 ISBN: 978-81-931039-3-7

ANALYSIS OF MULTICOLLINEARITY IN MULTIPLE REGRESSIONS

Ahmad A. Suleiman

Postgraduate Student, Department of Mathematic / 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. However, we introduced the concept of multicollinearity to check whether one of the assumptions of the linear regression model thatthere is no multicollinearity among the explanatory variables is satisfied. We also gave an example that indicated the presence of multicollinearity in the regression model using eview (statistical software).

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. Multicollinearity analysis 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:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon$$

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

The validity of the inference methods depends on the error term ε, satisfying these assumptions;

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

II. METHOD METHODOLOGY

In statistics, **multicollinearity** (also **collinearity**) is a phenomenon in which two or more predictor variables in a multiple regression model are highly correlated, meaning that one can be linearly predicted from the others with a substantial degree of accuracy. In this situation the coefficient estimates of the multiple regressions may change erratically in response to small changes in the model or the data. Multicollinearity does not reduce the predictive power or reliability of the model as a whole, at least within the sample data set; it only affects calculations regarding individual predictors.

Collinearity is a linear association between *two*explanatory variables. Two variables are perfectly collinear if there is an exact linear relationship between them. For example, X_1 and X_2 are perfectly collinear if there exist parameters λ_0 and λ_1 such that, for all observations i, we have

$$X_{2i} = \lambda_0 + \lambda_1 X_{1i}.$$

2.1 Sources of Multicollinearity

There are several sources of multicollinearity. As Montgomery and Peck note, multicollinearity may be due to the following factors:

1. *The Data collection method employed*, for example, sampling over a limited range of the values taken by the regressors in the population.

2. *Constraints on the model or in the population being sampled.* For example, in the regression of electricity consumption on income (X_2) and house size (X_3) there is a physical constraint in the population in that families with higher incomes generally have larger homes than families with lower incomes.

3. *Model specification*, for example, adding polynomial terms to a regression model, especially when the range of the *X* variable is small.

1. What is the nature of multicollinearity?

2. Is multicollinearity really a problem?

3. What are its practical consequences?

4. How does one detect it?

5. What remedial measures can be taken to alleviate the problem of multicollinearity?

2.2 The Nature of Multicollinearity

The term *multicollinearity*is due to Ragnar Frisch. Originally it meant the existence of a “perfect,” or exact, linear relationship among some or all explanatory variables of a regression model. For the k -variable regression involving explanatory variable $X_1, X_2, ..., X_k$ (where $X_1 = 1$ for all observations to allow for the intercept term), an exact linear relationship is said to exist if the following condition is satisfied:

$$\lambda_1 X_1 + \lambda_1 X_1 + ... + \lambda_k X_k = 0 \quad \ldots\ldots\ldots (3.1)$$

where $\lambda_1, \lambda_2, ..., \lambda_k$ are constants such that not all of them are zero simultaneously.

Today, however, the term multicollinearity is used in a broader sense to include the case of perfect multicollinearity, as shown by (3.1), as well as the case where the X variables are intercorrelated but not perfectly so, as follows:

$$\lambda_1 X_1 + \lambda_2 X_2 + ... + \lambda_k X_k + v_i \ldots\ldots\ldots (3.2)$$

where v_i is a stochastic error term.

Consequences of multicollinearity: One consequence of a high degree of multicollinearity is that, even if the matrix $X^T X$ is invertible, a computer algorithm may be unsuccessful in obtaining an approximate inverse, and if it does obtain one it may be numerically inaccurate. But even in the presence of an accurate $X^T X$ matrix, the following consequences arise.

A principal danger of such data redundancy is that of over fitting in regression analysis models. The best regression models are those in which the predictor variables each correlate highly with the dependent (outcome) variable but correlate at most only minimally with each other. Such a model is often called "low noise" and will be statistically robust (that is, it will predict reliably across numerous samples of variable sets drawn from the same statistical population).

