SXC JOURNAL
Volume 1, 2024

Regression Model in Social Science Research: The Issue of Multicollinearity, Detection Method, and Solution in SPSS

Dil Bahadur Gurung
Department of Major Arts, St. Xavier's College, Kathmandu, Nepal
Email: dilgurung@sxc.edu.np

ABSTRACT

One of the objectives of social science researchers in an inferential test is to build a reliableregression model. Multi-linear regression aims to find or predict the effect of predictor variables on predicted variables. However, when there is a high linear correlation between predictor variables in multi-linear regression, the predictor variables in the model cannot accurately define their impact on predicted variables. This statistical condition is called multicollinearity. Without testing and detecting multicollinearity and its precise treatment, the regression model can create difficulties in defining the impact of individual predictor variables on the predicted variable, leading to a faulty interpretation of the impact on the whole model. In this study, 36 primary-level teachers were selected randomly as the respondents. The respondents' data included their tentative salary, age, years of education, academic percentage in their final degree, and years of service. In the first round, the Karl Pearson correlation test is conducted among four independent variables: tentative current salary, age of respondent teachers, years of education completed, academic percentage in the final degree, and years of service s/he is involved in. SPSS version 25 is applied to find a correlation matrix between predictor variables, a matrix scatter plot, and linear regression with a collinearity diagnostic test. After finding a strong correlation between two variables, a collinearity diagnostic test is performed to locate and confirm the multicollinearity issue between the predictor variables. Once multicollinearity is confirmed, precise treatment is provided to solve the issue. The study found multicollinearity issues in two predictor variables; thus, further solutions were explained.

Keywords: Regression, multicollinearity, tolerance, VIF, condition index, variation proportion.

Introduction

When two or more independent variables in a regression model have a high degree of correlation with each other, statistically, it is known as a condition of multicollinearity. Regression analysis may be complicated as a result, as it becomes more challenging to precisely identify the contributions of each independent variable to the dependent variable. In other words, the statistical significance of independent variables cannot be achieved without confirming multicollinearity tests. This further reduces the predictive power of the regression model. Multicollinearity is one of the serious problems that should be resolved before starting the process of modeling the data (Daoud, 2017). In a nutshell, multicollinearity signifies a robust linear correlation between the predictor variables.

The regression model, for instance, is,

$$Y = a + b_1X_1^* + b_2X_2 + b_3X_3^*$$

Here, Y is a dependent variable, *a* is a constant (intercept), and b_1, b_2, and b_3 are the slopes of independent variables X_1, X2, and X3, respectively. The independent variables X_1 and X_3 (assigned with an asterisk) are highly correlated to each other. In such a situation, if multi-linear regression is conducted, changes in X_1 will also affect X_3 (the inflation effect), so there will be a problem observing the effect of each independent variable X_1 and X_3 on dependent variable Y. That means if there is a strong correlation between both the independent and predictor variables, then it would be difficult to find out which independent variable has a real impact on the dependent variable. Such a model can create a false interpretation in statistical inferential tests if multi-linear regression is not dissected with a collinearity diagnostic test.

Multicollinearity represents a high degree of linear inter-correlation between explanatory variables in a multiple regression model and leads to incorrect results of regression analyses. Diagnostic tools for multicollinearity include the variance inflation factor (VIF), condition index and condition number, and variance decomposition proportion (VDP) (Kim, 2019). In statistical terms, if both the independent variables are highly correlated, there is a high chance of multicollinearity.

Gujarati and Porter (2009) noted that a small sample size and their regression model can cause collinearity issues. Similarly, Neeleman (1973) states that if there is multicollinearity in the model, there is a possibility that the equation is under-identified, and consequently, it cannot estimate the impact of independent variables on dependent variables. Nonetheless, the possibility of the occurrence of multicollinearity is very rare in research if a proper sample size is picked. Raykov & Marcoulides (2006) write that the presence of multicollinearity in the model can easily lead to unstable regression coefficient estimates. So when multicollinearity 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 predictability of the model. Alin (2010) also adds that in such conditions, two or more independent variables are highly related.

Condition for detecting existence of multicollinearity in SPSS

Designing the best model in statistics is always the first priority of an individual researcher. By doing so, the predictability of independent variables can be calculated efficiently and accurately. There are basically two ways in SPSS to detect multicollinearity issues in the regression model. First off, after linear regression, in the ANOVA table, the F-statistics will be insignificant (P-value greater than 0.05). This tells us that there is some data-related fault in the whole regression model, and this can be because of multicollinearity issues in the model.

Second, after running correlation tests, if there is a strong correlation ($r > 0.8$) between two or more predictor variables, a multicollinearity issue is suspected in the model. This can lead the researcher to an uncertain situation in the result.

