COMPARISON BETWEEN MULTILEVEL AND MULTIPLE REGRESSION ANALYSIS: A CASE STUDY OF CHARSADDA DISTRICT

Allah Yar[1], Abdur Rahman[2], Iftikhar Ud Din[3]

[1,3] Department of Mathematics, Statistics and Computer Science, The University of Agriculture, Peshawar, [2]Government College Mathra, Peshawar, PAKISTAN.

[1]allahyar302@yahoo.com

ABSTRACT

The purpose of this study is to observe the impact of different factors on student's academic performance and make selection, comparison between multilevel and multiple regression regarding and disregarding the actual hierarchal or nested structure of the data. To obtain these objectives primary data were collected through a well-defined questionnaire from 2132 students of 48 different public and private schools in Charsadda District, Khyberpukhtun Khwa. It is recorded that both in multilevel and multiple context the effect of age, gender, mother education, family type, family size, family monthly income school standard and school type are found significant on student's academic performance. The results of both multilevel and multiple regression were found quit similar. It is recommended that proper simulation study is needed to observe difference in multilevel and multiple regression, also more extensive similar studies covering province/country will bring good impact on the standard of education.

Keywords: Multiple regression, Multilevel Regression, Education, Charsadda

INTRODUCTION

In health, educational and social sciences the hierarchal influences are commonly observed. For instance, because of social inequity in their lives, drug abusers consume drugs i.e. they are persuaded by different factors related to social/communal level. Social, biological and psychological processes that influence health which can be further divided into different factors. Similarly, environmental and community stressors are the main causes of depression diseases. Furthermore, the impact of class/school level factors on the academic performance of students cannot be ignored. In the above mentioned examples, the common factor is the effect of group level traits on individual level characteristics. The hierarchal structure is not natural only but may arise as a consequence of specific research design. For instance, data collected through multistage sampling design or longitudinal design refers to cluster or hierarchal structure data. For all the aforementioned problems, appropriate exploration and particular analytical tools are essential. The precise assessment of such type of problems is provided by multilevel modeling techniques.

The main part of statistical modeling is Regression Analysis which is used by researchers to draw inferences from the data and a statistical model is the simplified form of a complex real world phenomena, Draper and Smith (1998). Multilevel models are basically the generalization of common regression analysis, based on multilevel formats, Cohen and Cohen (1983). These models can be utilized for different purposes, like reduction of data, prediction and causal inference by means of observational and experimental studies, Raudenbush and Bryk, (1986), Kreft and de Leeuw (1998), Snijders and Bosker (1999) and Hox (2002). Recently however, the development of numerous statistical packages for hierarchal structured data makes this sort of analysis more accessible. Some of them include,

ISSN: 2223-9944, e ISSN: 2223-9553

VARCL Longford (1987), Longford (1990), HLM Bryk et al. (1988), MLn Rasbash and Woodhouse (1995) and MLwiN Goldstein et al. (1998), the latest version of MLn.

According to Emmeke (2014) the dependency of observation in nested structured data cannot be ignored as it violates the assumption of independency of various conventional statistical methods such as the ordinary least square regression, student t-test and many others. Ignoring this yields, an increase up to 80 percent probability of type one error and decrease the statistical power. Jennifer and David (2001) who debated that, computing the standard errors by means of multilevel models were more accurate and reliable for nested structure data then by ignoring the structure and the proper method. Also noticed that standard errors calculated through single-level technique shows 20 percent or greater downwardly biased. The situations in which intra cluster correlations, which is the ratio of group level variation to total variation are high and variables related to upper levels are involved best suits multilevel regression. Furthermore Jennifer and David (2001) suggest that, if the intra cluster correlation is zero then the advantages of using multilevel regression appears less.

