AN ANALYSIS OF RESIDUALS IN MULTIPLE REGRESSIONS

Ahmad A. Suleiman¹, Usman A. Abdullahi², Umar A. Ahmad³

¹, ², ³Postgraduate Student, Department of Mathematics / Statistics, Sharda University, Greater Noida, (India)

ABSTRACT
This paper concentrates on residuals analysis to check the assumptions for a multiple linear regression model by using graphical method. Specifically, we plot the residuals and standardized residuals given by model against predicted values of the dependent variables, normal probability plot, histogram of residuals and Quantile plot of residuals. Finally, we explained the concept of heteroscedasticity which we used to check the assumption that the residuals in regression model the same variance. As an example, a formal method to detect the presence of heteroscedasticity by Breusch Pagan method using eview was presented.

I. INTRODUCTION
The main aim of regression modelling and analysis is to develop a good predictive relationship between the dependent (response) and independent (predictor) variables. Regression diagnostics plays a vital role in finding and validating such a relationship. In this study, we discuss issues that arise in the development of a multiple linear regression model. Consider the following standard multiple linear regression model:

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon \]

where \( Y \) is a response variable and \( X \)'s are predictor variables, \( \beta \)'s are the (regression) parameters to be estimated from data, and \( \epsilon \) is the error or residual.

The validity of the inference methods depends on the error term \( \epsilon \), satisfying these assumptions:

- **Independence**: Observations (and hence residuals) are statistically independently distributed.
- **Normality**: The residuals are normally distributed with zero mean.
- **Homoscedasticity**: All the observations (and hence residuals) have the same variance.
- **Multicollinearity**: No linear correlation between independent variables

II. METHOD AND ANALYSIS
Here is a hypothetical data on Consumption, Export and GDP

<table>
<thead>
<tr>
<th>CONSUMPTION</th>
<th>EXPORT</th>
<th>GDP</th>
</tr>
</thead>
<tbody>
<tr>
<td>50.35718</td>
<td>1.436314</td>
<td>35.06</td>
</tr>
<tr>
<td>50.44603</td>
<td>1.414639</td>
<td>35.66</td>
</tr>
<tr>
<td>57.87973</td>
<td>1.529996</td>
<td>37.83</td>
</tr>
<tr>
<td>72.30876</td>
<td>1.746588</td>
<td>41.4</td>
</tr>
</tbody>
</table>
There are many diagnostics we can perform on the residuals. Here are the most important ones:

- **Normal Probability Plot**

To diagnose if the errors are normally distributed, we draw a normal probability plot of the residuals. The residuals should fall approximately on a diagonal straight line in this plot.
We can also plot a histogram to see if they are lumpy in the middle symmetric tails.

From our histogram, the residuals appear to be normally distributed.

We can make a Quantile Plot where the residuals are plotted against their percentage (empirical cumulative probability) point (0 to 1): if normality holds this should have an S-shape.
The S-shape of the curve seems to suggest that normality assumption is satisfied. To examine if the errors are independent. We look at the plot of residuals against estimated values to ensure that the residuals are randomly scattered above and below the 0 horizontal.

Although the mean of the residual may be accepted to be zero at each x-value, the variance seems to increase with x-value suggesting a possible violation of the homoscedasticity assumption.

- **Standardized Residuals**

The Standardized residuals is more useful since they are more easy to interpret. We plot Standardized residuals against estimated values to identified outliers in the dependent variable space. Large values (greater than 2 or 3 in absolute magnitude) indicate possible problems.
Case no. 34 with a Standardized value of 2.246 can be considered an outlier. If we delete this case from the data set and recompute the regression, the fit become better.

**2.2 Removal of Heteroscedasticity from Regression Mode**

We want remove heteroscedasticity from the regression model because heteroscedasticity is not desirable that is the residuals should be homoscedastic.

There are many ways to remove heteroscedasticity from the model. One of the most popular ways is to convert all the variables into log, which is known as log transformation.

Let us consider another multiple regression example with quite a few predictor variables. In our model, we have three variables such as consumption, export and growth domestic product (gdp). Here consumption is the dependent variable while the rest two are independent variables.

If we see heteroscedasticity in the model after estimation, we need to convert all the three variables into log that is: Consumption > log(consumption), export > log(export), gdp > log(gdp).

Once we run the model with log variables, heteroscedasticity will be removed and homoscedasticity will be appeared. As we know, homoscedasticity is desirable.

