DOI : https://doi.org/10.3126/kaladarpan.v4i1.62837

Multi-collinearity in Research and Wayforward

Sudeep Singh Nakarmi, PhD
Lecturer, Central Department of Sociology
Tribhuvan University
Email. sudipnakarmi@gmail.com

Abstract : *In multi-linear regression, there will be two or more independent variables also known as predictor variables. The main task in multi-linear regression is to find how the predictor variable impacts the dependent variable when there is a change in predictor variables by one unit. However, without testing multi-collinearity between these predictor variables, such a model will create difficulties in terms of defining the real impact of the predictor variable (independent) on the predicted variable (dependent). This situation makes us aware in terms of variable selection while conducting multi-linear regression to find the exact impact. To explain this issue, 14 students were selected and asked to fill up the form with their marks obtained in the recent examination, the number of study hours at home, the number of hours hanging out with friends, and the number of copies added during an exam. The research aimed to observe if there is a multi-collinearity issue between predictor variables (number of study hours, number of hours hanging out with friends, and number of copies added during an exam). SPSS-version 25 is used to find a correlation matrix between predictor variables, matrix scatter plot, the value of Tolerance, Variance Inflation Factor(VIF) value, condition index, and variation proportion. Multi-collinearity issue observed in two predictor variables thus further solution explained.*

Keywords: Regression, Multi-collinearity, Value of Tolerance, Variance Inflation Factor, Condition Index, variation proportion.

Introduction

Collinearity is often referred to as a testing situation in multi-linear regression where there is a high linear relationship between predictor variable. That means if there is a high or strong correlation between both the independent or predictor variables, then it would be difficult to find out which independent variable has a real impact on a dependent variable. Such a high linear relationship between predictor variables is called collinearity. For instance, a person starts a home remedy after having a runny nose but as he seems no chance of getting better after a week, then he visits the doctor and takes pills as suggested. The next day, the running nose starts getting better. He may not have an exact idea if his better health status is due to a week-long home remedy or the doctor's pills. It is unclear which (home remedy or doctor's pills) action made him

better. In statistical terms, such confusion between two or more independent variables is called multi-collinearity. If both the independent variables are highly correlated, there is a high chance of multi-collinearity. Neeleman (1973) states that the reason why multi-collinearity is of such importance is that, if it is present, there is a possibility that the equation in question is under-identified and consequently cannot be estimated. However, the occurrence of multi-collinearity is very rare in research. Gujarati and Porter (2009) observed that the inherent relationship between the predictor variable in the real world, as well as the small sample size, design of the models, and the trend of predictor variables, can cause collinearity. Raykov and Marcoulides (2006) states that in regression analysis the presence of multi-collinearity implies that one is using redundant information in the model, which can easily lead to unstable regression coefficient estimates. So when multi-collinearity exists in a data set, the data is considered deficient. Having said this, a high correlation between predictor variables will lead to an ambiguous relationship with the dependent variable and provide a faulty model. Allin (2010) also writes that multi-collinearity is a situation where two or more explanatory variables are highly related. The issue of multi-collinearity can be detected via a different method and in this paper, how SPSS is used to detect via four methods, will be discussed. To avoid such identified multi- collinearity issues in multi-linear regression analysis, either the predictor variable should be changed with a new one or both the predictor variables should be merged.

Condition for confirming presence of multi-collinearity

If the correlation between two or more predictor variables is very high (i.e. $r>0.8$), in such, multi-collinearity issue is suspected in the model which can lead the researcher to an uncertain situation in the result. To find out if there is the exact issue of multi-collinearity, a multi- regression analysis will be conducted where the value of Tolerance should be below 0.1 and Variance Inflation Factor (VIF) value should be larger than 10. Precisely, according to Paul (2006) practical experience indicates that if any of the VIF's exceeds 10, it is an indication that the associated regression coefficients are poorly estimated because of multi-collinearity. If the value of Tolerance is greater than 0.1 and for VIF, if the value is less than 10, we are confident that there is no multi-collinearity issue in the model. Finally, the condition index value must be larger than 15 in case of collinearity suspected and larger than 30 for serious multi-collinearity.