Indicators that multicollinearity may be present in a model:

1. Large changes in the estimated regression coefficients when a predictor variable is added or deleted
2. Insignificant regression coefficients for the affected variables in the multiple regression, but a rejection of the joint hypothesis that those coefficients are all zero (using an F -test)
3. If a multivariable regression finds an insignificant coefficient of a particular explanator, yet a simple linear regression of the explained variable on this explanatory variable shows its coefficient to be significantly different from zero, this situation indicates multicollinearity in the multivariable regression.
4. Some authors have suggested a formal detection-tolerance or the variance inflation factor (VIF) for multicollinearity:

$$tolerance = 1 - R_j^2,\ VIF = \frac{1}{tolerance}$$

5. where R_j^2 is the coefficient of determination of a regression of explanator*j* on all the other explanators. A tolerance of less than 0.20 or 0.10 and/or a VIF of 5 or 10 and above indicates a multicollinearity problem.
6. Condition Number is computed by finding the square root of (the maximum eigenvalue divided by the minimum eigenvalue). If the Condition Number is above 30, the regression may have significant multicollinearity; multicollinearity exists if, in addition, two or more of the variables related to the high condition number have high proportions of variance explained. One advantage of this method is that it also shows which variables are causing the problem.

We can derive what is known as the **condition number** k defined as

$$k = \frac{Maximum\ eigenvalue}{\text{Minimum eigenvalue}}$$

and the **condition index (CI)** defined as

$$Cl = \sqrt{\frac{Maximum\ eigenvalue}{\text{Minimum eigenvalue}}} = \sqrt{k}$$

Then we have the following rule: if k is between 100 and 1000 there is moderate to strong multicollinearity and if it exceeds 1000 there is severe multicollinearity. Alternatively, if the $Cl(=\sqrt{k})$ is between 10 and 30, there is moderate to strong multicollinearity and if it exceeds 30 there is severe multicollinearity.

2.3 Remedies for Multicollinearity

1. Make sure you have not fallen into the dummy variable trap; including a dummy variable for every category (e.g., summer, autumn, winter, and spring) and including a constant term in the regression together guarantee perfect multicollinearity.
2. Try seeing what happens if you use independent subsets of your data for estimation and apply those estimates to the whole data set. Theoretically you should obtain somewhat higher variance from the smaller

datasets used for estimation, but the expectation of the coefficient values should be the same. Naturally, the observed coefficient values will vary, but look at how much they vary.

3. Leave the model as is, despite multicollinearity. Drop one of the variables
4. Obtain more data, if possible.
5. Polynomial terms (i.e., for x_1, x_1^2, x_1^3 , etc.) can cause some multicollinearity if the variable in question has a limited range (e.g., [2,4]).
6. Ridge regression or principal component regression or partial least squares regression can be used.

III. ANALYSIS OF RESULT

We use a Longley data having six independent variables X1,…,X6 and one dependent variable Y.

	X1	X2	X3	X4	X5	X6	Y
1	0.94	0.8	2200	540	4000	140	17721
2	0.75	1.03	2300.06	790	9800	85	17768
3	0.6	0.95	1920	580	12343	56	17823
4	1	0.91	890	338	8070	41.8	15163
5	0.5	0.95	7343.3	3100	13290	250	17480
6	0.834	0.88	850.23	402	5110	78.1	15329
7	1	0.89	1678	590	7456	87	16141
8	0.75	0.89	739.1	560	3234	180.9	15326
9	1.5	0.93	1100	1200	3500	400	17115
10	1.5	0.89	274.6	500	1900	240.2	17117
11	0.71	0.86	360	201	4200	49	16127
12	1	0.94	1879.1	569	4975	115.78	17242
13	0.6	0.84	1965	1287	9000	189	17340
14	1.5	0.87	2300	1630	3000	560.5	15108
15	1.05	0.98	450.34	750	2134	370	16098
16	0.6	0.67	587	502	1501	376	15000
17	0.6	0.93	7598	3478.9	7564	460.01	18027
18	0.45	0.85	530	480	8700	65	17894
19	0.6	0.93	650	300	4997	60	12349
20	1	1.03	1550	380	11100	41.87	17011
21	0.45	0.76	1618.3	600	6501	96.78	16537
22	1.15	0.96	2009.8	280	6802	55.1	14123
23	1.15	0.87	1567	270	5234	54.34	13019
24	0.6	0.76	1298	520	2910	200.12	12980
25	1	1	14000	7323.3	25674	180.98	20513
26	1.35	0.87	3780	1801	13210	136	16089
27	0.75	0.91	545	980	3059	345	15944
28	1	0.79	2385.8	400	8312	55.4	14980

