STATISTICAL ANALYSIS OF THE DIFFERENT SOCIO-ECONOMIC FACTORS AFFECTING THE EDUCATION OF N-W.F.P (PAKISTAN)

Atta Ur RAHMAN[1]

PhD Candidate, Institute of mathematical methods in economics (EOS),
University of Technology Vienna, Austria

E-mail: rahman@eos.tuwien.ac.at; gujrati_stat@yahoo.com

Salah UDDIN[2]

PhD, University Professor, Department of Statistics,
University of Peshawar, Pakistan

E-mail: salahuddin_90@yahoo.com

Abstract: *A number of students in the urban and rural areas of N-W.F.P (Pakistan) and control group were collected to examine the various socio-economic factors which affect our education system. A logistic regression was applied to analyze the data and to select a parsimonious model. The response variable for the study is literate (illiterate) person(s) and the risk factors are Father literacy [FE], Father income[FI] Parents' attitude towards education[PA], Mother literacy [ME], Present examination system [PE], Present education curriculum [PC].*
The results of the analysis show that the factors Father Education combined with Parents' Attitude towards Education, Father Income combined with Mother Education, Father Income combined with Parents' Attitude towards Education are some of the factors which affect the education in N-W.F.P. Thus we concluded that there are a number of socioeconomic factors which affect our education.

Key words: *Logistic Regression; Stepwise Regression; Wald Test; Socio- economic Factors*

1. Introduction

Education is the basic need of human beings. It is also very important for the development of any country. Education is the responsibility of the state and government who should make every possible effort to provide it on an ever interesting and increasing scale in accordance with the national resources. The community should also realize its role in the development of education.

In Pakistan, 76% of the students in urban area are descendents of educated parents while in the case of the rural areas, this percentage is only 20%.

80% of the students in the urban area belong to high-income families while in the rural area 60% and 40% of the students belong to middle income and low-income families respectively, [8][3].

According to the 1998 census the adult literacy rate is 43.2%, this being a very low literacy rate. One of the main reasons for this low literacy rate, is the fact that only 1.1% of the GNP is invested in primary Education. That is why Pakistan is ranked 138 (out of 172) on the globe, [4]

The ultimate objective of development is to improve the living standards of people. In the present study, it has been tried to give a general structure of development particularly in Pakistan. In comparing the correlation of growth rates, GNP and literacy rates, Pakistan is found at the bottom of the world community. The physical quantity of life index (PQLI) is also computed, which is 40% in the case of Pakistan, being a very low value in comparison to other countries of the world,[1]

Pakistan is one of the countries of the world where the highest number of illiterates are concentrated. Being illiterate is not only an individual disability, it also has societal implications. Democratic institutions and values can hardly flourish in a society where half of the adult population is illiterate, and most of the voters cannot access information or read newspapers. The situation is particularly alarming for women and those living in rural areas. Illiteracy not only causes dependence, it deprives people of development of their fullest potential of participation in decision making at different levels, and ultimately rises to breed oppression and exploitation. Since its inception, governments in Pakistan have been endeavoring to eradicate illiteracy from the country. Although the overall literacy rate has increased gradually, the absolute number of illiterates has swelled significantly from 20.25 million in 1951 to 48.8 million in 1998, [3]

In Pakistan the strength of students in the class has little effect on the dropout rate. It is indicated that the dropout rate is higher in those schools where student/teacher ratio is lower. For instance in primary schools with strength of less than 70 students the students/teacher ratio is 28:1 and the dropout rate is 9%. On the other hand, in schools where the strength of students is around 1340 with a student/teacher ratio of 103:1, the dropout is amazingly less i.e. only one.

Looking at the opinions given by the teachers of GPS in Pakistan, in rural as well as urban areas, almost all the teachers agree that the main reasons for children dropping out from school at primary level are: limited opportunities of employment for educated youth and poverty i.e. boys from poor families have to help their fathers in farming and girls from poor families have to help their mothers in household activities. It is concluded that economic factors have a significant impact on children and they often drop out from schools due to poverty. It is also inferred that children often leave schools in early grades to become skilled workers.

