147

DETERMINANTS OF DROPPING OUT FROM SCHOOLS IN ETHIOPIA

BY
MESERET MAMO

A Thesis Presented to the School of Graduate Studies, Addis Ababa University in (Part) Fulfillment for the Degree of Master of Science in Statistics

Addis Ababa,
June, 1998

የአዲስ አበባ ዩኒቨርሲቲ
ቤተ መጻሕፍት
ADDIS ABABA UNIVERSITY
LIBRARIES

ACKNOWLEDGMENT

I wish to express my sincere thanks to my advisor, Dr. Abebe Tessera, for his many, constructive comments and suggestions throughout the study.

I would like also to thank Dr. Jacky Galpin for her valuable comments and suggestions.

Finally, I wish to express my heartfelt gratitude to all my friends and my family for their continuous support and encouragement.

Table of Contents

	Page
ACKNOWLEDGMENT	
ABSTRACT	
CHAPTER I	1
1. Introduction	1
1.1 Statement of the Problem	1
1.2 Objectives of the Study	4
1.3 Significance of the Study	5
CHAPTER II	6
2. Literature Review	6
CHAPTER III	16
3. THE DATA, THE VARIABLES AND METHODS OF STUDY	16
3.1 The Data	16
3.2 Limitations of the Data	17
3.3 Variables included in the Study	17
3.4 Methods of Study	21
3.4.1 Preliminary Analysis	21
3.4.2 Discriminant Analysis	21
3.4.3 Logistic Regression	21
3.4.4 Cox Proportional Hazards Model	22
CHAPTER IV	23
4. Analysis Results	23
4.1 Results of Preliminary Analysis	23
4.2 Discriminant Analysis	25

4.2.1 Assumptions 25

4.2.2 Classification Rules 26

4.2.3 Variable Selection 27

4.2.4 Hit-rate Estimation 28

4.2.5 Assessing Effectiveness of Classification Rules 28

4.2.6 Discriminant Analysis Results for Primary Schools 29

4.2.7 Discriminant Analysis Results for Secondary Schools 38

4.3 The Logistic Regression Model 44

4.3.1 The Model 44

4.3.2 Assessing the Goodness-of-fit of the Model 48

4.3.3 Logistic Regression Diagnostics 48

4.3.4 Logistic Regression Results for Primary Schools 49

4.3.5 Logistic Regression Results for Secondary Schools 54

4.4 The Cox Regression Model 58

4.4.1 The Model 63

4.4.2 The Final Proportional Hazards Model 65

4.4.3 Diagnostics 69

CHAPTER V 73

5. DISCUSSION AND CONCLUSION 73

REFERENCES 78

ANNEXES 84

Figure 1.....Normal probability plot of standardized residuals for primary schools 85

Figure 2.....Normal probability plot of deviance for primary schools 86

Figure 3.....Normal probability plot of standardized residuals for secondary schools 87

Figure 4.....Normal probability plot of deviance for secondary schools 88

Figure 5.....Plot of martingale residuals against age of a student 89

Figure 6.....Plot of martingale residuals against educational level of the household head 90

Figure 7..... Plot of martingale residuals against distance to school 91

Figure 8..... Plot of martingale residuals against family size 92

Figure 9..... Plot of martingale residuals against monthly income of the household head 93

Figure 10....Plot of martingale residuals against linear predictor scores (X'Beta) 94

ABSTRACT

Dropping out has been Ethiopia's problem since the inception of modern education in 1906. But little is known about the underlying causal factors that lead students to dropping out of school. In this study, the effects of external factors-factors that are not directly related to the school system-on dropping out were examined. The sample consisted of 7636 students in the age group 5-24 who were enrolled in primary and secondary schools of Ethiopia during the 1994/95 academic year. The effect of several factors on dropping out was assessed by using the logistic regression model, Cox's proportional hazards model and discriminant analysis. Using the discriminant analysis technique, we developed classification rules, for primary and secondary schools, by which group membership (stay-in & dropout) of students is predicted. About 93 percent of the students in primary schools and 90.02 percent of the students in secondary schools were correctly classified when we used linear classification rules. The percentages of correctly classified observations using the logistic regression model were quite similar to those in discriminant analysis. The results obtained using these results indicated that the major factors that contribute significantly to dropping out of primary schools are urban/rural residence, educational level of the household head, monthly income of the household head, marital status of parents of the household head, age of the student and relation to household head while factors that significantly affect dropping out in secondary schools include age of the student and family size. The Cox proportional hazards model was also fitted to examine the effects of factors on dropping out probabilities at various grade levels. Based on the findings,some conclusions and recommendations were made.

CHAPTER I

1. Introduction

1.1 Statement of the problem

The dropout problem is a common phenomenon in all countries of the world though the degree varies from country to country. A recent study conducted by Lockheed, et al. (1991) indicated that despite relatively high gross enrollment rates worldwide, including developing countries, fewer than 60% of those who enter school in low-income countries and only about 70% of those who enter school in the lower middle-income countries reach the terminal year of the primary cycle. Coombs (1985) also reported that developing countries suffer from high rate of dropouts.

Dropping out has a considerable effect on individuals and society at large. Individual dropouts suffer because many of them find it difficult to secure steady employment and an adequate income not just when they leave school but over their entire lifetimes. The situation may even be worse in the future because without a sound, basic education, dropouts will have less opportunities to learn new skills and adapt to a changing new environment. Society suffers as well because inadequate education which implies unemployment, and lost earnings lower tax revenues. Other social consequences include increased crime and poorer levels of health (Rumberger, 1987).

Dropping out or failure of the system to hold the children who are recruited at the beginning of each cycle is an indicator of educational wastage, that is, wastage in human learning, school buildings and equipment, and the labor of teachers (Brimer & Pauli, 1971). Mialaret (1979) and Perraton (1983) also explained about the consequences of high dropout rates as they result

in wastage of public resources. When educational resources are in limited supply which is not an uncommon situation in most developing countries, fewer places will be available for other children not yet provided for, whose educational potentiality and progress might be better than that of the repeaters and dropouts. Thus the cost of providing classrooms, equipment, materials & teachers for pupils who will either dropout of the school system before acquiring even functional literacy or will repeat several grades of schooling before completing a given cycle of education would be a serious hindrance to educational advancement..

Dropping out is not always disastrous as there are some young people who present difficulties in the school setting but will fit adequately into the world of work. It could even be beneficial to the schools in the sense that some of the dropouts would have been disruptive to those who want to remain in school and learn (Hathaway, et al., 1963; Rumberger, 1987). In their study of high school graduates and dropouts, for instance, Wehlage & Rutter (1986) found that dropouts showed equal or greater improvements in self-esteem and a sense of control than high school graduates.

Dropping out has been Ethiopia's problem since the inception of modern education in 1906 (Tadesse, 1975). However, there are no records on dropout rates for the period prior to the academic year 1987/88. Hence the rates from 1987/88 to 1994/95 were used to assess trends in dropout rates in grade 1. Grade 1 was selected because the dropout rate is very high in this grade as compared to other grades.

Table 1 Repeaters & dropouts in grade 1 by year & gender

		Year					
Rates	Gender	1987/88[1]	1988/89[1]	1991/92[2]	1992/93[3]	1993/94[3]	1994/95[3]
Repeaters	Boys	17.6	17.6	13.6	14.4	12.4	16.3
	Girls	20.1	20.1	14.6	17.3	16.2	20.4
	Total	19	19	14	16	13.7	17.8
Dropouts	Boys	28.5	30.8	35.6	25.1	35.7	28.6
	Girls	30.1	34.1	32.5	28.5	42.6	28.4
	Total	29	32	34	26.5	38.2	28.5

[1] Government schools only

[2] Based on samples of 1,243 schools

[3] Government and non-government schools

The above table shows that the dropout rate is higher than the repetition rate and both repetition and dropout rates are generally higher for girls than boys. But no systematic pattern, that is either increasing or decreasing, was observed.

