CHAPTER – 5

Factors Explaining the Variations in the School Enrolment of the Students

In the preceding chapter we have made a descriptive analysis of enrolment of the students in the schools under study. In this analysis we have noted how it has varied from year to year and also across the schools. In some schools there has been more or less steady rise in enrolment over the period from 2013 to 2016, i.e., over the past four years. In some schools there have been variations over years and in some either no change or no fall. Now the question why the enrolment of students over time and across schools and in different locations has behaved like this – needs an answer.

In the following section we make an attempt to identify the factors responsible for the variation in enrolment across the schools. The other part of the question will be addressed in the last section of this chapter by considering panel data – (combination of time series and cross section data) that comprise 4 years data (2013-16) and 16 schools generating total number of 64 observation (16*4). The number of cross sectional units 16(16 schools) - number is significant as to generate results to be accepted at 5% level of significance for the independent variables under consideration. But the time series data for each school are only 4 covering the period of 4 years from 2013 to 2016. Only with four years data no time series analysis is suitably possible. So we take up panel data and carry out the statistical analysis.

5.1 The variables in the total enrolment of students in schools may be explained by a number of factors as we understand from the review of literature in this area, and also from my own experience as a school teacher where we constantly deal with the behavior of the students. We consider the following variables as the variables expected to significantly explain the variations in the enrollment of students in schools. These variables are : 1) academic qualifications of the parents of the students (X_2), 2) household average annual income (X_3), 3) school fees (X_4), 4) learning achievement score (X_5), 5) quality of teaching index (X_6) , 6) school infrastructure index (X_7), 7) female-teacher ratio (X_8), 8) location of the school (D_1), 9) school management or organization (D_2), 10) availability of UP (Upper Primary) and or Secondary or Higher secondary schools within 1 to 2 kms of the primary school (s) (D_3) where the students seek admission school, and 11) headmaster as a leader (X_9). X represent quantitative variables and D represent dummy or qualitative variables.

From the definitions of the variables, such, LAS, QTI, school Infrastructure index, and headmaster as a leader, it appears that they are correlated because many of the components or elements of these four variables are directly or indirectly related to one another. However, only guesses cannot be used as a basis for making such an assumption. So to know about their exact nature and extent of relationship among these variables the dependent variable - students' enrolment in schools is regressed upon all the following explanatory variables. The model can be written as

$Y = \beta_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\beta_5X_5+\beta_6X_6+\beta_7X_7+\beta_8X_8+\beta_9X_9+\gamma_1D_1+\gamma_2D_2+\gamma_3D_3+u$ **5.1**

The results of the estimated regression equations are given below (see Table 5.1). $X_2,X_3,\ldots\ldots\ldots\ldots\ldots,X_8$ are quantitative variables, $D_{1,}$ D_2 and D_3 are dummy or qualitative variables, u is the disturbance term.

The regression results are given below.

Table 5.1

Result of Regression of Y on 11 Explanatory variables:

Model: 1

	Constant and independent variables	β	t	p
1	Constant	59.658	.098	.926
2	Academic qualification of parents(X_2)	11.442	1.445	.244
3	Household average income(X_3)	-8.697E-5	-.774	.495
4	School Fees($X_{4)}$	-1.427	-1.809	.168
5	Learning achievement Score(X_5)	3.593	1.036	.376
6	Quality of teaching Index(X_6)	6.124	0.812	.476
7	School infrastructure index(X_7)	13.196	2.026	0.136
8	Location(D_1)	222.383	1.193	.319
9	Organization(D_2)	854.696	1.687	.190
10	Female Teacher(%)(X_8)	1.029	.201	.854
11	Availability of Up Pri/Sec school(D_3)	308.723	2.244	.111
12	Headmaster as a leader(X_9)	-10.112	-1.227	.307