Our hypothesis:

Null hypothesis: Homoscedasticity

Alternative hypothesis: Heteroscedasticity

**Basic Software output yield:**
Now, we are to check whether this model has heteroscedasticity or not. Using Breusch-Pagan-Godfrey test we have p-value corresponding to $R^2$ is 0.0394 which is less than 5%, hence we reject the null hypothesis and accept the alternative hypothesis. Meaning that this model got heteroscedasticity. In other words, the residuals are heteroscedasticity so that the model is not desirable.

After transforming the variables Breusch-Pagan-Godfrey shows that the p-value corresponding to the observed $R^2$ is 0.7207, which is more than 5%. Meaning that, we cannot reject the null hypothesis. Hence the model is homoscedasticity which is desirable.

2.3 Multicollinearity Problem and Regression Model

How multicollinearity affects any estimated regression model?

Here is our model:

$$Y = C + X_1 + X_2 + X_3 + X_4 + X_5 + X_6 \ldots$$

Here $Y$ is the dependent variable and the rest are independent variables. After estimating Model (1.1), we saw that only $X_5$ is significant while others are not. We suspect that, there is a problem of multicollinearity in Model (1.1) that is why most of the variables have become insignificant.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>133.22</td>
<td>3292.415</td>
<td>4.046549</td>
<td>0.0004</td>
</tr>
<tr>
<td>X1</td>
<td>-426.9264</td>
<td>1037.802</td>
<td>-0.411376</td>
<td>0.6839</td>
</tr>
<tr>
<td>X2</td>
<td>876.4252</td>
<td>3796.148</td>
<td>0.230872</td>
<td>0.8191</td>
</tr>
<tr>
<td>X3</td>
<td>-0.835029</td>
<td>0.434370</td>
<td>-1.922391</td>
<td>0.0648</td>
</tr>
<tr>
<td>X4</td>
<td>1.061489</td>
<td>1.007373</td>
<td>1.053721</td>
<td>0.3010</td>
</tr>
<tr>
<td>X5</td>
<td>0.409495</td>
<td>0.131346</td>
<td>3.117691</td>
<td>0.0042</td>
</tr>
<tr>
<td>X6</td>
<td>3.631322</td>
<td>3.989728</td>
<td>0.910168</td>
<td>0.3705</td>
</tr>
</tbody>
</table>

- R-squared: 0.457529
- Adjusted R-squared: 0.341285
III. CONCLUSION

1. We estimate the value of the regression residuals for each value of $y$:

$$\hat{e} = y - \hat{y}$$

Which is the observed value – the predicted (or expected) value.

2. We made sure the removal of multicollinearity by dropping the appropriate highly correlated independent variables before studying the residuals.

3. We have dealt with an example dealing with GDP in chapter four where we detected and removed heteroscedasticity by the methods suggested by Bruesch-Pagan.

   Another problem that can affect adversely our study of multiple regression is heteroscedasticity.

4. In conclusion we have dealt with
   - estimates of regression coefficients;
   - their standard errors, confidence intervals and tests of their significance;
   - analysis of variance for regression which produces an overall test of significance and an estimate of the error variance as the residual (error) mean-square;
   - multiple correlation coefficient, its square, and its adjusted value, which give a measure of how much of the variation has been captured by the predictor variables and hence how useful the regression is;
   - Using the saved residuals we can
     - make suitable plots to examine the assumption of normality;
     - carry out a formal test of significance for normality;
     - make suitable plots to examine the assumption of homoscedasticity;
     - make suitable plots to examine the assumption of independence.

Thus some basic diagnosis can be performed of the validity of assumptions under which a standard regression analysis is carried out using this regression output.

IV. ACKNOWLEDGEMENTS

Our infinite gratitude goes first and foremost to Almighty Allah for sparing my life till this moment.

Many special thanks and appreciation goes to our honorable supervisor and at the same time our esteemed lecturer Prof. U.V. Balakrishnan who his suggestions, corrections and encouragements made this work reality.

We are also grateful to the entire lecturers in the Department of Statistics in person of Dr. N.M. Chahda, Dr. Sweta Srivastav, Dr. Krushidalam and others whose wealth of experience and knowledge I have benefited.

REFERENCES


[4]. D. A. Belsley, E. Kuh, and R. E. Welsch, Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.


[7]. Douglas Montgomery and Elizabeth Peck, Introduction to Linear Regression Analysis