In a school cohort of 1000 pupils in primary schools, the estimated average number of years dropouts spent in the system before dropping out is 2.0 years for boys while it is 2.4 years for girls (MoE, 1997).

The dropout problem is more serious in our country because many children are leaving the school system before they acquired basic literacy and numeracy. It is less likely that these dropouts will return to school because while they were in school, they were very young to fully understand the benefits of education.

Many researchers have identified factors that are closely associated with dropping out. Family-related factors related to dropping out include low educational and occupational attainment level of parents, low family income, speaking a language other than English in the home, single-parent families, and the absence of learning materials and opportunities in the home (Ekstrom, et al., 1986; Rumberger, 1983; Steinberg, Blinde & Chan, 1984; Tadesse, 1975). According to Ymen (1967), "transportation problems, lack of suitable materials, dissatisfaction with school programme, financial problem of parents, illness and academic difficulty" are the most important factors associated with dropping out of school.

A clear understanding of the effects of some of these and other variables on dropping out may help identify potential dropouts at an early age, when effective interventions could be designed and implemented. Thus, in this study we attempt to examine the effect of some of these variables on dropping out using some sophisticated statistical techniques.

1.2 Objectives of the study

The main objectives of the study are:

i) to examine the effects of some variables that may contribute to high rate of school dropouts in Ethiopia and identify subsets of variables which discriminate best between dropouts and stay-ins;

ii) to make regional and gender comparisons among school dropouts in Ethiopia;

iii) to make comparisons among the three statistical techniques used for analysis namely, discriminant analysis, logistic regression model and Cox's proportional hazards model in terms of their merits and demerits in explaining the data under consideration;

iv) to provide valuable information for educational planners and policy makers.

1.3 Significance of the study

Dropping out is a social problem which will remain in existence as long as there are schools. All our effort is to reduce its incidence (Tadesse, 1975). To design and implement effective prevention strategies, one has to understand the nature of the problem better. Thus, the aim of the study is to meet this need, that is, to provide valuable information, recommendations & suggestions for educational planners and policy makers so as to reduce and alleviate the problem.

It is hoped that these suggestions could help educational planners and policy makers to re-evaluate their approach towards the planning of education at the national level. It is also hoped that the study will stimulate researchers to further investigation of related problems.

CHAPTER II

2. Literature Review

2.1 General Background

A large body of empirical research has identified a wide range of factors that are associated with dropping out of school. There is a general agreement that dropping out is common among students who come from families with low income, low educational and occupational attainment levels. Other family related factors associated with dropping out include family size, marital status of parents, language spoken at home, and ethnicity (Brimer & Pauli, 1971; Coombs, 1985; Douglas, 1977; Ekstrom, et al., 1986; Hughe, 1973; Levy, 1971; Rosenbaum, 1986; Rumberger, 1983; Russell, 1968).

Numerous studies have found that personal characteristics such as age and sex of a child, and the distance a child has to travel to school are also important determinants of dropping out of school (Austin, 1975; Ekstrom, et al., 1986; Rosenbaum, 1986; Tadesse, 1975; World Education Report, 1995; Ymen, 1967).

It was also found that school-related factors such as academic failure, behavioral problems in school including absenteeism, truancy, tardiness and disciplinary problems are associated with dropping out (Edigar, 1987; Rosenbaum, 1986). That is, dropouts are more likely to have lower school grades, to have cut classes, to have had disciplinary problems or to have been suspended from school.

2.2 Evidence from other countries

2.2.1 Developed countries

Most of the literature from developed countries which display the causes of dropping out is reported from the United States of America.

Numerous studies reviewed by Rumberger (1987) pointed out that dropping out of school is common among young people from low socio-economic groups, no matter what factors are used to measure socio-economic status. Students from low income families tend to discontinue their education because their families cannot afford school fees or other expenses and because they want to or feel they have to work to help their families. Such students usually face shortages of books & stationery and they are denied proper attention and guidance at home. As a result their progress in school slows down and eventually they drop out (Mialaret, 1979).

Another important determinant of dropping out of school is parental educational level. Lichter, et al., (1962) states that students from families with low or no educational level may have limited aspirations on schooling and quite comfortably drop out. Uneducated parents are more likely to hold back their children from school than educated ones. It is not that parents, who have not been to school or despite having been in school remain illiterate, are antagonistic to their children's education but simply that they fail to understand in what way most of what the children do will be of benefit to them (Brimer & Pauli, 1971).

A study conducted by Elliott (1966) emphasizes on causes not directly connected with the school situation and points to the significance of the social class or the minority group to which

a student belongs. It points out that the pupil's personalities and the attitude of families to education interact in such a way as to encourage dropping out. It was also found that parents' occupation have influence on a child's educational pursuit. Edington (1970), for instance, observed a higher frequency of dropouts among high school rural students whose fathers were irregularly employed than in those families in which the father has steady employment.

Separation, divorce, or death of one or both parents might create crucial problems which can force the children to drop out of school (Brimer & Pauli, 1971; Hughe, 1973). Division of family might create tension, anxiety and instability in the life of the student. Thus it might be impossible for him/her to concentrate on school work (Edigar, 1987). Neill (1979) found that students with single parents are twice as likely to drop out of school as are students living with both parents.

Douglas, Ross, & Simpson (1968) observed that those who came from large families performed, on the average, poorly than those who came from medium or small sizes; the reasons being inadequate home conditions, lack of parental interest and poor schooling which are associated with large families. They also observed that even in the most prosperous homes where all the children would be well cared for and adequately fed, the performance of those who came from large families was poor.

In his study on dropping out of high school, Rosenbaum (1986) found that most dropouts were older in age. In another study, Austin (1975) reported a negative association between age and academic success.

The relationship between dropping out and language spoken at home was studied by Steinberg, Blinde & Chan (1984). They found that dropping out of school is common among students who speak a language other than English in the home.

A number of studies reviewed by Kablaoui & Pautler (1991) suggested that the dropout rate is higher among working students. However, the studies also indicated that employed students gained in practical business knowledge and they also developed social skills.

Behavioral problems in school including absenteeism, truancy, tardiness and discipline problems are also associated with dropping out (Edigar, 1987; Rosenbaum, 1986). According to Wehlage & Rutter (1986) , these problems are associated with low socio-economic status and other background factors. In relation to this, Crenbach (1962) reported that lack of motivation, student's laziness, poorness in reading and study habits, dislike of certain subjects and teachers contributed to school dropouts.

Some students may have trouble adjusting themselves to the school system. The performance of such students may not be satisfactory and hence they leave school. Lichter, et al., (1962) observed that " when school problems have reached the point where school is predominantly ungratifying and an unhappy experience, leaving school is an escape hatch, and the wish to drop out is very strong." On the other hand, those who can not adjust themselves to the school system may develop negative attitude towards teachers, and rules and regulations of the school. Consequently, they leave or pushed out of school.

Students who lack the mental ability to do the required work, and who may have the potential ability but lack the requisite skills to perform effectively leave schools early often before entering the senior high school. Frey (1970) categorized such students as retarded dropouts. They are easily identified by their low scores on achievements and intelligence tests.

According to Fine (1986), schools with poor facilities and inadequate teaching staffs have negative effect on students' performance in school and hence on their aspirations to learn. She reported also that many adolescents leave high school because their kin need assistance for health or social services.

In comparing dropouts with stayers Ekstrom, et al., (1986) found that dropouts: (1) are more likely to be older and to be males rather than females, (2) had fewer study aids present in their homes, (3) had less opportunity for non-school related learning, (4) were less likely to have both natural parents living at home, (5) had mothers with lower levels of formal education, (6) had mothers with lower educational expectations for their offspring, (7) had mothers who were more likely to be working, and (8) had parents who were less likely to be interested in or to monitor both in-school and out-of-school activities.

Developing a complex path-analysis model, they also found that the demographic variables related to the decision to stay in or drop out of high school were in approximate order of importance: intact family, socio-economic status, race/ ethnicity, region and sex. One school related variable, having behavioral problems, was also related to dropping out .