R^2 = .907, Adjusted R^2 .566, F = 2.663, p=.228 DW=1.606 K=11

Dependent variable: Total students’ enrolment in school

From the regression results given in Table 5.1 we see that none of the regression coefficients are significant even at as high as 10% probability level. The correlation matrix also shows high correlation among some of the explanatory variables. The p- value of the estimated regression equation itself is so high as 0.228. This means that the model is valid at 23% level of significance. We, therefore, reject the model at 10 % probability level. The justification is, as already explained that in the case of qualitative variable such as education, level of significance at 10 % probability level is not too high a risk to land us in trouble. The level of significance indicates the level of accuracy of prediction i.e., the level of confidence. In the case of 10% level of significance, the confidence of level is 90%, which we think no harmful. This means that this model cannot be used to explain the variations in the dependent variable -the students' enrolment in the school. One reason of the model showing very bad fit to the data sets is, without any doubt, is very low degree of freedom which is just 4, here in this model.

The other reason might be the presence of multicollinearity among the explanatory variables. To counter these problems we have to drop some variables - the variables that are correlated with other variables. So considering the nature and degree of correlations among the explanatory variables and their levels of significance (not more than 10% probability level), along with the degree of their association with the dependent variable (explained in detail in the chapter 3) we include the variables such as (i) Academic qualification of the parents of the students (ii) location,(iii) female teacher ratio,(iv) Availability of upper primary(UP)/secondary school and (v) master as a leader in the new regression model. Multicollinearity among the explanatory variables was detected by VIFs. The regression results are given below.

Table 5.2

Regression Result using a set of only 5 Explanatory Variables

Model: 2

Constant and independent variables	β	t	p
1. Constant	-440.58	-.045	0.321
2. Academic qualification	1.478	.384	0.709
3. Location	74.468	.935	0.372
4. Female Teacher Ratio	3.333	2.087	0.06
5. Availability of Primary/Secondary school	167.611	2.017	0.071
6. Headmaster as a leader	3.289	1.228	0.24

R^2 = .711, Adjusted R^2 = .587, F = 4.294, p=0.015 K=5

Dependent variable: Students' enrolment

The regression result shows that in this model only two explanatory variables are found to have significant influence on the dependent variable, one at 6.3% and the other one at 9.1% probability levels. The model however, gives a good fit at 1.5% probability level. This indicates the presence of strong multicollinearity.

To improve the goodness of fit of the model and the level of confidence of making inferences about the influence of the explanatory variables we rerun the regression equation using two new sets of explanatory variables. While excluding or including explanatory variable(s) from the original set of 11 explanatory variables we considered two things – minimizing the degree of association among the variable (i.e. the problem of multicollinearity) and improving the degrees of freedom. Thus, we obtained two estimated regression models- 5.3 and 5.4 that are given below.

Table - 5.3

Regression Result using a set of only 4 Explanatory Variables

Model:3

Constant and independent variables	β	t	p
Constant	-507.029	-1.373	0.197
Location	73.242	0.258	0.358
Female Teacher Ratio	3.544	2.459	0.032
Availability of UP Pri/Sec school	160.616	2.063	0.064
Headmaster as a leader	3.785	1.686	0.121

R^2 = .707, Adjusted R^2 = .600, F = 5.632, p=0.006 K=5

Dependent variable: Students' enrolment

Table 5.4

Regression Result using a new set of only 5 Explanatory Variables

Model: 4

Constant and independent variables	β	t	p
Constant	-204.340	.128	.286
LAS	2.133	0.822	.431
QTI	-1.388	-.460	.655
School infrastructure	4.630	2.159	.056
Female Teacher Ratio	5.399	3.911	.003
Availability of UP Pri/Sec school	172.166	2.427	.036

R^2 = .762, Adjusted R^2 = .643, F = 6.400 p=0.006 K=5 DW = 1.278

Dependent variable: Students' enrolment

Comparing the results of the four Models presented in Tables 5.1 through 5.4 we see that the model 4 is the best model. A look at the following table demonstrates why it is the best.

