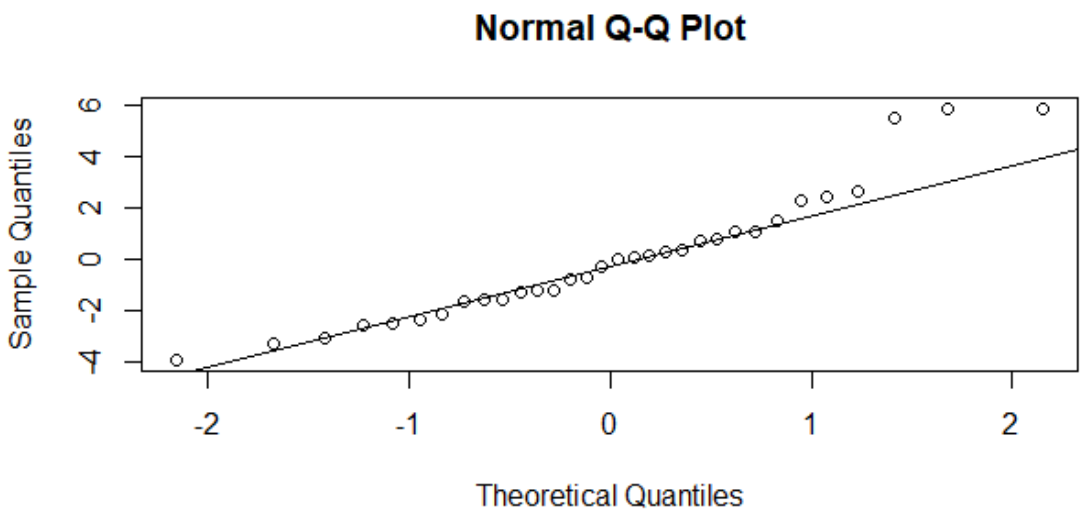
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.22727    1.59879  23.285  < 2e-16 ***
hp          -0.03177    0.00903  -3.519  0.00145 **
wt          -3.87783    0.63273  -6.129 1.12e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.593 on 29 degrees of freedom
Multiple R-squared:  0.8268,    Adjusted R-squared:  0.8148
F-statistic: 69.21 on 2 and 29 DF,  p-value: 9.109e-12
```

The R output of a multiple regression model with the displacement variable removed is shown above. It is evident that both the model and each of its parameters have significant values. Also, after resolving the multicollinearity issue, the model's efficiency is marginally increased to 81%.

Linearity
There is no visible pattern in the residual plot that suggests a linear relationship between the independent and dependent variables.



Normality
The points on the QQ plot roughly follow a straight diagonal line thus the normality assumption is met.



Homoscedasticity

The residual plots already show that there is no heteroscedasticity. On the other hand, homoscedasticity can also be verified using the Breusch-Pagan Godfrey test.

H_0: constant error variance

H_1: not H_0

```
        studentized Breusch-Pagan test

data:  m2
BP = 0.88072, df = 2, p-value = 0.6438
```

Do not reject the null hypothesis because the p-value is greater than 0.05. As a result, the error terms exhibit constant variance.

Autocorrelation

H_0: $\rho = 0$ i.e., no autocorrelation

H_1: $\rho \neq 0$

```
 lag Autocorrelation D-W Statistic p-value
   1       0.2954091      1.362399   0.032
 Alternative hypothesis: rho != 0
```

As p-value is less than 0.05, the null hypothesis is rejected. As a result, the regression model's residuals exhibit autocorrelation.

IV. Conclusion

An essential tool for predicting observations with the aid of an independent variable is linear regression. As mentioned in the article, before utilizing the regression model for prediction, make sure its assumptions are met. If the model does not meet the assumptions, it will estimate the regression coefficients, which will be misleading and forecast incorrect values in the future. It has been noted that inaccurately calculated regression coefficients also have an impact on the model's overall effectiveness. Therefore, while linear regression is a useful tool, it is important to take into account its assumptions and limits when using it to solve real-world issues.

References

[1] Turoczy Zsuzsanna, Liviu Marian. Multiple regression analysis of performance indicators in the ceramic industry. Procedia
Economics and Finance 3 (2012) 509-514

[2] Gulden Kaya Uyanik, Nese Guler. A study on multiple linear regression analysis. Procedia - Social and Behavioral Sciences
106 (2013) 234 – 240

[3] https://www.statology.org/multiple-linear-regression-assumptions/

[4] https://medium.com/@emilywinslet/understanding-linear-regression-with-real-life-examples-dff7cd851e4e