Wehlage & Rutter used two discriminating functions, academic function and social context of schooling function, to discriminate between dropouts and stay-ins. The academic function indicated that dropouts had low expectation for future schooling, low achievement/ability, low socio-economic status, and low grades. They observed also that the second function, the social context of schooling function, discriminate best between the two groups. This function indicated that dropouts differ from their academically similar peers in terms of their high truancy (the number of school days missed when not sick), discipline problems (being suspended or placed on probation and cutting classes every once in a while), lateness, and hours worked (number of hours a week the student works at a job).

2.2.2 Developing Countries

In his study of primary school dropouts in developing countries, Levy (1971) examined the relationship between dropout rates and independent variables closely associated with dropping out by means of stepwise linear regression. The results of the study indicated that high levels of dropout rates can be explained primarily by socio-economic factors. Higher average rate, higher degrees of tension, higher fertility rates and lower levels of urbanization were associated with high dropout rates. He concluded also that school systems with rigid examination and promotion procedures which lead to high repetition rates also deter children from finishing primary school.

Brimer & Pauli (1971) reviewed a number of studies conducted in Asia. The main cause for dropping out and non-attendance was asserted to be poverty. Other causes were employment of children in agriculture, lack of religious education in schools, migration, unsatisfactory methods of teaching, hostility among the landlords, lack of means of communication and school failure.

In several studies it has been suggested that inadequate teacher training, low salaries for teachers, large class sizes and high pupil-teacher ratios are strongly associated with high dropout rates. For instance, Halpern (1986) observed in his study of Latin America that a child's success at school can be hampered by overcrowded classes, unavailability of text books, and minimally trained teachers.

In a number of studies dropout rates have been observed to be higher in rural than in urban areas (Coombs, 1968; World Education Report, 1995). Edington (1970) suggested that socio-economic problems contribute much to this high dropout rate. It might also be because of the distance some rural students have to travel to school. Meerman (1980) and Nag (1977) observed how transport costs to distant schools create additional burden to parents in rural areas. Besides, home-school distances may compel students to look for shelter and food items around their schools. This additional cost may force parents to hold back their children.

According to the World Education Report (1995), in developing countries, dropout rates for girls are higher than those for boys. It is because girls are usually responsible for tasks ranging from the care of younger siblings to fetching water and collecting firewood. Failure to send girls to school have deep-rooted socio-cultural basis such as early marriage, child bearing responsibility, and unwillingness to allow girls to travel because of fear of abduction for marriage (UNESCO, 1975). Broadfoot & Shutherland (1987) reported that some girls are likely to have little education because their parents favor boy's education than their daughters. Some girls drop out of school because they may face psycho-social problems such as premarital sex, teenage pregnancy, teenage abortions and the like (UNESCO, 1988).

Fafunwa studied the effect of language spoken at home, in Nigeria, on a child's school performance. He observed that in most schools, where a child starts out with the mother tongue as medium of instruction in the first two or three years and then switches to English, many children dropout before the end of the primary school education.

2.3 Studies in Ethiopia

The study conducted by Tadesse (1975) addressed the dropout problem in primary schools of rural Ethiopia. It involved a sample of 4,000 dropout students in the 1968-69 academic year from both the rural and urban areas of Ethiopia. He found that for 4% of the dropouts in rural areas and 73% of the dropouts in urban areas, the main cause for dropping out was academic failure. He also found that the largest number of dropouts in rural areas occurred during the harvest months of December, January and April.

Contrary to the general literature, the results in this study indicated that parental death and broken homes were not among the major reasons mentioned by dropouts in rural areas of Ethiopia. The death of a father or a mother or both has no significant effect on the child's education because, in rural areas, immediate families (parents, grandparents, uncles, etc.) live in the same village and the upbringing of children is the responsibility of the whole group. Similarly, divorce was not a major reason because it is socially disapproved in rural areas.

In his study, performance in school measured in terms of staying in school, completing school and dropping out of school was dependent on the income level, educational status and religion of the parents, family size, and tribal background of the students at 0.10 level. He reported also that students who spoke 'Amharic' (the language of instruction in primary schools of Ethiopia at the

time) in the home were more likely to complete primary grades than those who spoke other languages at home.

High dropout rate among rural youth is observed perhaps due to daily walk of ten or more kilometers and because they are forced to spend more time on their parents' farm, especially during harvest time, than in school (MoE, 1969; Tadesse, 1975).

Disillusionment with prospects for finding employment after graduation may cause parents to refrain from sending their children to school, and students to be indifferent to schooling. Desta Asayehegn (1977) pointed out that " after observing the unemployed high school graduates who went before them, students who are currently enrolled in school have come to doubt the widely held myth that schooling in Ethiopia society will lead to future employment and social mobility." Since students' motivation is strongly influenced by their perception of future job opportunities, unavailability of employment even if they complete the intended level of education might discourage, and lead them to dropping out of school.

In a study involving six sample schools in Cheha district in Gurage zone, Yelfign, et al., (1995) observed higher dropout rates for girls than boys. A logit model was used to explain and generate probabilities of female participation at the six sampled primary schools. It was found that girls who have educated mothers, chances to repeat in school and parents willing to buy stationery and clothing have better chances to continue their schooling. In the same study, they observed that both female and male dropouts were of an older age.

A recent report for the PHRD Project Office (1996) revealed that percentage of males who drop out of primary and junior secondary schools in Ethiopia was higher than that of girls in rural areas. Beyond grade 8, the percentage was higher for girls. This is perhaps due to the higher opportunity cost of sending male students to school. On the other hand, the percentage of females discontinuing schooling was higher at all grades in urban centers.

The study points out also that larger families may derive less utility from sending an additional child to school because some children may be required to help their family while others are sent to schools. Hence, students from larger families are less likely to go to school.

Factors included in the study

Some of the factors mentioned above are directly related to the school system. These include poor academic performance, disciplinary problems, inadequate teacher training, low salaries for teachers, large class sizes, large pupil-teacher ratios and rigid examination and promotion procedures. Others are not directly related to educational performance. Brimer & Pauli (1971) called them external factors. In this study we attempt to examine the effect of these factors on dropping out of school. It is because

i) school- related factors have been studied by a number of researchers particularly because many of these factors are the ones that can be manipulated through practice and policy (Rumberger, 1987).

ii) most analysts and commentators have stressed that out-of-school factors in the family and community are more important than in-school factors, especially for developing countries (World Education Report, 1995).

CHAPTER III

3. The Data, the Variables and Methods of Study

3.1 The Data

The data for the study were obtained from the 1995/96 Household Income, Consumption and Expenditure and Welfare Monitoring Surveys carried out by the Central Statistical Authority. The surveys covered the population in the country on a sample basis except the non-sedentary population in Affar and Somali regions. Residents of collective quarters and homeless persons were not covered in the surveys.

The surveys, conducted in June and July 1995, covered 6743 households selected from 571 rural enumeration areas and 4731 households selected from 320 urban enumeration areas. In rural areas, a two stage sample design was used for sample selection, the enumeration areas (EAs) and households being the primary and ultimate sampling units, respectively. The number of sample EAs allocated to each region was reallocated to zones with probability to their sizes, where size is defined as the total number of households of the zones. Similarly, in the ten major urban centers such as Mekele, Bahir Dar, Gondar, Dessie, Jimma, Debre Zeit, Nazareth, Harar, Dire Dawa and Addis Ababa, a two stage sample design was used for sample selection with EAs and households as primary and ultimate sampling units, respectively. The sample EAs were reallocated to Keftegnas/Weredas in proportion to their sizes, size being the total number of households of the Keftegnas/Weredas. On the other hand, a three stage stratified sample design was used for sample selection in the other urban centers. Urban centers, EAs and households were the primary, secondary and ultimate sampling units, respectively. Sample selection of urban centers was carried out with probability proportional to size in their respective regions.

Fifteen households in each sampled urban EA, and 12 households from each sampled rural EA were selected using systematic sampling technique.

When complex survey designs are used and the sample is small, the estimates and test results in logistic regression model and Cox's proportional hazards model may be unreliable. However, data from large-scale surveys such as the data under consideration produce reliable results (Lehtonen & Pahkinen, 1995).