Multilevel models got popularity in recent decades and used in various fields of science due to its tremendous improvement both in applied and methodological sections. The interest in analyzing and interpreting multilevel data in education has historical background Bauer and Cai (2009). Robinson (1950) was the first to recognize the need for multilevel analysis through his study based on ecological process where a statistical discussions occurred in 1970. Lindely and Smith (1972) were the first to use the term hierarchical linear models. Extensive literature is available to study the application of multilevel models in education. Interested reader might see, Bryk et al. (1988), Willms (1992), Muthen (1994), Muthen and Satorra (1995), Goldstein (1997, 2003), Hox (2002), Heck and Thomas (2009) and Fan et al. (2011). In this study, the concept of Khan and Kamal (2013) were extended such that the variation can reduced and model can predicted more efficiently by maximizing the number of hierarchy levels. But variation may be reduced and model may predicted more efficiently by adding significant explanatory variables to each level of hierarchy. Also in this study comparison is made between the selection of multiple and multilevel regression by disregarding and regarding nested structure of data, when actually the data have natural hierarchal structure. Consequently, the current study was designed to determine the impact of age (measuring in years and rounds up), gender (male=0, female=1), mother education (educated=0, uneducated=1), father education (educated=0, uneducated=1), family size, family type (single=0, joint=1) and family monthly income in Pakistani rupees (up to 25000, up to 35000, up to 45000 and above 45000) at level one and class size, school type (public=0, private=1) and school standard (derived from overall students grades), categorized (good, normal and low) at level two on student's academic performance by disregarding and regarding hierarchal structure of the data. Application of proper statistical methods has significant importance in finding insight from collected data. This paper emphases on comparative assessment of multilevel and multiple regression techniques with reference to their application to the collected data from education sector.

MULTILEVEL MODEL

Let Y_{ij} symbolize the ij observation of dependent variable Y, observed at lowest level or level one of the hierarchy then the null or random intercept only multilevel model (two level) is,

$$Y_{ij} = \beta_{0j} + \varepsilon_{ij} \text{ (Level -1) } i = 1, 2, 3, \ldots, n \text{ and } j = 1, 2, 3, \ldots, k \quad (1.1)$$

$$\beta_{0j} = \gamma_{00} + u_j \text{ (Level -2)}$$

ISSN: 2223-9944, e ISSN: 2223-9553

This will estimate $k+2$, number of parameters i.e. one estimated intercept for each considered group and an estimates for the level one ε_{ij} and two u_j error terms. The intercept is taken as random and γ_{00} represents the average values of intercept and u_j shows the variations in intercepts between different clusters. Let X_{ij} and X_j, represent the explanatory variables (fixed) at level one and two simultaneously then the model (1.1) become,

$$Y_{ij} = \beta_{0j} + \beta_1 X_{ij} + \beta_2 X_j + \varepsilon_{ij} \quad \text{(Level -1)} \quad i = 1, 2, 3, \ldots, n \text{ and } j = 1, 2, 3, \ldots, k \qquad (1.2)$$

$$\beta_{0j} = \gamma_{00} + u_j \text{ (Level -2)}$$

Specialized software multilevel modeling MLWin, provide the facility of considering intercept or explanatory variable fixed or random, if we consider X_{ij} random then the model looks like,

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \beta_2 X_j + \varepsilon_{ij} \quad \text{(Level -1)} \quad i = 1, 2, 3, \ldots, n \text{ and } j = 1, 2, 3, \ldots, k \quad (1.3)$$

$$\beta_{0j} = \gamma_{00} + u_j \text{ (Level -2)}$$

$$\beta_{1j} = \gamma_{01} + u_{1j} \text{ (Level -2)}$$

Similarly the effect of β_{1j}, is varying around's the different groups, γ_{01} and u_{1j} represents the average effect of X_{ij} and averagely variation in the estimated values (β_{1j}) of each group.

METHODOLOGY

The metric (9^{th} and 10^{th}) students of Charsadda district of KPK province constitute the population. Primary data were selected through a well-designed questionnaire from 2132, secondary school certificate (SSC) level students of the district. The information were gathered by using two stage sampling procedure, consequently, the study is limited to two levels. In first stage of sampling procedure, 48 schools (clusters) is identified randomly from total 191 and in second stage the calculated number of students were chosen randomly. The impact of different factors related to students their socio-economic status and school were identified on student's academic performance by making use of the multilevel regression model. The gain in precision is assessed from the corresponding estimates from the multiple linear regression model. Furthermore, for the best exploration of the data is equivalent to adopting best and appropriate modeling approach. For the goodness of fit the criteria of deviance were used. The results and discussions section constitute selection between multilevel and multiple regression on the basis of the stated criteria, comparison of the estimated parameters and their standard error of multilevel model with corresponding multiple regression.