Method

To find out the issue of multi-collinearity in SPSS and how it can be resolved, 14 students from S. college were selected as respondents on a simple random basis. The respondents were asked to fill in the information in terms of MO, HS, HSWF, and NOCA. After data was collected from respondents, the information is tabulated as below in Table 1 where MO denotes Marks Obtained by students, HS denotes Study Hours spent at home, HSWF denotes the number of Hours spent with Friends and NOCA stands for the number of copies added during an exam.

Table 1: Descriptive statistics of test variables

Case Number	MO	HS	HSWF	NOCA
1	77.00	5.00	4.55	4
2	56.00	3.00	3.10	3
3	67.00	4.00	4.13	3
4	98.00	6.00	5.47	4
5	49.00	3.00	3.25	2
6	76.00	5.10	5.13	4
7	57.00	3.00	2.50	3
8	87.00	5.50	5.10	4
9	66.00	4.00	4.05	3
10	87.00	5.00	4.86	3
11	71.00	4.30	4.20	3
12	68.00	4.30	4.10	3
13	81.00	5.30	5.10	4
14	92.00	5.50	5.20	4

Source: S. College Student

All the data are entered in SPSS version 25. In the first step, Karl Pearson correlation is applied among three predictor variables as all the variables are in scale measurement. After finding a high correlation between two independent or predictor variables (Table 2), further multi-linear regression is conducted by selecting multicollinearity diagnostics.

Results

SPSS version-25 was used to find multicollinearity between the predictor variables (Study hours at home and Hours spent with friends).

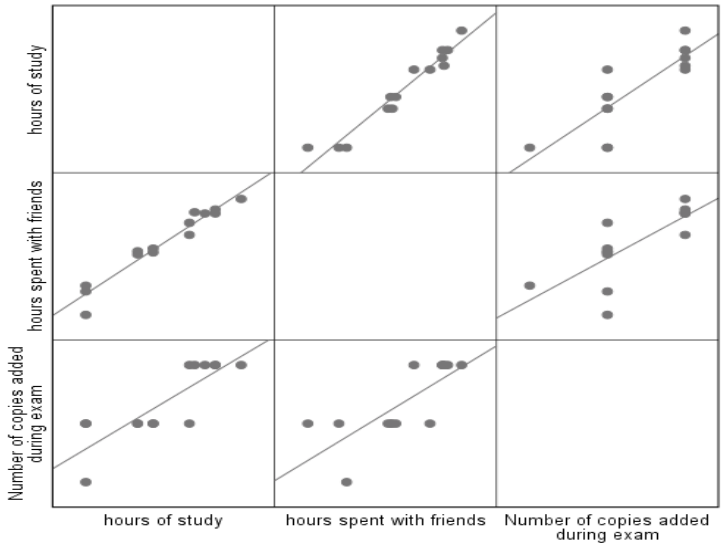
Table 2. Correlation between predictor variables

		hours spent with friends	hours of study	Number of copies added during exam
hours spent with friends	Pearson Correlation	1	.971**	.760**
	Sig. (2-tailed)		.000	.002
	N	14	14	14
hours of study	Pearson Correlation	.971**	1	.836**
	Sig. (2-tailed)	.000		.000
	N	14	14	14
Number of copies added during exam	Pearson Correlation	.760**	.836**	1
	Sig. (2-tailed)	.002	.000	
	N	14	14	14

**. Correlation is significant at the 0.01 level (2-tailed).

Source: SPSS data computation

Fig. 1 *Matrix Scatter Plot*



Pearson Correlation coefficient value (.971) between two predictor variables (hours of study and hours spent with friends) was found to be a strong positive correlation (also shown in Matrix scatter plot, Fig.1). If the correlation coefficient value between predictor variables is larger than 0.850, multicollinearity issue between such predictor variables is suspected. Whereas from table 1, the correlation coefficient value of NOCA is less than 0.850 with the other two variables, and hence multicollinearity is not suspected at least from this variable. Hence, as mentioned in the method section, multi-linear regression analysis is conducted selecting multicollinearity diagnostics from the statistics tab to explore the multicollinearity issue. After the computation of variables, the below tables are generated.

Table 3. Variable entered/removed

Model	Variables Entered	Variables Removed	Method
1	Number of copies added during exam, hours spent with friends, hours of study[b]	.	Enter
a. Dependent Variable: Marks obtained			
b. All requested variables entered.			