Table 5.5:

A comparison of the Regression Models

Model	No. of significant parameters and their levels of significance (within parentheses)	R Square	Adjusted R Square	Significance of the Model	Comment
1	None (minimum level of significance 11.1%)	0.907	0.566	0.228	Not Significant
2	Only two (63% and 7.1%)	0.711	0.587	.015	Significant
3	Only two (8.2% and 6.4%)	0.707	0.600	.006	Significant
4	Three (5.6%, 0.3% and 3.6%)	0.762	0.643	.006	Significant

On the basis of the result as obtained from the above 4 models, we select the Model 4 as the best one and interpret its results in the following way:

The three variables that have significant influence on the students enrolment in schools are, namely, (i) school infrastructure (ii) Female teacher ratio(%) and (iii) availability of upper primary and secondary school within a distance of one to two k.m. The effects of the variables measured by the estimated values of the coefficient of this variable on the students' enrolment are respectively 4.630, 5.399 and 172.166 and their levels of statistical significance are 5.65%, 0.3% and 3.6% respectively.

The goodness of fit of the model is the best for model 4, Adjusted R_4^2=0.643, while Adjusted R_3^2=0.600 for the model 3, Adjusted R_2^2=0.567 for the model 2 and Adjusted R_1^2=0.566 for the model 1. Model 3 and 4 have the same level of significance. But comparing the levels of significance of the selected variables and the number of significant variables, we finally select model 4 as the best model identifying the explanatory variables for the dependent variable-students' enrolment in schools.

Thus, the variables/factors that are found to have highly significant effect on the enrolment of students in school are: 1) School infrastructure index 2) Female teacher ratio 3) Availability of Upper Primary/ Secondary School within distance of maximum 2 kms.

Henceforth, to attract the students to the schools, the school must 1) improve the school infrastructure (explained in the Methodology chapter-3), 2) increase the ratio of female teachers to the male teachers.

The third variable which, however, is beyond the control of the authority of the school under consideration is the proximity of the primary school under consideration to the upper primary/secondary school. The first two variables may be used as policy variables for school administration (local as well as state), while the third variable is beyond the control of the concerned authorities.

5.2. Now we make an attempt to identify the factors that influence the enrolment of students across schools and over time (2013-16). For this we use pooled data consisting of total 16 schools and 4 years (2013-16), resulting in the total number of observations N = 64. The pooled regression results are given below:

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-150.0564	42.58152	-3.523978	0.0009
LEARNING_ACHIEVEMENT	0.789419	0.476202	1.657741	0.1030
LOCATION	49.45625	14.21332	3.479571	0.0010
ORGANISATION	147.5943	35.58430	4.147736	0.0001
QUALIFICATION_GUARDIAN	2.446155	0.776388	3.150687	0.0026
SCHOOL_FEES	-0.315064	0.092142	-3.419321	0.0012
SCHOOL_INFRA	1.820608	0.563693	3.229784	0.0021
UP_PRIMA	47.19984	13.84765	3.408508	0.0012
R-squared	0.594797		Mean dependent var	45.16276
Adjusted R-squared	0.544147		S.D. dependent var	26.63512
S.E. of regression	17.98321		Sum squared resid	18110.17
F-statistic	11.74319		Durbin-Watson stat	1.250000
Prob(F-statistic)	0.000000			

The pooled data have been used to determine the variables that will explain the variation or the changes in the enrolment of the students over time and across schools .The regression result show that there is no over time variations in the students enrolment, but there is significant variables in the enrolment data across schools. The factors explaining this cross – section variations are identified by running the pooled regression equation

F(7, 56)	=	39.48
Prob > F	=	0
R-square	=	0.8315
Adj R-square	=	0.8105
Root MSE	=	26.113

enrolment	**Coef.**	**Std. Err.**	**t**	**P>t**
qualificationofguardians	2.33	0.428	5.44	0.00
school fees	-0.28	0.050	-5.63	0.00
schoolinfrastructuralindex	2.08	0.297	7.00	0.00
location	38.72	9.311	4.16	0.00
organisation	133.00	21.213	6.27	0.00
female teacher	0.42	0.220	1.91	0.06
availabilityofupss	46.43	7.468	6.22	0.00
_cons	-121.35	20.076	-6.04	0.00

So the results from multiple regression analysis and pooled regression model are as here under:-

In the multiple regression model, the variable-Learning achievement score is found to be significant at 10.3 % probability level (the value of the R^2 = 0.594 and adjusted R^2 = 0.544).