RESULTS AND DISCUSSIONS

The analysis started from a null model and the equivalent random intercept in the multilevel context. It is evident that significant changes are observed in the standard error, degrees of freedom and t-ratio. Though, the difference between the estimates is marginal but the standard error of the estimate is much higher in case of random intercept model as compared to null regression model. Consequently, the t-value reduces drastically along with the degrees of freedom. The remarkable reduction in the -2loglikelihood ratio indicates the superiority of multilevel over traditional regression model.

There is overwhelming indication of school effect and null hypothesis of no group effect is rejected. This established the preference of multilevel model over classical model.

ISSN: 2223-9944, e ISSN: 2223-9553

Table 1. Null Regression Model

Parameter	*Estimates*	*Std. Error*	*df*	*t-ratio*	*P-value*	*95% C.I.*	
Intercept	19.613	0.0335	2132	585.820	0.000	Lower	Upper
Residual	2.349	0.0733				19.57	19.70

*-2loglikelihood = 79122.189

Table 2. Multilevel Regression (Random Intercept Model)

Parameter		*Estimates*	*Std. Error*	*df*	*t-ratio*	*P-value*	*95% C.I.*	
Intercept		19.681	0.158	48	123.54	0.000	Lower	Upper
	σ_u^2	1.178	0.036500				19.36	20.00
Residual	σ_e^2	1.171	0.244502					

*-2loglikelihood = 6582.003

$$LRT = 7912.189 - 6582.003$$

$$LRT = 1330.186$$

$$\chi^2_{(1)} = 4.77 \text{ (p-value= 0.028)}$$

When the best set of predictors are included in the multilevel model depict the following picture. Also the concept of Khan and Kamal (2013) were extended for reduction of errors in multilevel regression from only increasing in the number of hierarchy by adding different significant variable at different level. As the result of multilevel in table 2 shows variation at level one 1.171 and at level two 1.178 with standard errors 0.244 and 0.036 respectively. However the multilevel model present in table 3 and 4 shows significant reduction in errors at both levels as 1.129 and 1.058 respectively at level one and two with standard errors 0.244 and 0.036 respectively. Also it is noted that, a significant variable from level two or school level will decreases variation of level two as well as level one but a significant variable from level one will only decrease variation at level one. This can be observed form table 3 and 4 also.

Table 3. Multilevel Regression (by Adding level one variable)

Parameter		*Estimates*	*Std. Error*	*t-ratio*	*P-value*
Intercept		24.288	0.493	49.26	0.000
Students age		-0.297	0.030	-9.9	0.000
	σ_{u0}^2	1.058	0.036		
Residual	σ_e^2	1.129	0.244		

*-2loglikelihood = 6488.007

ISSN: 2223-9944, e ISSN: 2223-9553

Table. 4 Multilevel Regression (by Adding level two variable)

Parameter		*Estimates*	*Std. Error*	*t-ratio*	*P-value*
Intercept		23.086	0.623	37.05	0.000
School type		-0.436	0.103	-4.23	0.000
	σ_{u0}^2	0.701	0.148		
Residual	σ_e^2	1.178	0.036		

*-2loglikelihood = 6347.056

Table 5 presents the results of multilevel and multiple regression. Different factors were considered related to student's level or level one and to school level or level two regarding multilevel context. The impact of these factor on the academic performance of students were present in table.5 regarding hierarchal or nested structure and disregarding the structure of the data i.e., in both multilevel and multiple regression format. We are slightly interested in the impact of these factors on student's academic performance, as main interest of the study is the selection and importance of appropriate statistical method. The difference in multilevel and multiple regression is observed using the results of estimated coefficient and their standard errors. First we will compare the estimated coefficients and then their standard errors. The estimated values of the impact of student's age on student's academic performance are recorded -0.228 and -0.239 in multilevel and multiple regression respectively. It is clear that both the calculated values have same sign (minus) with a difference of 0.011 points i.e., multiple regression shows higher coefficient value as compare to multilevel regression for the factor age. Similar results (-0.513 and -0.515) for the female (gender) factor are observed in multilevel and multiple regression respectively. In case of educated mother (mother education) factor, the estimated coefficient value in multilevel regression is higher (+33) than in multiple regression but has the same sign. This result is clearly opposite to the above two results i.e., the impact of age and female (gender). Similar indications were observed in the case of level two variables, i.e., class size and private (school type). In general, all estimated coefficients have the same sign in both multilevel and multiple regression format. A maximum difference of 0.20 is observed in the value of estimated coefficient which is too low to affect the significance or insignificance of a factor. Secondly if standard error of the coefficients is taken for comparison, the factors age and female (gender) show a higher values for standard errors in multilevel regression but the factors mother education and father education show a higher values for standard errors in multiple regression. A maximum difference of 0.172 is observed in the standard errors of multilevel and multiple regression. The results are quit surprising and similar for both type of regression and for data structure, drawing an exact conclusion about the appropriate method, is not possible. However if the selected factors will consider random in multilevel context the result will be surely different.