In the Household Income, Consumption and Expenditure survey, data were collected on characteristics such as age, sex, household size, and other demographic variables, consumption of food & drinks, household expenditure and payment on various items (consumption and non-consumption), and household income & receipts. On the other hand, data on educational and health status of households, housing facilities, household assets were collected in the Welfare Monitoring Survey.

The sample data includes 7636 students in the age group 5-24 who were enrolled in primary and secondary schools of Ethiopia during the 1994/95 academic year. Grade 12 students were not included in the study. Of those, 581 students (7.6 percent) dropped out of school for various reasons.

3.2 Limitations of the Data

There are two major limitations in the data used for analysis:

(i) A considerable number (about 82.7 percent) of the rural sampled households were selected from only three regions namely, Amhara, Oromiya and SNNPR. The sampled households

selected from the other regions accounted only for 17.3 percent. The observed small number of dropouts in the country (only 7.6 percent) can be attributed to this limitation of the data because , for students in regions like Affar, Somali, Benshangul/Gumuz, Gambella and Harari, the probability that their aspirations to learn can severely be affected by in-school as well as out-of-school factors is high and we expect the percentage of dropouts to be higher in these regions;

(ii) Important factors such as school type, educational level, working status and occupation of spouse were excluded from the analysis because there were many missing values associated with these variables. When these variables were included, the statistical procedures used for analysis rejected more than half of the observations. Even when they are excluded, about 21.7 % of the students in primary schools and 33.4 % of the students in secondary schools were rejected because of missing values.

3.3 Variables

An individual who registered in the academic year 1994/95 is taken as the unit of analysis. The criterion variable is a dummy coded 1 if the child did not return to school in the academic year 1995/96 and 0 otherwise.

The explanatory or independent variables included in the study can be classified into two categories:

i) **Personal characteristics** : includes variables such as age, sex, type of school, and relation to household head.

ii) **Family-related factors:** includes variables like educational level and occupation of the head of the household, of spouse, family size, marital status of the household head, estimated gross family income per month, distance to school, urban/rural residence and region.

Because of lack of reliability of some of the predictor variables like age, distance to school, household size and income, ordered categories are imposed on them for the purpose of classification. The variables with their corresponding categories are presented in the table below.

Table 2 Variable Description

Variable	Variable Label	Category
Age	Age of child	In completed years 1: 5-6 years 2: 7-17 years 3: 18 & above
Sex	Sex of child	1: female 0: male
Dst_schl	Distance to school	1: less than 1 km 2: 1-3 kms 3: greater than 4 kms
Relation	Relation to household head	1: Common child 0: Otherwise
Educhd	Educational level of household head	1: Illiterate 2: Grade 1-6 3: Grade 7-12 4: Above 12
Workhd	Working status of household head	1: Not working 0: Working
Occuphd	Occupation of household head	1: Farmer 0: Non-farmer
Educsp	Educational level of spouse	1: Illiterate 2: Grade 1-6 3: Grade 7-12 4: Above 12

Worksp	Working status of spouse	1: Not working 0: Working
Occsp	Occupation of spouse	1: Farmers 0: Non-farmers
Mar_st	Marital status of the household head	1: Non-married 0: Married
Income	Income of the family	In Birr per month 1: less than 170 2: 170 - 1,000 3: 1,000 or more
Fam_size	Family size	1: size of 1-2 2: size of 3-5 3: size of 6 & above
Urb_rur	Urban/rural residence	1: Rural 0: Urban
Region	Region	1: Tigray 2: Affar 3: Amhara 4: Oromiya 5: Somali 6: Benshangul/Gumuz 7: SNNPR 8: Gambella 9: Harrari 10: Addis Ababa 11: Dire Dawa

The above coding scheme is used only for preliminary analysis. The same coding scheme, except for variable age and region, will be used for other analyses.

3.4 Methods of Study

3.4.1 Preliminary Analysis

Chi-square tests were used to test the null hypothesis that the probability that a student falls in either of the two groups, dropouts and stay-ins, is not affected by the class to which the student happens to belong.

3.4.2 Discriminant Analysis

With two groups of students, a set of stay-ins and a set of dropouts, a predictive discriminant analysis technique is used to formulate a prediction rule by which group membership (stay-in or dropout) of a student is predicted. The SAS DISCRIM procedure was employed for this purpose.

The SAS STEPDISC procedure was also used to determine which variables contribute most to the separation of students who remain in school from those who do not.

3.4.3 Logistic Regression

A logistic regression model was also fitted to the data. This method has two advantages:

(i) the analysis and interpretation are quite similar to the well known procedure of multiple regression (Draper & Smith, 1981);

(ii) it requires far fewer assumptions than discriminant analysis; and even when the assumptions required for discriminant analysis are satisfied, logistic regression performs well (Norusis, 1990).

In logistic regression, we estimate the probability of dropping out of school on the basis of the predictive measures. If the estimated probability is less than 0.5, we predict that the student will remain in school; and we predict that the student will be a dropout if it is greater than 0.5.

The analysis was performed using the statistical package SPSS.

3.4.4 The Cox regression model

The proportional hazards model, which was first introduced by Cox (1972), is used to examine the effects of a variety of factors on dropping out probabilities at various grade levels. Cox regression models are similar to the well-known regression models in that we predict a dependent variable (length of time to the occurrence of the event) as a function of a set of independent variables. However, unlike ordinary regression models, Cox regression models can be used to analyze data that contain censored observations. Censored cases are those students who have not experienced dropping out by the time of the survey, but may in the future.

Unlike the logistic regression model, in which the dependent variable takes the value 1 if the event occurred and 0 if it is not, the Cox regression model takes into account the fact that the probability of experiencing an event differs with the duration of exposure to risk. The Cox regression model allows us, for instance, to treat two individuals who have identical covariates except that one was in school for 10 years and the other for only 1 year differently.

The statistical package SPSS is used for analysis.

CHAPTER IV

4. Analysis and Results

4.1 Results of Preliminary analysis

Of those students who were enrolled in primary schools during the academic year 1994/95, 6.33 percent left school for various reasons. Students who dropped out of junior and senior secondary schools accounted for 8.4 and 12.5 percent, respectively. A chi-square test(X^2-value = 50.25 with d.f=2) indicated a significant association between staying in or dropping out of school and the grade level(primary, junior & senior secondary levels) they were registered at $\alpha = 0.05$ level. Thus the analyses were performed separately for primary and secondary school dropouts.

Of the 15 variables, 11 of them are significantly associated with staying in or dropping out of school at 0.05 level of significance. These variables are age, distance to school, relation to head, educational level of head, occupation of head, educational level and occupation of spouse, monthly income of the household head, family size, urban/rural residence and region.

It is shown in the Table 3 that sex and working status of parents are not significantly associated with the educational status of students (stay-ins and dropouts).

Though educational level and occupation of spouse are significantly associated with staying in or dropping out of school, they are excluded from the analyses because there are many missing values for these variables; and inclusion of such variables, that is, variables with many missing values has two disadvantages: (i) estimates based on small samples are usually quite unstable,

and (ii) the resulting estimates may be too biased. Working status of the household head and the spouse were excluded from the analysis not only because they were not significantly associated with staying in or dropping out of school but also there were many missing values associated with these variables. Besides, the variable occuphd indirectly shows the household head's working status.