Now we get the pooled regression of Y (the students enrolment score) on the explanatory variables, namely, (i) academic qualification of the parents / guardians of the school students, (2) school fees, (3) School Infrastructure, (4) Location, (5) Organization, (6) female teacher, and (7) availability of upper primary /secondary or HS school within a distance of 1 to 2 k.m.s.

Interpretations of the results as obtained from the pooled regression model.

We present below the estimated pooled model for convenience of interpretation of the results.

Y(estimated value of enrolment) = -121.35+ 2.33 (Academic qualification of the parents) - 0.28 (School Fees) + 2.08 (School infrastructure) + 38.72 (Location) + 133.00 (Organization) +0.42 (female teacher) + + 40.82594 (Availability of upper primary / Madhyamik / HS school)

The estimated pooled regression model shows that the enrolment of students in the primary school depends on academic qualifications of the parents of the students. The higher the academic qualification of the parents, the enrolment of the students in the school increases. Locality, here urban, has very highly significant effect i.e., at 0.0% probability level. The variable – location (urban) is very significant at 0.0 % probability level. The parents prefer urban location of the school to the rural location. School infrastructure also plays a very important role. Because they believe that the schools located in urban areas enjoy some special advantages as compared to the schools in rural areas as well as good infrastructure facilities, sports ground, good library with books liked by the children below the age of 10/11 years. So they normally show greater tendency for urban schools and schools with good infrastructure for their children. Sex composition of teachers also is considered as a special and important factor in the Indian context as it is generally considered that the presence of female teachers in the schools helps increase attendance and retention student in the primary stage. This is particularly considered important in the rural areas where the attendance of girls is very poor.

The regression result shows that the enrolment of the students in urban areas increases by about 38 more over the schools located in the rural areas.

This means that the total number of students enrolled in urban schools is 38 more than the total number of students enrolled in rural areas schools for the period from class I to class IV.

Organization or the school management plays a very important role. Organization of a school refers to whether it is a public school (i.e. state-aided or state run or managed or run and managed by local government) or a private school, i.e. whether it is funded and managed by some individuals with spirit of improving the quality of education and looking for the overall development of the students. Regression results show that the private schools attract the students much more than the public school. The enrolment of the students in the private schools are higher than that other students in public school by a number of 133 students, and this increase is statistically significant at 0.0% probability level, i.e. significant at 100 percent confidence level.

School fees play a negative role. The higher is the school fee, lower is the tendency to get enrolled in the schools charging very high fees. The negative tendency as reflected in the estimate at the regression coefficient (-0.28) is statistically highly significant at 0.0% probability level, i.e. at 100 percent confidence level.

Here it may be noted that though school fees is a significant deferent to enrolment of students in the schools, mainly private school which charge higher school fees, the joint effect of all the variables (all variables other than school fees are positive and statistically highly significant (see the result above) taken together is positive and the net positive value showing net positive effect is very high.

The school infrastructure has positive and highly significant effect on the enrolment of the students. The positive effect of this variable means that the enrolment of the students relatively much higher in the school offering excellent or good infrastructure (as measured by or reflected in the higher infrastructural indices) than in the schools with low infrastructure facilities offered to the students. The level or significance of this effect is found to be 0.000 (p).

Availability of upper primary and / or Madhyamik or HS schools also plays a very significant role. The estimate of the regression coefficient is obtained as 46.43, meaning thereby that the primary schools which have in their close proximity (within one k.m.) upper primary and / or Madhyamik and/ or HS schools are very likely to get higher enrolment of the students than those which do not have this facility. The coefficient of this variable is statistically significant at 0.000 %(p =.000) probability level.

The pooled regression model's goodness of fit is given by R^2=0.8315, Adjusted R^2 = 0.8105. This means that as much as 83% of the variables in the enrolment of the students is explained jointly by these variables. All of them play very significant roles in explaining the variables in the enrolment of students.