To further confirm the existence of multicollinearity issues in the model, a multi-regression analysis needs to be conducted. If the value of tolerance is below 0.1 and the variance inflation factor (VIF) value is greater than 10, then it can be confirmed that there is multicollinearity in the model. According to Paul (2006), practical experience indicates that if any of the VIF's exceed 10, it is an indication that the associated regression coefficients are poorly estimated because of multicollinearity. In other words, the variance inflation factor can estimate how much the variance of a regression coefficient is inflated due to multicollinearity. Variance inflation factors enable a rapid assessment of the degree to which a variable contributes to the regression's standard error. The variance inflation factor for the variables involved will be very substantial when there are significant multicollinearity difficulties.

Table 1. Tolerance and VIF interpretation

Tolerance		VIF	
Tolerance ≤ 0.01	Multicollinearity exist	VIF value ≥ 10	Multicollinearity exist
Tolerance ≥ 0.01	Multicollinearity do not exist	VIF value ≤10	Multicollinearity do not exist

Table 2. Descriptive statistics of test variables

ID	Tentative Salary (DV)	Age (IV1)	Service Year (IV2)	Education completed Year (IV3)	Academic percentage (IV4)
R1	35000	25	4	16	71
R2	40000	27	5	18	66
R3	41000	26	4	18	60
R4	25000	22	2	12	70
R5	30000	24	3	16	65
R6	50000	28	5	18	59
R7	35000	25	4	16	65
R8	25000	22	2	12	60
R9	45000	29	6	18	55
R10	40000	26	4	18	57
R11	33000	26	4	16	50
R12	41000	28	5	16	70
R13	25000	22	2	12	60
R14	25000	23	2	12	55
R15	30000	24	3	12	55
R16	30000	25	4	12	67
R17	35000	27	5	16	55
R18	45000	29	6	18	50
R19	30000	25	4	12	56
R20	35000	26	4	12	44
R21	45000	27	5	18	62
R22	45000	29	6	18	58
R23	51000	30	6	18	61
R24	42000	26	4	18	59
R25	35000	25	4	16	45
R26	33000	25	4	16	58
R27	30000	23	2	12	56
R28	25000	22	2	12	64
R29	22000	22	2	12	71
R30	27000	24	3	12	60
R31	40000	26	4	18	55
R32	34000	24	3	12	54
R33	40000	27	5	18	68
R34	30000	25	4	12	59
R35	30000	25	4	12	50
R36	31000	25	4	12	52

Source: Survey questionnaire: Teachers from primary level education, 2022.

Method

To find out the issue of multicollinearity and how it can be resolved, a teaching profession assessment file prepared during M.Phil. field work is used. From a total of 242 teachers' profiles, a simple random sampling technique is applied through SPSS, where approximately 15% of the sampling cases are selected. Through this technique, 36 teachers were selected. The respondent's other data, such as their current tentative salary (DV), their current age (IV1), their service year in that particular educational institution (IV2), their total education year completed (IV3), and the academic percentage (IV4) they received from their last degree, are used in this study.

Results

SPSS version-25 was used to find strong correlation between the predictor variables (Age of respondents, Years of service s/he is involved in, Years of education completed and academic percentage they received from their final degree).

The correlation coefficient value between Age (IV1) and Years of service (IV2) found strong correlation (0.975). However, the correlation between Age (IV1) and Education year (IV3); and correlation between Years of service (IV2) and Education year (IV3) found moderate correlation (0.772 and 0.743 respectively). The scatter plot matrix of predictor variables (Figure 1 and 2) also demonstrates that there is high correlation between Age of respondent and Years of service s/he is involved in.

Table 3. Correlation between predictor variables

		IV1	IV2	IV3	IV4
IV1	Pearson Correlation	1	.975**	.772**	-.162
	Sig. (2-tailed)		.000	.000	.345
IV2	Pearson Correlation		1	.743**	-.141
	Sig. (2-tailed)			.000	.413
IV3	Pearson Correlation			1	.069
	Sig. (2-tailed)				.688
IV4	Pearson Correlation				1
	Sig. (2-tailed)				