It is concluded from the data collected that the average number of siblings in about 82% of the families of dropout children is four or more, whereas only 18% of the families have less than four children. Most of the fathers of dropout children are laborers, shopkeepers, helpers or attendants. Also a majority of them are either illiterate or have studied up to middle level only. In Pakistan poverty is major cause of dropout and thus students with a large number of sibling tend to dropout from school as the resources few,[5].

2. Methods and Materials

A sample of 500 students in the urban and rural area institutions of N.W-F.P (Pakistan) and control group was collected to examine the different socio-economic factors which affect our education system.

The response variable for the study is literate (illiterate) person(s).

Yi = 1, if literate
0, if illiterate

Risk factors selected for the study are

1. Father literacy [FE]
   FE= 1, if father is literate
   0, if father is illiterate
2. Father income[FI]
   FI= 1, if father income is more than Rs, 2000
   0, if father income is less than Rs, 2000
3. Parents' attitude towards education [PA]
   PA= 1, if parent's attitude is positive towards education.
   0, if parent's attitude is negative towards education.
4. Mother literacy[ME]
   ME= 1, if mother is literate
   0, if mother is illiterate
5. Present examination system[PE]
   PE= 1, if students like present examination system.
   0, if students not like present examination system.
6. Present curriculum of education[PC]
   PC= 1, if students like the present curriculum of education.
   0, if students not like the present curriculum of education.

The appropriate technique, used for model selection, in the case of binary response variable is logistic regression. Here we consider general linear models in which the outcome variables are measured on a binary scale.

The logistic regression model was first introduced by Berkson [2], who showed how the model could be fitted using iteratively reweighed least squares. Logistic regression is now widely used in social science research because many studies involve binary response variable. We look at basic notation underlying a logistic regression model.

The logistic model can now be written as

$$P = P(x) = \frac{e^{(\beta_0 + \sum \beta_i x_i)}}{1 + e^{(\beta_0 + \sum \beta_i x_i)}}, i = 1,2,...,m \qquad (1)$$

For a single explanatory variable X, the above model takes the form

$$P = \frac{e^{(\beta_0 + \sum \beta_1 x_1)}}{1 + e^{(\beta_0 + \sum \beta_1 x_1)}} \qquad (2)$$

We can draw inference from logistic regression. The main contribution in this case is that of [6], who provided general asymptotic results for maximum likelihood estimator, it follows that parameter estimator in the logistic models having large sample normal distribution. Thus a large sample 100(1-α) % confidence interval for parameter has the form

$$\hat{\beta} + z_{\alpha/2}\sigma(\hat{\beta}) \tag{3}$$

Where $\sigma(\hat{\beta})$ is the estimated asymptotic standard error.

Let $\gamma = (\gamma_1, \gamma_2, \gamma_3, \ldots\ldots\ldots\ldots\gamma_q)$ denote a subset of normal parameters. Suppose we want to test H_0. Let M_1 denote the fitted model, and M_2 denote the simpler model with $\gamma=0$. Large sample test can use [7], likelihood ratio approach, with statistic test based on twice the log of ratio of maximized likelihood's for M_1 and M_2. Let L_1 denote the maximized log likelihood for M_1 and let L_2 denote the maximized likelihood for M_2 under H_0, the statistic test $-2(L_2 - L_1)$ has a large sample Chi –squared distribution with df=q. Alternatively, by the large sample normality of parameters estimator, the statistic

$$\gamma'\left(Cov(\gamma)^{-1}\right)\gamma \tag{4}$$

has some limiting null distribution in large sample [6], This is called Wald statistic. When γ has a single element, this Chi square statistic with df=1 is the square of ratio of parameter estimate to its estimated standard error, that is

$$wald = \left[\frac{(estimate)}{s.e(estimate)}\right]^2 \tag{5}$$

In order to estimate the parameter we suppose that binomial data of the form y_i out of n_i trails $i = 1,2,3,....,n$ are available. Where the logistic transform of the corresponding success probability p_i, or $\log it(p_i)$ is to be modeled as a linear combination of n explanatory variable, $x_{1i}, x_{2i}, x_{3i}, ..., x_{ni}$. So that

$$\log it(P_i) = \beta_0 + \beta_{1i}x_{1i} + \beta_{2i}x_{2i} + \beta_{3i}x_{3i} + ... + \beta_{ni}x_{ni} \tag{6}$$