Furthermore, the impact of level one factors in multilevel context i.e. age, gender, mother education, family type, family size and family monthly income is observed. Negative and highly significant effect of student's age were observed on student's academic. As age increases the performance of students decreases and vice versa in both multilevel and multiple regression. The factor gender (coded female=1, male=0) shows negative effect on student's academic performance which means that, on the average female students obtained -0.513 and -0.515 lower marks in multilevel and multiple regression respectively as compare to male students. Similarly the effect of family type and family size is recorded negatively

ISSN: 2223-9944, e ISSN: 2223-9553

significant in both cases i.e. multilevel and multiple on student's academic performance. The impact of family monthly income on student's academic performance is recorded positive and significant.

At level two the effect of school standard, class size and school type were observed on student's academic performance. Results revealed that, school standard has positive and significant effect on student's academic performance in both multiple and multilevel format. As the good standard category is taken base for the remaining two low and normal standard categories and normal category shows lower value compare to low category which suggest that, the normal category is better than lower category. The effect of class size in both multilevel and multiple regression were observed negative and statistically insignificant on student's academic performance. Public schools are considered base for identifying the effect of school type on student's academic performance which shows that, on the average students from public schools secure more marks as compare to the performance of private schools.

Table 5. Comparison of Multilevel and Multiple Regressions

Variables	*Multilevel Regression*			*Multiple Regression*		
Model	*Coefficients*		*t, (p) values*	*Coefficients*		*t, (p) values*
	B	*Std. Er*		*B*	*Std. Er*	
(Constant)	24.604	0.603	40.80(0.00)	24.619	0.513	47.968(0.000)
Age	-0.228	0.029	-8.44(0.00)	-0.239	0.026	-9.231(0.000)
Gender (female)	-0.513	0.095	-5.40(0.00)	-0.515	0.061	-8.490(0.000)
Mother education	0.133	0.044	-3.02(0.002)	0.095	0.045	-2.124(0.016)
Joint system	-0.348	0.050	-6.96(0.00)	-0.363	0.052	-6.997(0.000)
Family size	-0.056	0.007	-8.00(0.00)	-0.06	0.007	-8.114(0.000)
Family monthly income						
Up to 35000	0.007	0.060	0.11(0.912)	0.020	0.062	0.330(0.742)
Up to 45000	0.157	0.083	1.89(0.028)	0.231	0.084	2.758(0.002)
Above 45000	0.791	0.133	5.94(0.00)	0.991	0.116	8.559(0.000)
School standard						
Normal	-0.945	0.131	-7.21(0.00)	-0.926	0.062	-14.948(0.000)
Low	-1.672	0.271	-6.16(0.00)	-1.621	0.099	-16.291(0.000)
Class size	-0.006	0.007	-0.85(0.395)	-0.004	0.003	-1.415(0.078)
Private	-0.197	0.066	-2.98(0.022)	-0.170	0.048	-3.562(0.000)
Variances						
Level one	0.876	0.027				
Level two	0.094	0.023				

*-2loglikelihood = 5437.103

ISSN: 2223-9944, e ISSN: 2223-9553

CONCLUSION AND RECOMMENDATIONS

In this study author oppose the importance and selection of appropriate and proper statistical method for the analysis of data and also the impact of different factors on student's academic performance. The result of multilevel model in this study is identical to multiple regression, however slight difference in estimates and their standard error are recorded. The recorded differences in estimates and their standard errors are not in the same order i.e. increasing and decreasing in all, some parameter estimates and standard errors shows increase and some shows decrease. The results will be different from multiple regression, if the factors included in this study consider random in multilevel format. However, the effect of age, gender, mother education, family type, family size and family monthly income at level one and at level two the effect of school standard and school type is found significant on student's academic performance. Furthermore, it is observed that response varies due to individual effect but the environmental effect also plays a significant role. On the basis of conclusion the following recommendations are made.