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

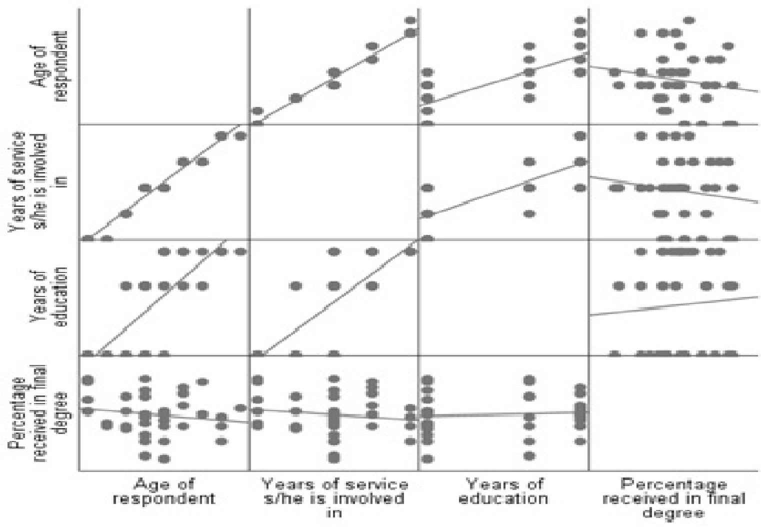

Figure 1. Scatter plot matrix of explanatory variables

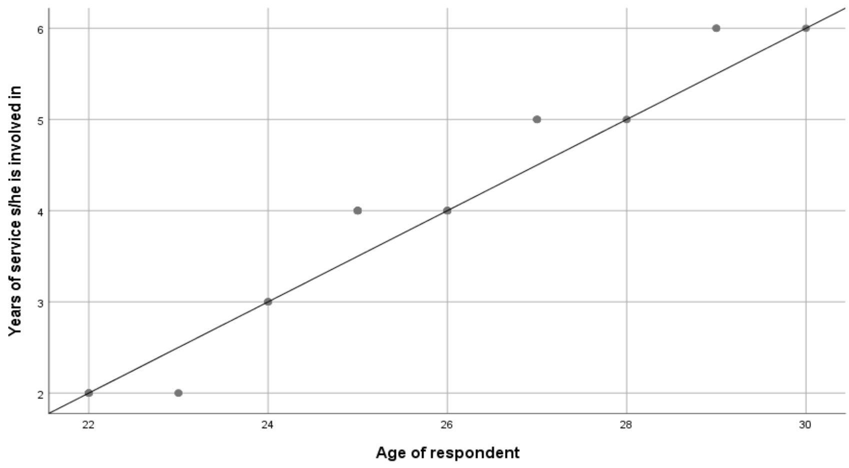

Figure 2. Scatter plot matrix: Correlation between age of respondent and years of service

If the correlation coefficient value between predictor variables is greater than 0.850, a multicollinearity issue between such predictor variables is suspected. Since there is a strong correlation between the age of the respondent and the years of service he or she is involved in, multicollinearity is suspected in the regression model. However, from Table 3, the correlation coefficient value between IV1 and IV3 (0.772) and IV2 and IV3 (0.743) signifies a moderate correlation, thus a multicollinearity issue between these variables is not suspected. Further, there is no correlation between IV1 and IV4 (-0.162), IV2 and IV4 (-0.141), and IV3 and IV4 (0.069), as their P-values are greater than the 0.01 significant level (0.345, 0.413, and 0.688, respectively). In these variables, there is zero chance of multicollinearity issues.

As there is a strong correlation between two independent or predictor variables (Table 3), further multi-linear regression is required to confirm multicollinearity issues in the model. To verify this, as mentioned in the method section, multi-linear regression analysis is conducted by selecting the collinearity diagnostics test option from the statistics tab to explore the multicollinearity issue in this regression model. After the computation of variables, the below tables are generated.

Table 4. Variable entered/removed

Model	Variables Entered	Variables Removed	Method
1	Percentage received in final degree, Years of education, Years of service s/he is involved in, Age of respondent	.	Enter
Note * Dependent Variable: Current salary of respondent **All requested variables entered.			

Table 5. Model summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.959*	.920	.910	2285.363
Note * Predictors: (constant), percentage received in final degree, years of education, years of service s/he is involved in, age of respondent				

Table 6. Analysis of variance

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	1872396147.718	4	468099036.929	89.625	.000**
	Residual	161909407.838	31	5222884.124		
	Total	2034305555.556	35			

Note * Dependent variable: Current salary of respondent

**Predictors: (constant), percentage received in final degree, years of education, years of service s/he is involved in, age of respondent

Table 7. Correlation coefficient

Model	Unstandardized Coefficients		t	Sig.	Collinearity Statistics	
	B	Std. Error			Tolerance	VIF
(Constant)	-69408.093	16855.129	-4.118	.000		
Age of respondent	4027.608	872.481	4.616	.000	.042	23.984
Years of service	-3186.158	1425.178	-2.236	.033	.048	20.827
Years of education	1014.535	236.030	4.298	.000	.361	2.769
Percentage in final degree	-11.909	59.784	-.199	.843	.870	1.150

Note: *Dependent Variable: Current salary of respondent

Table 8. Collinearity diagnostics

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions				
				Constant	Age of respondent	Years of service	Years of educa-tion	Percentage in final degree
1	1	4.914	1.000	.00	.00	.00	.00	.00
	2	.068	8.476	.00	.00	.04	.00	.04
	3	.010	21.876	.00	.00	.04	.82	.00
	4	.007	26.147	.02	.00	.05	.06	.88
	5	.000	164.700	.98	.99	.88	.12	.08

a. Dependent Variable: Current salary of respondent

Interpretation of the result

Table 4, "variable entered/removed," shows that all requested variables have been entered. The percentage received in the final degree, years of education, years of service s/he is involved in, and the respondent's age is entered in the independent box. The respondent's current salary is entered in the dependent box.