Table 3 Predictor variables and educational status of students

Variables	Variable label	d.f	X^2-value	P-value
Age	age of child	2	298.9947	.00000*
Sex	sex of child	1	0.00011	.99176
Relation	relation to head	1	37.002	.00000*
Dst_schl	distance to school	2	19.9992	.00005*
Educhd	educational level of head	3	87.003	.00000*
Workhd	working status of head	1	.02526	.87372
Occuphd	occupation of head	1	56.08	.00000*
Mar_st	marital status of the household head	1	2.69833	.10045
Incomehd	income of family	2	51.163	.00000*
Fam_size	family size	2	38.944	.00000*
Urb_rur	urban/rural residence	1	78.9909	.00000*
Region	region	10	63.9471	.00000*
Educsp	educational status of spouse	3	35.121	.00000*
Worksp	working status of spouse	1	.4623	.49656
Occusp	occupation of spouse	1	17.83	.00002*

* significant at $\alpha = 0.05$ level

4.2 Discriminant Analysis

Given the samples from the two populations (a set of dropouts and a set of stay-ins) of size n_g, g =1,2 with p predictor variables on each of the $n = n_1 + n_2$ units; the basic purpose of a predictive discriminant analysis (PDA) is to set up a rule, using the n x p data matrix, so as to determine from which of the two populations an $(n+1)^{st}$ unit is most likely to have been randomly sampled.

4.2.1 Assumptions

For the linear discriminant function to be "optimal"- that is, to provide a classification rule that minimizes the probability of misclassification- certain assumptions about the data must be met. Each group must be a sample from a multivariate normal population, and the population covariance matrices must be equal.

The assumption of normality is not tenable in our study since some of the predictor variables included are dichotomous (urb_rur, sex, occuphd, mar_st, relation, tigray, amhara, oromiya, affar, somali, ben_gum, SNNPR, gambella, harari, addis and dire) and some are ordered categorical variables (dst_schl, educhd, income, fam_size and age). In order to use the normal-based rules, the dichotomous variables were scaled using a 0-1 assignment, and ordered categories were coded using an integer scaling method. Though the results obtained using the linear discriminant function may not be "optimal ", due to lack of normality, the function performs reasonably well if the variables are handled in a way described above (Huberty, 1994).

The null hypothesis

$$Ho : \Sigma_1 = \Sigma_2 = \Sigma$$

may be tested using either a chi-squared or an F statistic. The SAS DISCRIM procedure employs a chi-squared statistic which is slightly different from the chi-squared statistic usually used to test the homogeneity of within covariance matrices. The expression for the test statistic, with its degrees of freedom, is given by Morrison (1976).

There are two problems associated with both the chi-squared and F-tests: (i) the degrees of freedom will be quite large even for medium-sized data sets, and (ii) the tests are sensitive to lack of multivariate normality of the outcome variable vectors. A level of 0.01 was used to improve the testing power of the procedures (Huberty, 1994).

4.2.2 Classification Rules

The maximum probability (or minimum distance) rule assigns a unit to that population for which the posterior probability of membership is greatest. This rule minimizes the total number of misclassification errors since it incorporates prior probabilities.

The maximum probability rule in terms of the classification functions is: Assign unit u to that population whose sample yields the largest linear classification function (LCF) score (with equal covariance matrices), or the largest quadratic classification function (QCF) score (with unequal covariance matrices).

4.2.3 Variable Selection

A SAS STEPDISC procedure was used to select a subset of variables that produce a good discrimination model using stepwise selection. The stepwise selection method begins with no variables in the model. The model is examined at each step. If the variable in the model that

contributes least to the discriminatory power of the model as measured by the Wilks' lambda fails to meet the criterion to stay (F-to-stay), then the variable is removed. Otherwise, the variable not in the model that contributes most to the discriminatory power of the model is entered. When all the variables in the model meet the criterion to stay and none of the other variables meets the criterion to enter (F-to-enter), the stepwise selection process stops.

Wilks' lambda is the likelihood ratio statistic for testing the hypothesis that the means of the classes on the selected variables are equal in the population. The mathematical formulas for Wilks' lambda and the associated F approximation are given in Rao (1973). Lambda is close to zero if the two groups are well separated.

A significance level of 0.15 was used for adding variables in the forward selection mode. This value lies within the interval $0.10 \leq \alpha \leq 0.25$ recommended by Costanza and Afifi(1979). The default value, a significance level of 0.15, was used for retaining variables in the backward elimination mode.

The partial R**2 (the square of correlation between the grouping variable and X_i when the other variables are partialed out of X_i) and F values (which reflects the group separation yielded by X_i in addition to that attributable to the other variables), common approaches for variable ordering in many analyses, pertain to group separation but not to classification accuracy. Hence another criterion based on the total group hit rates is used for variable ordering. A variable, say V_1, is considered " most important" if the analysis conducted with V_1 deleted yields the lowest hit rate.

4.2.4 Hit rate Estimation

A hit results when a unit emanating from group g, g = 1,2 is assigned, by means of the classification rule obtained, to group g. Let n_{gg} be the number of units in group g that are correctly classified. The hit rate for group g is n_{gg}/n_g and the total-group hit rate is $\Sigma\, n_{gg}/n$.

Huberty (1994) and Lachenbruch (1975) compared two types of hit-rate estimation methods: the holdout and the leave-one-out (L-O-O) methods. They recommended the use of the L-O-O method for hit-rate estimation because this method is fairly robust to distribution violations, and it yields an almost unbiased estimate of the expected actual hit rate unless n is extremely small.

The L-O-O method involves two steps: (i) one unit is deleted and LCFs are determined on the remaining n-1 units; (ii) the LCFs are used to classify the deleted unit into one of the two groups. The process is carried out n times and the proportion of deleted units correctly classified are used as hit-rate estimates.

4.2.5 Assessing Effectiveness of Classification Rules

Let q_g denote the estimated prior probability of membership in group g. The **Proportional Chance Criterion** for the total-group hit rate is given by

$$e/n = 1/n \,\Sigma\, q_g\, n_g, \quad g=1,2$$

where $e = \Sigma\, q_g\, n_g$ is the overall(or total-group) chance frequency of hits.

An observed total-group (or separate group) hit is compared with the corresponding proportional chance criterion to determine if we have better than chance classification. The standard normal

statistic may be used to test the hypothesis that the observed classification accuracy is better than what may be expected by chance:

$$Z = \frac{o - e}{\sqrt{e(n-e)/n}}$$

where $O = \Sigma n_{gg}$ is the overall observed frequency of hits.

The **Improvement over chance index** is defined as

$$I = \frac{H_o - H_e}{1 - H_e}$$

where H_o is the observed hit rate and H_e is the hit rate expected by chance. This index shows by how much we predict group membership better than chance. That is, by using the classification rule we developed, about I(100)% fewer classification errors would be made than those made if classification were done by chance.

4.2.6 Discriminant Analysis Results for Primary Schools

We begin our analysis by examining univariate statistics. This helps us to see group differences with respect to each variable. The table presented below shows the means for the predictor variables for students who dropped out and those who remained in school, with their corresponding standard deviations.

Table 5 Group means and standard deviation

Status of students

	Dropouts		Stay-ins	
Variable	Mean	Stan. dev.	Mean	Stan. dev.
Urb_rur	.49618	.50094	.25074	.43349
Sex	.41221	.49318	.46053	.49851
Dst_schl	1.0115	.13793	1.0005	.02313
Educhd	1.5000	.77682	1.9939	.94847
Occuphd	.51145	.50083	.26412	.44092
Income	1.4580	.58407	1.7819	.66321
Fam_size	2.6221	.53817	2.7118	.49207
Mar_st	.26718	.44333	.24619	.43085
Relation	.51527	.50072	.38159	.48584
Age	2.6947	.49353	2.4030	.52839
Tigray	.04580	.20945	.08777	.28300
Amhara	.16794	.37453	.18464	.38806
Oromiya	.30534	.46143	.28445	.45121
Affar	.01908	.13708	.00776	.08776
Somali	.00763	.08720	.00482	.06924
Ben_gum	.03435	.18248	.01258	.11145
SNNPR	.21374	.41073	.10918	.31190
Gambella	.03435	.18248	.02462	.15498
Harari	.04580	.20945	.03693	.18861
Addis	.07634	.26604	.19963	.39977
Dire	.04962	.21757	.04763	.21301

From the above table, we can see that 49.6 % of the dropouts came from rural areas, 41.2 % were female, 26.7 % came from non-married parents, and 51.5 % were not common child to parents. Dropouts came from lower-income earning families, and parents with lower educational level than stay-ins. They were also older in age.