1. The study used a selected sample data, for obtaining more accurate results, a proper simulation study is needed.
2. To work out the economic problems of the families, it is recommended that government should provide maximum number of scholarships at SSC level.
3. It is further recommended that more extensive similar studies covering province/country will bring to focus the issues which are urgently needed to be solved to lift the standard of education in the country.

REFERENCES

[1] Aarts, E., Verhage, A., Jesse, V. V., Dolan, C. V., & Sluis, S. V. D. (2014). A solution to dependency: using multilevel analysis to accommodate nested data. *Nature neuroscience*, *17* (4), 491-496.

[2] Bauer, D. J., & Cai, L. (2009). Consequences of un-modeled nonlinear effects in multilevel models. *Journal of Educational and Behavioral Statistics, 34*, 97–114.

[3] Bryk, A.S., Raudenbush, S.W., Seltzer, M., & Congdon, R.T. (1988). *An introduction to HLM. Computer Program and User's Guide* (Version 2.0). USA: University of Chicago.

[4] Cohen, J., & Cohen, P. (1983). *Applied multiple regression analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

[5] Draper, N.R., & Smith, H. (1998). *Applied regression analysis* (3rd Ed.). USA: John Wiley.

[6] Fan, W., Williams, C. M., & Cokin, D. M. (2011). A multilevel analysis of student perceptions of school climate: The effect of social and academic risk factors. *Journal of Psychology in the Schools, 48*(6), 632-647.

[7] Galbraith, S., Daniel, J. A., & Vissel, B. (2010). A Study of clustered data and approaches to its analysis. *The Journal of Neuroscience, 30* (32), 10601–10608.

[8] Goldstein, H. (1997). Methods in school effectiveness research. *School Effectiveness and School Improvement*, *8*, 369-395.

[9] Goldstein, H. (2003). *Multilevel statistical models* (3rd Ed.). London: Edward Arnold.

ISSN: 2223-9944, e ISSN: 2223-9553

[10] Goldstein, H., Rasbash, J., Browne, W., Yang, M., Plewis, I. & Healy, M. (1998). *A user's guide to MLWin.* London: University of London.

[11] Heck, R. H., & Thomas, S. L. (2009). *An introduction to multilevel modeling techniques*. (2nd Ed.). New York: Routledge.

[12] Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

[13] Jennifer, L. K., & David, P. M. (2001). Multilevel modeling of individual and group level mediated effects. *Multivariate Behavioral Research, 36* (2), 249-277.

[14] Khan, R. A., & Kamal, S. (2013). Generalization of random intercept multilevel models. *Pakistan journal of statistics and operation research, 4*, 205-211.

[15] Kreft, I. G. G., & Leeuw, J. (1998). *Introducing multilevel modeling*. Newbury Park, CA: Sage.

[16] Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, 34*, 1–41.

[17] Longford, N. T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Journal Biometrika, 74*, 817–827.

[18] Longford, N. T. (1990). *VARCL*: *Software for variance component analysis of data with nested random effects* (Maximum Likelihood). Princeton, New Jersey: Educational Testing Service.

[19] Muthen, B. O. (1994). Multilevel covariance structure analysis. *Journal of Sociological Methods and Research, 22*, 376–398.

[20] Muthen, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Journal of Sociological Methodology, 5*, 267–316.

[21] Rasbash, J., & Woodhouse, G. (1995). *MLN command reference*. London: Institute of Education.

[22] Raudenbush, S. W., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Journal of Sociology of Education, 59*, 1–17.

[23] Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15*, 351–357.

[24] Snijders, T. A. B., & Bosker, R. (1999). *Multilevel analysis an introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

[25] Willms, J. D. (1992). *Monitoring school performance: A guide for educators*. Lewes: Falmer Press.

ISSN: 2223-9944, e ISSN: 2223-9553