Table 4. Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.972[a]	.945	.929	3.84479

a. Predictors: (Constant), Number of copies added during exam, hours spent with friends, hours of study

Table 5. ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2547.033	3	849.011	57.434	.000[b]
	Residual	147.824	10	14.782		
	Total	2694.857	13			

a. Dependent Variable: Marks obtained

b. Predictors: (Constant), Number of copies added during exam, hours spent with friends, hours of study

Table 6: Coefficient

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	18.481	6.807		2.715	.022		
	hours of study	23.069	5.736	1.607	4.022	.002	.034	29.114
	hours spent with friends	-8.575	5.436	-.533	-1.577	.146	.048	20.844
	Number of copies added during exam	-3.389	3.329	-.149	-1.018	.333	.256	3.908

a. Dependent Variable: Marks obtained

Table 7: Collinearity Diagnostics

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions			
				(Constant)	hours of study	hours spent with friends	Number of copies added during exam
1	1	3.963	1.000	.00	.00	.00	.00
	2	.027	12.013	.58	.01	.01	.00
	3	.009	21.295	.11	.01	.05	.67
	4	.001	69.611	.31	.98	.95	.33

a. Dependent Variable: Marks obtained

Interpretation of the Result

Table 3 variable entered/removed shows that all requested variables have been entered. Marks Obtained entered in the dependent box whereas Study Hours, Hours spent with Friends and number of copies added during an exam in Independent box.

In Model summary, table 4, coefficient of determination (R^2 value 0.945) revealed that 94% of marks obtained by the students is determined by independent variables, study hours, hours spent with friends, and the number of copies added during the exam. The high coefficient of determination also indicates the level of suspicion in terms of multicollinearity between the predictor variables. ANOVA table somehow shows that F-statistics (85.351) is significant as the p-value (0.000) is less than ά-value (0.05) at 0.05 significance level which denotes that the regression model is fit.

However, in Coefficient table 6, collinearity statistics suggest that values of tolerance for study hours (.034) and Hours spent with Friends (.048) are less than 0.1, and VIF values for study hours (29.114) and Hours spent with Friends (20.844) are larger than 10. This confirms that there is a multicollinearity issue between two predictor variables in the model. For the third independent variable both the tolerance value greater than 0.1 and VIF value, less than 10 indicate that there is a multicollinearity issue from this variable.

In table 7, the condition index two values are greater than 15 and as variance proportions are high in study hours (.98) and hours spent with friends (.95), this helped the researcher to locate the position of the multicollinearity issue in the model. That means as these two predictor variables in variance proportions cross the threshold (.50), it confirms the multicollinearity problem in these two predictor variables. This also shows that both these variables are collinear to each other.

Conclusion and Discussion

The overall finding demonstrates that it is difficult to claim if Marks Obtained is statistically shaped by the study hours, number of Hours spent with Friends and number of copies added during exam variables. With such multicollinearity between two predictor variables, there will be an issue with modeling in regression analysis. This will lead the researcher in a blur state to confirm the relationship between independent variables and dependent variables.

After locating such a problem, it is better not to run regression analysis with the same data set. Either a new independent variable should be selected in the place of one existing predictor variable or both existing predictor variables (where multicollinearity is located) should be merged to avoid multicollinearity issues in further regression modeling. In the first option, eliminating one predictor variable and replacing this variable with new one, there is no certainty if the effect of the multicollinearity issue will be resolved after this action. If this continues, the same process must be repeated. Another solution to resolve or decrease the effect of the multicollinearity issue is increasing the sample size in existing variables.

Reiterating the above statement, the multicollinearity problem brings objectionable

effects on the estimation in the regression model so this issue is considered a serious problem. That means, with multicollinearity issues in the model, variable Y (dependent variable) cannot be predicted with existing independent variables where both or all predictor variables are highly correlated among themselves. Thus in multiple regression, identifying the best subset among different variables to include in a model is the hardest part of model building.

References

Alin, A. (2010). *Multicollinearity.* – WIREs Computational Statistics.

Ho. R. (2006). *Handbook of Univariate and Multivariate Data Analysis and Interpretation with SPSS*. Queensland University Rockhamption.

Neeleman. D. (1973). *Multicollinearity in linear economic models*. Springer.

Paul, R. K. (2006). Multicollinearity: Causes, effects and remedies. *IASRI, 1*(1), 58-65.

Raykov, T., & Marcoulides, G.A. (2006). *A First Course in Structural Equation Modeling* (2nd ed.). Lawrence Erlbaum Associates, Inc.