Differences between the two groups can also be examined using Wilks' lambda and the F values from a One-way-ANOVA procedure. When variables are considered individually, the Wilks' lambda is the ratio of the within-groups sum of squares to the total sum of squares. Small values of lambda which correspond to large values of the F-statistic indicate that the group means do appear to be different.

The significance tests for the equality of group means for each variable revealed that 12 of the 21 variables (urb_rur, dst_schl, educhd, occuphd, incomehd, fam_size, relation, age, tigray, ben_gum, SNNPR, and addis) vary significantly between the two groups at $\alpha = 0.05$ level (Table 6).

Table 6 Tests for Equality of Group Means using Wilks' lambda & Univariate F-ratio with one and 3997 degrees of freedom

Variable	Wilks' Lambda	F	Significance
Urb_rur	.98115	76.8092	.0000*
Sex	.99942	2.3031	.1292
Dst_schl	.99583	16.7408	.0000*
Educhd	.98331	67.8342	.0000*
Occuphd	.98143	75.6086	.0000*
Income	.98539	59.2645	.0000*
Fam_size	.99800	8.0264	.0046*
Mar_st	.99986	.5788	.4468
Relation	.99540	18.4604	.0000*
Age	.98153	75.2221	.0000*
Tigray	.99861	5.5486	.0185*
Amhara	.99989	.4555	.4998
Oromiya	.99987	.5233	.4695
Affar	.99907	3.7258	.0536

Variable	Wilks' Lambda	F	Significance
Somali	.99990	.3902	.5322
Ben_gum	.99790	8.4206	.0037*
SNNPR	.99347	26.2568	.0000*
Gambella	.99976	.9418	.3319
Harari	.99987	.5338	.4651
Addis	.99399	24.1654	.0000*
Dire	.99999	.0212	.8842

* - Significant at 0.05 level.

The chi-squared statistic, employed by the SAS DISCRIM procedure, was used to test the equality of group covariance matrices. Since the Chi-square value (X^2-value = 6118.89 with 231 d.f) is significant at the 0.01 level, we reject the null hypothesis that the covariance matrices are equal. Under such a circumstance, the quadratic rule may be used for classification purposes. However, when the condition $\min_g (n_g) \geq 5P$, g = 1, 2 is met, the linear rule is better, because of its greater stability, than the quadratic rule (Huberty, 1994). In our case, the condition is satisfied for the two groups; hence a linear classification rule was used to classify units into one of the two groups (dropouts and stay-ins).

A number of values between 0.1 & 0.25, and between 0.80 & 0.93 were used as prior probability estimates for group 1 and group 0, respectively because (i) the sample proportion estimates underestimate the true prior probabilities, and (ii) dropout rates for primary and secondary schools are not available. Among these, estimated prior probabilities of 0.15 and .85 were chosen for group 1 (dropouts) and group 0 (stay-ins), respectively. These values were chosen because (i) they reflect the relative sizes of the respective populations, and (ii) they improve the total as well as the separate-group correct classifications.

The stepwise selection method begins only with the constant. The variable with the largest F statistic value among all variables (urb_rur) is selected first. Then this variable is tested for removal before selecting a new variable for entry. Since the variable urb_rur meets the criterion to stay, the variable not in the model that contributes most to the discriminatory power of the model (age) is entered. The selection process continues in such a way until no variables are eligible either for entry or removal. Twelve of the 21 variables were selected by the SAS STEPDISC procedure. These include urb_rur, age, relation, incomehd, dst_schl, educhd, mar_st, tigray, addis, amhara, gambella and oromiya. The Wilks' lambda and the associated F values for the variables selected by the stepwise method are summarized below (Table 7)

Table 7 Summary Table

Variables	Partial R**2	F-statistic	Wilks' Lambda	Prob < Lambda
Urb_rur	0.0190	78.068	0.9810	0.0001
Age	0.0114	46.467	0.9698	0.0001
Relation	0.0057	23.066	0.9643	0.0001
Incomehd	0.0042	16.798	0.9603	0.0001
Tigray	0.0037	14.907	0.9568	0.0001
Dst_schl	0.0028	11.233	0.9541	0.0001
Educhd	0.0025	10.105	0.9517	0.0001
Mar_st	0.0018	7.447	0.9499	0.0001
Addis	0.0015	5.862	0.9486	0.0001
Amhara	0.0009	3.504	0.9477	0.0001
Gambella	0.0007	2.770	0.9471	0.0001
Oromiya	0.0010	3.994	0.9461	0.0001

The predictor variables selected by the stepwise method differ from those obtained from the significance tests for the equality of group means. This is because the univariate tests do not incorporate interdependencies among the predictors.

The 12 variables selected by the stepwise method are ordered according to their relative importance in contributing to predictive accuracy (Table 8).

Table 8 Hit rates and Variable Ranks

Variable deleted	L-O-O total group hit rate	Rank
Tigray	0.9267	1
Dst_schl	0.9275	2
Educhd	0.9283	3
Amhara	0.9298	4
Addis	0.9300	5
Age	0.9304	7
Mar_st	0.9304	7
Oromiya	0.9304	7
Income	0.9310	9
Gambella	0.9316	10
Urb_rur	0.9321	11
Relation	0.9364	12

Statistically speaking, however, these variables except the last one are equally important in contributing to prediction accuracy. This is because variables with $(Z_{(i+1)}- Z_{(i)})$ values, where $Z_{(i)}$ is defined as

$$Z_{(i)} = \frac{nH_{(i)} - e}{\sqrt{e(n-e)/n}}$$

and $H_{(i)}$ denotes the observed (L-O-O) total group hit rate obtained with predictor X_i deleted, less than $.5\sqrt{2}$ are assigned the same rank.

The linear classification function, L_{ug}, u =1,..., n, g = 0,1, is a linear composite of the X scores with the row vector of weights

$$b_g' = \bar{X}_g' S^{-1}$$

and constant

$$c_g = -\frac{1}{2} \bar{X}_g' S^{-1} \bar{X}_g + \ln q_g$$

That is,

$$L_{ug} = b_g' X_u + c_g$$

The maximum probability rule is then to assign unit u to population 1 (dropouts) if $L_{u1} > L_{u0}$, otherwise assign u to population 0 (stay-ins). The weights b_g and the constant c_g for the variables selected by the stepwise method are presented in Table 9 .

Table 9 Linear Discriminant Function

	Group 0	1
Constant	-234.3591	-239.7489
Urb_rur	2.0469	2.7918
Dst_schl	428.0574	430.7277
Educhd	1.9656	1.7036
Income	5.4222	5.0675
Relation	-0.0992	0.4734
Age	8.5508	9.3391
Addis	6.2333	5.5071
Gambella	7.6235	6.6719
Oromiya	5.8757	5.4825
Amhara	6.9767	6.3982
Tigray	6.9812	5.8936
Mar_st	2.0575	1.7347

The Leave-One-Out (L-O-O) method or cross-validation was used for hit-rate estimation. The L-O-O results appear in the form of a classification table in Table 10 .

Table 10 Leave-One-Out Results

		Predicted Group 0	1	Total
Actual	0	4765	53	4818
Group	1	316	9	325
	Total	5081	62	5143

The overall chance frequency of hits is given by

$$e = \Sigma\, q_g\, n_g = 4144.05$$

Thus the chance hit rate or the hit rate expected by chance would be

$$H_e = e/n \quad = 0.806$$

This is the proportional chance criterion for the total-group hit rate. It is compared with the observed total-group hit rate, denoted by H_o, to determine if we have better-than-chance classification. The observed total-group hit rate is given by

$$H_o = o/n = 0.928$$

The observed total number of correct classifications is significantly (Z = 15.72) greater than the chance number at the 0.05 level. Then we computed the improvement over chance index to see by how much we predicted group membership better than chance. By using the linear classification rule we developed, about 63 % fewer classification errors would be made than if classification were done by chance.