 ISBN: 978-81-931039-3-7

29	1.15	0.98	1700	350	3898	110.45	14980
30	1	0.87	550.98	200	8309	20.2	17743
31	0.85	0.71	1769.1	460	9000	52.099	16890
32	0.5	0.79	1100	654	2510	262.7	14980
33	1.3	0.98	845	1010	3208	310.09	18014
34	0.75	0.85	1678.67	734	16340	47.09	22000
35	0.71	0.85	167.67	350	6988	56.67	20744

How multicollinearity affects any estimated regression model?

Here is our model:

$$Y = C + X_1 + X_2 + X_3 + X_4 + X_5 + X_6 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.1)$$

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 multicollinearioty in Model (1.1) that is why most of the variables have become insignificant.

What is Multicollinearity?

If there is exits a high correlation between any two independent variables, problem of multicollinearity arises. A multicollinearity problem makes significant variables insignificant by increasing its standard error. And if the standard error goes up, t-value goes down and hence comes up with high p-value. So, that particular variable becomes insignificant but in reality it is not. In particular,

$$t\text{-statistic} = \frac{estimated\ coefficient}{standard\ error}$$

hence absolute t-statistic and p-value has always opposite relationship. Normally, if the p-value is more than 5% (0.05), then variable is insignificant.

How to Detect Multicollinearity?

We run correlation analysis using all our independent variables only given in Model (1.1) and find there exits high correlation between X_3 and X_4. As a result, problem of multicollinearity arises, so we have to drop one variable from the model, either X_3 or X_4 to solve the problem of multicollinearity.

How to Solve the Problem of Multicollinearity

The guideline is that, we shall drop that variable which has higher p-value out of X_3 and X_4. Higher the p-value, lower the level of significance.

What is Multicollinearity?

If there is exits a high correlation between any two independent variables, problem of multicollinearity arises. A multicollinearity problem makes significant variables insignificant by increasing its standard error. And if the standard error goes up, t-value goes down and hence comes up with high p-value. So, that particular variable becomes insignificant but in reality it is not. In particular,

$$t\text{-statistic} = \frac{estimated\ coefficient}{standard\ error}$$

hence absolute t-statistic and p-value has always opposite relationship. Normally, if the p-value is more than 5% (0.05), then variable is insignificant.

How to Detect Multicollinearity?

We run correlation analysis using all our independent variables only given in Model (1.1) and find there exits high correlation between X_3 and X_4. As a result, problem of multicollinearity arises, so we have to drop one variable from the model, either X_3 or X_4 to solve the problem of multicollinearity.

How to Solve the Problem of Multicollinearity

The guideline is that, we shall drop that variable which has higher p-value out of X_3 and X_4. Higher the p-value, lower the level of significance.

After estimating Model (1.1), and after removing X_4, we saw that more variables have become significant such as X_3, X_5 and X_6. Consequently, problem of multicollinearity have been removed. Normally in a good regression model most of the independent variables should be significant. Therefore, since out of five variables, three are significant so we are happy about the model.

Dependent Variable: Y

Method:

Least Squares

Date: 08/21/15 Time: 05:35

Sample: 1 35

Included observations: 35

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	12307.52	3154.222	3.901920	0.0005
X1	-435.8796	1039.740	-0.419220	0.6781
X2	1273.139	3784.611	0.336399	0.7390
X3	-0.435085	0.211627	-2.055908	0.0489
X5	0.467108	0.119651	3.903908	0.0005
X6	6.477944	2.941538	2.202230	0.0358

R-squared	0.436018	Mean dependent var	16534.71
Adjusted R-squared	0.338779	S.D. dependent var	2062.459

Correlation Analysis

X1 1.000000	0.314350	-0.034371	0.018090	-0.121271	0.220694
X2 0.314350	1.000000	0.282192	0.292033	0.272359	0.005813
X3 -0.034371	0.282192	1.000000	0.953609	0.728938	0.170192
X4 0.018090	0.292033	0.953609	1.000000	0.658299	0.342713
X5 -0.121271	0.272359	0.728938	0.658299	1.000000	-0.343221
X6 0.220694	0.005813	0.170192	0.342713	-0.343221	1.000000