In order to fit a linear logistic model to a given set of data the (n+1) unknown parameters, $\beta_0, \beta_1, \beta_2, \beta_3, ..., \beta_n$ have first to be estimated. These parameters are estimated using the method of maximum likelihood .The likelihood function is given by

$$L(\beta) = \Pi \overset{n_i}{C}_{y_i} P_i^{y_i}(1-p_i)^{n_i - y_i} \tag{7}$$

This likelihood depends on the unknown successes probabilities p_i which in turn depends on the β's the likelihood function can regarded as a function of β. The problem now

is to obtain those values $\beta_{0,}\beta_{1,}\beta_{2,}\beta_{3,}.....,\beta_n$ which maximize L (β), or equivalently, $\log L(\beta)$

The logarithm of likelihood function is

$$\log L(\beta) = \sum_i \left[\log \overset{n_i}{\underset{y_i}{C}} + y_i \eta_i + n_i \log\left[1 + e^{\eta_i}\right] \right] \tag{8}$$

$\eta_i = p_i \big/ 1 - p_i = \sum \beta x_{ij}$ and $x_{0i} = 1$ for all values of i. the derivative of log likelihood function with respect to $n+1$ unknown β parameters are

$$\frac{\partial \log L(\beta)}{\partial \beta_j} = \sum y_i x_{ij} - \sum \eta_i x_{ji} e^{\eta_i (1+e^{\eta_i})}; j = 1,2,...,n \tag{9}$$

Once $\hat{\beta}$ has been obtained, the estimated value of the linear systematic component of the model is

$$\eta_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + ... + \hat{\beta}_n x_{ni} \tag{10}$$

this is unknown as the linear predictor. From this the fitted probabilities $\hat{p}_i$ can be found using

$$\hat{p}_i = e^{\eta_i} \big/ 1 + e^{\eta_i} \tag{11}$$

To fit an appropriate logistic model, which fits the data well, first we applied the forward methods to select the initial model for the backward elimination procedure. Logistic regression modeling is the appropriate statistical technique for this case because we wish to relate the chosen risk factors, which affect our education, which are binary variable. The objective of the study was to construct a model that could be used to predict the value of binary response variable. For fitting the models, we use the SPSS package. The variables used to determine an initial model are FE, MI, PA, ME, PE and PC. Different factors which were significant at different stages (at 5% level of significance) are FE is significant in one factor variable and the two factors which are significant $\mathrm{FE} * \mathrm{PA}, \mathrm{FI} * \mathrm{ME}, \mathrm{FI} * \mathrm{PE}, \mathrm{FI} * \mathrm{PA}$. While three factors FE*ME*PA, FI*PA*PE are significant, in four factors the interactions which are significant are FI*ME*MI*PA and FI*FE*PA*PE. For backward elimination we get model (FE, FE*PA, FI*ME, FI*PE, FI*PA, FE*ME*PA, FI*PA*PE, FE*ME*MI*PA, FI*FE*PA*PE).

Backward Elimination procedure

The best model is selected in one step automatically using SPSS. The model selected through SPSS is (FE*PA, FI*ME, FI*PA, FI*PA*PE). This model contains the main effect and two factor interaction. The following table gives information on log likelihood for the models selected in one step through backward elimination procedure.

Using SPSS we fit the model (FE*PA,FI*ME,FI*PA,FI*PA*PE) to estimate the model parameters and their standard error along with likelihood of the model. Also the odd ratios have been calculated and 95% confidence interval for odd ratio. The following table gives model summary.

Table 1. Variables in the equation

Step1		B	S.E	Wald	df	Sig.	Exp(B)	95% C.I For Exp(B)	
								Lower	Upper
	FE*PA	.243	0.325	14.671	1	.000	0.288	0.153	0.545
	FI*ME	0.797	0.321	6.158	1	.013	0.451	0.240	0.846
	FI*PA	0.734	0.368	3.982	1	.046	2.084	1.013	4.287
	FI*PA*PE	1.243	0.433	8.220	1	.004	3.465	1.482	8.103
	Constant	2.228	0.245	82.641	1	.000	9.280		

We see that the coefficients of the model parameters are highly significant. Hence the final selected model is

$$\text{Logit}(P) = 2.228 + 1.243\ \text{FI*PA*PE} + 0.734\ \text{FI*PA} + 0.797\ \text{FI*ME} + .243\ \text{FE*PA}$$

3. Conclusions

We have investigated the factors which affect our education in the model with one explanatory variable the main effect FE (father education) has a significant (p=.000) effect on education.