Though we have an observed total group hit rate that is significantly greater than the chance number, the separate-group hit rate for the smaller group (dropouts) is not as good. It is because, for groups of unequal sizes that tend to reflect relative population sizes, use of unequal priors increase the hit rates for the larger groups and decrease the hit rates for the smaller groups.

When a quadratic rule is used for classification, about half of the dropouts are classified correctly. But the observed total number of classifications as well as the observed number of correct classifications for stay-ins are significantly (Z = -7.69 and Z = -13.01, respectively) less than the chance numbers. Moreover, the improvement over chance index for dropouts (I = 0.41) is not good.

4.2.7 Discriminant Analysis Results for Secondary schools

The differences between groups were examined using univariate statistics such as means of the predictor variables for the two groups, their corresponding standard deviations and F values from a One-Way ANOVA table. The group means and standard deviations of the 21 predictor variables for each group are summarized in Table 11 .

Table 11 Group means and standard deviation

	Status of students			
	Dropouts		Stay-ins	
Variable	Mean	Stan. dev.	Mean	Stan. dev.
Urb_rur	0.1418	0.3501	0.0635	0.2440
Sex	0.4894	0.5017	0.5314	0.4992
Dst_schl	1.0000	0.0000	1.0007	0.0269
Educhd	1.8794	0.9218	2.2519	0.9658
Occuphd	0.1560	0.3642	0.0859	0.2804
Income	1.7872	0.6189	1.9531	0.6551
Fam_size	2.6454	0.5089	2.7473	0.4879
Relation	0.5461	0.4997	0.4361	0.4961
Mar_st	0.2908	0.4557	0.2592	0.4384
Age	2.4184	0.5366	2.1906	0.4543
Tigray	0.0142	0.1187	0.0361	0.1866
Amhara	0.1206	0.3268	0.1827	0.3865
Oromiya	0.3333	0.4731	0.2737	0.4459
Affar	0.0213	0.1448	0.0036	0.0600

	Status of students			
	Dropouts		Stay-ins	
Variable	Mean	Stan. dev.	Mean	Stan. dev.
Somali	0.0071	0.0842	0.0029	0.0537
Ben_gum	0.0000	0.0000	0.0022	0.0465
SNNPR	0.0993	0.3001	0.0513	0.2206
Gambella	0.0000	0.0000	0.0065	0.0804
Harari	0.0567	0.2322	0.0585	0.2347
Addis	0.3192	0.4678	0.3227	0.4677
Dire	0.0284	0.1666	0.0599	0.2374

As shown in Table 11, dropouts from secondary schools are more likely to come from rural areas, parents with lower educational and income level, and non-married parents than stay-ins. They were also older in age than stay-ins. Dropouts from secondary schools in Oromiya, Addis Ababa and Amhara regions accounted for 33.3, 32.0 and 12.1 percent, respectively. It is interesting to note that the proportions of dropouts in the Benshangul /Gumuz and Gambella regions are zero.

The Wilks' lambda, the F values and their significance for the predictor variables under consideration are presented below.

Table 12 **Tests for Equality of Group Means using Wilks' lambda & Univariate F-ratio with one and 1524 degrees of freedom**

Variable	Wilks' Lambda	F	Significance
Urb_rur	0.9922	12.010	0.0005*
Sex	0.9994	0.9071	0.3410
Dst_schl	0.9999	0.1017	0.7498
Educhd	0.9876	19.198	0.0000*
Occuphd	0.9951	7.5276	0.0061*
Income	0.9946	8.2626	0.0041*
Fam_size	0.9964	5.5360	0.0188*
Relation	0.9959	6.2836	0.0123*
Mar_st	0.9996	0.6590	0.4170
Age	0.9800	31.053	0.0000*
Tigray	0.9988	1.8674	0.1720
Amhara	0.9978	3.3924	0.0657
Oromiya	0.9985	2.2660	0.1324
Affar	0.9949	7.6874	0.0056*
Somali	0.9996	0.6920	0.4056
Ben_gum	0.9998	0.3057	0.5804
SNNPR	0.9963	5.6253	0.0178*
Gambella	0.9994	0.9210	0.3374
Harari	1.0000	0.0071	0.9329
Addis	1.0000	0.0076	0.9307
Dire	0.9985	2.3713	0.1238

* - Significant at 0.05 level.

Nine of the 21 variables differ significantly for the two groups (dropouts and stay-ins) at 0.05 level. These include urban /rural residence, educational level, occupation and income of the household head, family size, relation to household head, age, and two regional variables: affar and SNNPR.

The null hypothesis that the covariance matrices are equal was rejected (X^2 = 8377.93 with 231 d.f) at the 0.01 level. However,since the condition that the number of observations in the smaller group should be greater than or equal to 5p is satisfied, we can use the linear discriminant function for classification.

For similar reasons described in section 4.2.6, estimated prior probabilities of 0.15 for dropouts and 0.85 for stay-ins were used.

Eleven of the 21 variables were selected by the stepwise method. These include age, educational level of the household head, affar, urban/rural residence, gambella, relation, marital status, tigray, amhara, dire and family size. Table13 shows the Wilks' lambda and the associated F values for these variables.

Table 13 Summary Table

Variable	F-statistic	Wilks' lambda	Prob < lambda
Age	31.870	0.9798	0.0001
Educhd	14.357	0.9707	0.0001
Affar	5.784	0.9671	0.0001
Urb_rur	3.738	0.9648	0.0001

Variable	F-statistic	Wilks' lambda	Prob < lambda
Gambella	3.247	0.9627	0.0001
Relation	3.401	0.9606	0.0001
Mar_st	3.486	0.9584	0.0001
Tigray	3.565	0.9562	0.0001
Amhara	3.988	0.9537	0.0001
Dire	3.845	0.9513	0.0001
Fam_size	3.027	0.9495	0.0001

The order of the variables with regard to their contribution to prediction accuracy differ according to the classification rule used. The L-O-O total-group hit rate when a variable is deleted ($H_{(i)}$) and rank for the variables are summarized in Table 14 .

Table 14 Hit Rates and Variable Ranks

Using LCF			Using QCF		
Variable deleted	$H_{(i)}$	Rank	Variable deleted	$H_{(i)}$	Rank
Mar_st	0.8926	1	Tigray	0.0998	1
Relation	0.8952	2	Mar_st	0.1071	2
Affar	0.8979	3	Educhd	0.1379	3
Educhd	0.9002	4	Age	0.1388	6
Gambella	0.9011	5.5	Affar	0.1388	6
Amhara	0.9011	5.5	Amhara	0.1388	6
Tigray	0.9015	7	Dire	0.1388	6
Dire	0.9020	8	Fam_size	0.1388	6
Age	0.9034	9.5	Urb_rur	0.1393	9
Urb_rur	0.9034	9.5	Relation	0.8649	10
Fam_size	0.9043	11	Gambella	0.8702	11

When a linear classification is used, the hit rates are fairly close. Thus it may be concluded that all the variables have equal importance in terms of their contribution to prediction accuracy. The same is true, except for the last two variables, when a quadratic classification rule is used.

The L-O-O results when a quadratic classification rule is used are presented in Table 15 .

Table 15 **Leave-One-Out Results with QCF**

		Predicted Group		
		0	1	Total
Actual	0	103	1893	1996
Group	1	5	203	208
	Total	108	2096	2204

Using the quadratic discriminant function for classifying units into one of the two groups has one advantage, the percent of correct classifications for dropouts is as high as 97.6. However it has two major disadvantages: (i) the observed total number of correct classifications is significantly (Z = -73.59) less than the total number expected by chance, and (ii) the observed number of correct classification for stay-ins is significantly (Z = -80.63) less than the chance number. Because of these, it is hard to say we have a statistically better-than-chance classification.

Cross-validation results using linear discriminant function are summarized below.

Table 16 Leave-One-Out Results with LCF

		Predicted Group		
		0	1	Total
Actual	0	1980	16	1996
Group	1	204	4	208
	Total	2184	20	2204

The observed total number of correct classifications is significantly (Z = 13.26) greater than the chance number. Moreover, the observed number of correct classifications for stay-ins is significantly (Z = 17.76) greater than that expected by chance. Thus we may conclude that we have a statistically better-than-chance classification. By using the linear classification rule, about 54 percent fewer classification errors would be made than if classification were done by chance. The separate-group hit rate for dropouts is , however, not good. This is mainly the effect of using unequal prior probabilities.