Dependent Variable: Y

Method:
Least Squares
Date: 08/21/15 Time: 05:35
Sample: 1 35
Included observations: 35

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	12307.52	3154.222	3.901920	0.0005
X1	-435.8796	1039.740	-0.419220	0.6781
X2	1273.139	3784.611	0.336399	0.7390
X3	-0.435085	0.211627	-2.055908	0.0489
X5	0.467108	0.119651	3.903908	0.0005
X6	6.477944	2.941538	2.202230	0.0358
R-squared	0.436018	Mean dependent var	16534.71	
Adjusted R-squared	0.338779	S.D. dependent var	2062.459	

IV. CONCLUSION

1. We estimate the value of the regression residuals for each value of $y : \hat{\varepsilon} = y - \hat{y}$

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.

V. ACKNOWLEDGEMENTS

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

Many special thanks and appreciation goes to my honourable supervisor and at the same time my esteemed lecturer Prof. U.V. Balakrishnan who his suggestions, corrections and encouragements made this work reality. I am also grateful to the entire lecturers in the Department of Statistics, Sharda University, India in person of Dr. N.M. Chahda, Dr. SwetaSrivastav, Dr. Krushidalam and others whose wealth of experience and knowledge I have benefited.

My sincere acknowledgement also goes to my seasoned and most wonderful elder brother, Suleiman Abubakar Suleiman for his constant support and motivation.

Last but not the least, I would like to thank my relatives, friends, course mates and entire kwankwasiyya students of sharda university India, particularly, AliyuIsma'il, AbubakarUsman, UsmanAliyu, Aminu Suleiman, Umar Alhassan and Ameer Hassan Abdullahi for a memorable and lovely time together, thanks for the jokes, piece of advice and constructive criticism. I will not forget you all.

REFERENCES

[1]. AshishSen and Muni Srivastava, Regression Analysis: Theory, Methods, and Applications,Springer-Verlag, New York, 1990, p. 92. Notation changed.

[2]. Belsley, David A.; Kuh, Edwin; Welsch, Roy E. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: Wiley. ISBN 0-471-05856-4.

[3]. C. R. Rao, Linear Statistical Inference and Its Applications, John Wiley & Sons, New York, 1965, p. 258.

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

[5]. D. E. Farrar and R. R. Glauber, "Multicollinearity in Regression Analysis: The Problem

[6]. Revisited," Review of Economics and Statistics, vol. 49, 1967, pp. 92–107.

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

[8]. H. Glejser, "A New Test for Heteroscedasticity,'' Journal of the American Statistical Association, vol. 64, 1969, pp. 316–323.

[9]. Hill, R. Carter; Adkins, Lee C. (2001). "Collinearity". In Baltagi, Badi H. A Companion to Theoretical Econometrics. Blackwell. pp. 256–278. doi:10.1002/9780470996249.ch13. ISBN 0-631-21254-X.

[10]. J. T. Webster,"Regression Analysis and Problems of Multicollinearity," Communications in Statistics A, vol. 4, no. 3, 1975, pp. 277–292; R. F. Gunst.

[11]. Johnston, John (1972). Econometric Methods (Second ed.). New York: McGraw-Hill. pp. 159–168.

[12]. John Wiley & Sons, New York, 1982, pp. 289–290. See also R. L. Mason, R. F. Gunst.

[13]. Kmenta, Jan (1986). Elements of Econometrics (Second ed.). New York: Macmillan. pp. 430–442. ISBN 0-02-365070-2.

[14]. Maddala, G. S.; Lahiri, Kajal (2009). Introduction to Econometrics (Fourth ed.). Chichester: Wiley. pp. 279–312. ISBN 978-0-470-01512-4.

[15]. R. Koenker, "A Note on Studentizing a Test for Heteroscedasticity," Journal of Econometrics, vol. 17, 1981, pp. 1180–1200.

[16]. R. L. Mason, "Advantages of Examining Multicollinearities in Regression Analysis," Biometrics, vol. 33, 1977, pp. 249–260.

[17]. T. Breusch and A. Pagan, "A Simple Test for Heteroscedasticity and Random Coefficient

[18]. Variation,'' Econometrica, vol. 47, 1979, pp. 1287–1294.S