In the model summary table 5, the coefficient of determination (R2 value = 0.920) revealed that 92% of the respondent's current salary is determined by independent variables (percentage received in the final degree, years of education completed, years of service s/he is involved in, and age of the respondent).

The high coefficient of determination (Table 5) also indicates the level of suspicion regarding multicollinearity between the predictor variables. ANOVA table (Table 6) somehow shows that F-statistics (89.625) is significant as the p-value (0.000) is less than the ά-value (0.05) at the 0.05 significance level, which denotes that the regression model is fit.

However, in Coefficient (Table 7), collinearity statistics suggest that values of tolerance for age of respondent (.042) and years of service (.048) are less than 0.1, and their VIF (value inflation factor) are 23.984 and 20.827, respectively, which are both larger than 10. This confirms a multicollinearity issue between these two predictor variables in the model. For the third and fourth independent variables, both the tolerance values are greater than 0.1, and their VIF values are less than 10, indicating no multicollinearity issue with these variables.

In Table 7, the variance proportions are high in the age of the respondent (0.99) and years of service (0.88), which further confirms the existence of the multicollinearity issue in the model. That means these two predictor variables in variance proportions have crossed the threshold (0.50), confirming the multicollinearity problem in these two predictor variables.

Solution

Increasing the number of sample sizes can resolve multicollinearity issues in the regression model. By doing so, the tolerance test value will increase above 0.1. Further, the values in the variance inflation factor will decrease and remain below 10. Since no further samples are available, this study does not apply this method.

Another way to resolve such multicollinearity issues in a regression model is to remove the independent variable whose P-value is larger than the P-value of other independent variable (where collinearity issues exist). There is no certainty that this action will resolve the effect of the multicollinearity issue. For instance, in Table 7 above, the P-value of years of service (0.033) is larger than the P-value of the age of the respondent (0.000), so after removing the "years of service" variable, the further table is observed as below.

Table 9. Correlation coefficient after treatment

Model B	Unstandardized Coefficients		Standardized-Coefficients	t	Sig.	Collinearity Statistics	
	Std. Error	Beta				Tolerance	VIF
(Constant)	-35220.924	7518.480		-4.685	.000		
Age of respondent	2194.429	316.143	.624	6.941	.000	.357	2.799
Years of education	1069.196	248.994	.382	4.294	.000	.365	2.739
Percentage received in final degree	-26.317	63.039	-.024	-.417	.679	.880	1.136
Note: Dependent variable: Current salary of respondent							

In Table 9, both the tolerance values (0.357) and VIF (2.799) are greater than 0.1 and less than 10, respectively.

This denotes that the multicollinearity issue is resolved in this regression model. By doing so, the predictability of independent variables on dependent variables can now be accurately explained.

Conclusion

The overall finding (before treatment) demonstrates that it was difficult to explain the salary of respondents (a dependent variable) by other independent variables such as age, years of education, years of service in that educational institute, and percentage received in the final degree due to the collinearity issue in the model. This will lead the researcher into a blurred state to confirm the relationship between independent variables and dependent variables. After treating an independent variable (with a less significant value than another independent variable), the regression model is corrected, and the multicollinearity issue is resolved.

The multicollinearity problem is a major issue since it adversely impacts the regression model's estimation. In cases where both or all predictor variables have a high degree of correlation, the dependent variable (Y) cannot be predicted using the independent variables in the same model. Therefore, the most challenging aspect of developing a multiple regression model is determining which subset of the available variables to include in the model.

References

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

Daoud, J. I. (2017). Multicollinearity and regression analysis. *Journal of Physics: Conference Series.949.* https://iopscience.iop.org/article/10.1088/1742-6596/949/1/012009/pdf#:~:text=Multicollinearity%20is%20detected%20by%20examining,and%20it%20is%20in%20fact

Gujarati, D. N., & Porter, D. C. (2009). *Basic Econometrics* (5th ed.), 320-364. McGraw-Hill. https://ucanapplym.s3.ap-south-1.amazonaws.com/RGU/notifications/E_learning/0nline_study/Basic-Econometrics-5th-Ed-Gujarati-and-P.pdf

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

Kim, J. H. (2019). Multicollinearity and misleading statistical results. *Korean Journal of Anesthesiology, 72*(6), 558-569. https://doi.org/10.4097/kja.19087

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.