While the model having two factor variable FE*FI (father education and father income), variable FI*ME (father income and mother education), FI*PA (father income and parents attitude) has significant (p value=.000) on education.

The model having three factors, the factor FE*FI*ME (father education, father income and mother education), has significant effect on the education.

To obtain a best possible fitted logistic model, we use backward elimination procedures using SPSS package. We start form the model (FE, FE*PA, FI*ME, FI*PA, FE*ME*PA, FI*PA*PE, FE*ME*FI*PA, FI*FE*PA*PE), having four factors, three factors, two factors interaction and the main effect.

The best model is selected in one step automatically using SPSS. The model selected through SPSS is (FE*PA, FI*ME, FI*PA, FI*PA*PE). This model contains the main effect and three two factor interaction.

The factor which affects our education is "FE*PA", which means that the education of the child is depend on the education of the father and attitude of parents. The other factor are FI*ME means that father income and mother education also affect the education of the child. The father income and parents' attitude also affect the education of the child. The

three factors are father income, parents' attitude and present examination system also affect our education.

References

1. Bahrawar, J., Iftihar, D. and Saima, M. **A comparative study of socio-Economic measures for development,** J. Sc and Tech., University of Peshawar, 1998
2. Berkson, J. **Application of the logistic function to bioassay,** Journal of American Statistical Association, 39, 1994, pp. 357-365
3. Breines, I. **Literacy trends in Pakistan,** UNESCO Office, Islamabad, 2003
4. Wald, A. **Test of Statistical Hypothesis concerning several parameters when the number of observation is large,** Transaction of American Mathematical Society, 54, 1943, pp. 426-482
5. Wilks, S.S. **The large-sample distribution of likelihood ratio for testing composite hypothesis,** Ann. Math. Stat, 9, 1938, pp. 60-62
6. Yousaf, M. **A comparative study of performances of rural and urban Students in science subjects at Secondary level in Mardan,** Bed thesis, Allama Iqbal Open University Islamabad, 2002
7. * * * **Census Report,** Statistical Bureau of Pakistan, 1998
8. * * * **Technical report,** National commission for human development Pakistan, 2005

[1] He holds a M.SC and a M.Phil from University of Peshawar and currently he is pursing a PhD Program in Econometrics at Technical university of Vienna, Austria under the supervision of Prof Dr, Manfred Deistler. His Research topic is "Modeling and Forecasting of Financial Time series". Before starting his PhD Program he served as Statistics Lecture in Directorate of Colleges N-W.F.P (Pakistan).

[2] Dr.Salah Uddin has completed his M.Sc and Ph. D from United Kingdom and remained Chairman of the Department of Statistics, University of Peshawar from 1996 to 1999 , from 2001 to 2004 and from 2007 until now.

[3] Codification of references:

[1]	Bahrawar, J., Iftihar, D. and Saima, M. **A comparative study of socio-Economic measures for development,** J. Sc and Tech., University of Peshawar, 1998
[2]	Berkson, J. **Application of the logistic function to bioassay,** Journal of American Statistical Association, 39, 1994, pp. 357-365
[3]	Breines, I. **Literacy trends in Pakistan,** UNESCO Office, Islamabad, 2003
[4]	* * * **Census Report,** Statistical Bureau of Pakistan, 1998
[5]	* * * **Technical report,** National commission for human development Pakistan, 2005
[6]	Wald, A. **Test of Statistical Hypothesis concerning several parameters when the number of observation is large,** Transaction of American Mathematical Society, 54, 1943, pp. 426-482
[7]	Wilks, S.S. **The large-sample distribution of likelihood ratio for testing composite hypothesis,** Ann. Math. Stat, 9, 1938, pp. 60-62
[8]	Yousaf, M. **A comparative study of performances of rural and urban Students in science subjects at Secondary level in Mardan,** Bed thesis, Allama Iqbal Open University Islamabad, 2002