4.3 The Logistic regression model

4.3.1 The Model

The outcome or dependent variable Y is a dummy coded 1 if a child dropped out of school before completing the compulsory cycle and 0 otherwise. Let the conditional probability that dropping out has occurred given **X** be denoted by $P(Y=1/X) = \pi(\mathbf{X})$ where $\mathbf{X}=(X_1, X_2,...,X_p)'$.

The logistic regression model is then

$$\pi(X) = \frac{e^{g(X)}}{1+e^{g(X)}}$$

or equivalently,

$$\pi(X) = \frac{1}{1 + e^{-g(X)}}$$

where $g(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$

The logit, $g(X) = \ln\left(\frac{\pi(X)}{1-\pi(X)}\right)$, is linear in its parameters, may be continuous, and may range from $-\infty$ to ∞, depending on the range space of **X**.

The value of the outcome variable can be expressed as

$$Y = \pi(X) + \epsilon$$

The error term ϵ may assume only two values. If Y=1 then $\epsilon = 1 - \pi(X)$ with probability $\pi(X)$, and if Y=0 then $\epsilon = -\pi(X)$ with probability $1 - \pi(X)$. Thus ϵ has a distribution with mean zero and variance equal to $\pi(X)[1 - \pi(X)]$. That is, the binomial, not the normal, distribution describes the distribution of errors.

For variables with more than two categories such as age, educhd, dst_schl, fam_size and incomehd, we created new variables to represent the categories. The number of variables required to represent a categorical variable is one less than the number of categories. The deviation coding scheme was used. It differs from indicator-variable (dummy variable) coding scheme only in that the reference category is coded as -1 for each of the new variables. For each new variable, the coefficients tell us how much better or worse each category is compared to the overall effect of all the categories. Interaction terms created for categorical variables will be of interest if this coding scheme is used instead of indicator- variable coding scheme. Interaction terms included in the analysis are sex x urb_rur, dst_schl x urb_rur, tigray x urb_rur

amhara x urb_rur, oromiya x urb_rur, somali x urb_rur, ben_gum x urb_rur, SNNPR x urb_rur gambella x urb_rur, harrari x urb_rur, addis x urb_rur, and dire x urb_rur.

In logistic regression the estimates for betas, or regression coefficients are obtained by the method of maximum likelihood. That is, the coefficients that make the observed results most likely are selected. Since the model is non linear, an iterative algorithm is necessary for parameter estimation.

The odds of an event are defined as the ratio of the probability that it will occur to the probability that it will not. The logistic model in terms of the log of the odds is

$$\log\left(\frac{\pi(X)}{1-\pi(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

Thus, the β_i 's can be interpreted as the change in log odds associated with a unit change in the independent variables. The logistic model can also be written in terms of odds as

$$\frac{\pi(X)}{1-\pi(X)} = e^{\beta 0} \cdot e^{\beta 1 X 1} \ldots\ldots e^{\beta p X p}$$

Thus $e^{\beta i}$, i = 1,2,..., p is the factor by which the odds change when the i^{th} independent variable increases by one unit. If β_i is positive, the odds are increased; if β_i is negative, the odds are decreased; and if $\beta_i = 0$ then the odds remain unchanged.

The FORWARD STEPWISE SELECTION method is used to select those variables which best explain the data. It starts with a model that contains only the constant. At each step, the variable

with the smallest significance level for the **score statistic**, provided it is less than the chosen value, 0.05, is entered into the model.

The score statistic, in the univariate case, is

$$ST = \frac{\sum_i x_i (y_i - \bar{y})}{\sqrt{\bar{y}(1-\bar{y})\sum_i (x_i - \bar{x})^2}}$$

All variables that have been entered are then examined to see if they meet removal criteria. The process continues until either a previously generated model is encountered or no variables meet entry or removal criteria.

The likelihood-ratio (LR) test is used as a removal criterion. This criterion was chosen because it performs better than the Wald statistic which is defined as the square of the ratio of the coefficient to its standard error for a variable with a single degrees of freedom (Hauck & Donner, 1977; Jennings, 1986).

The likelihood is defined as the probability of the observed results, given the parameter estimates. Under the null hypothesis that the coefficients of the terms removed are 0, the LR is defined as

$$LR = \frac{\text{(likelihood for the reduced model)}}{\text{(likelihood for the full model)}}$$

If the null hypothesis is true and the sample size is sufficiently large, -2logLR has a chi-square distribution with r degrees of freedom, where r is the difference between the number of terms in the full model and the reduced model.

4.3.2 Assessing the goodness of fit of the model

A classification table is used to assess how well the model fits by comparing predicted values to the observed outcomes. It consists of the number of cases correctly classified, and misclassified cases. It also gives the overall percent of correctly classified cases.

Another measure of how well the model fits is -2 times the log of the likelihood(-2LL) statistic. A good model is one that results in a high likelihood of the observed results, or equivalently, the one that results in a small value for -2LL.

There is also another measure: model chi-square. The model chi-square is the difference between -2LL for the model with only a constant and -2LL for the current model. It tests the null hypothesis that the coefficients for all the terms in the current model, except the constant, are zero. The degrees of freedom for the model chi-square are the difference between the number of parameters in the two models.

4.3.3 Logistic regression diagnostics

The residual is the difference between the observed probability of an event and predicted probability of an event based on the model.

The standardized residual is defined as

$$Z_i = \frac{\text{Re}\,sidual_i}{\sqrt{p_i(1-p_i)}}$$

where p_i denoted the i^{th} predicted probability.

If the sample size is large, the Z_i's are approximately normally distributed with mean zero and standard deviation 1.

The deviance is calculated as the square root of

-2 x log (predicted probability for the observed group).

It has a negative sign if the event did not occur for that case. Large values for deviance indicate that the model does not fit the case well. The deviance is approximately normally distributed for large sample sizes.

4.3.4 Logistic Regression Results for Primary Schools

The new variables created to represent categorical variables such as educhd, age, incomehd, fam_size and dst_schl are presented in Table 17.

Table 17 Deviation Coding Scheme for categorical variables

Variables	Value	Freq	Parameter Coding (1)	(2)	(3)
Educhd					
Illiterate	1	1622	1.000	.000	.000
Grade 1-6	2	1198	.000	1.000	.000
Grade 7-12	3	949	.000	.000	1.000
Over Grade 12	4	260	-1.000	-1.000	-1.000
Age					
5-6 years	1	76	-1.000	-1.000	
7-12 years	2	2175	1.000	.000	
Above 12	3	1778	.000	1.000	
Incomehd					
Less than 170 birr	1	1481	1.000	.000	
170-1000 birr	2	2030	.000	1.000	
More than 1000 birr	3	518	-1.000	-1.000	
Fam_size					
Size of 1-2	1	77	-1.000	-1.000	
Size of 3-5	2	1029	1.000	.000	
Size of 6 & above	3	2923	.000	1.000	
Dst_schl					
Less than 1 km	1	4025	-1.000	-1.000	
1-3 kms	2	3	1.000	.000	
Greater than 4 kms	3	1	.000	1.000	

The final logistic regression model

The variables (including interaction terms) selected by the forward stepwise method were age, relation to head, educational level of the household head, income of the family, marital status of parents, urban/rural residence, tigray (a dummy coded 1 if a student lives in Tigray and 0 otherwise), addis (a dummy coded 1 if a child is from Region 14 and 0 otherwise), and affar x urb_rur (coded 1 if a student lives in rural Affar areas and 0 otherwise).

Goodness of fit of the model

i) The Classification Table

Of the 4029 primary school students included in the analysis (1119 cases were rejected because of missing data), 93.42% were correctly classified. This clearly shows that the final logistic model performs well.

Table 18 Classification Table for Y

		Predicted		
		Stay-ins	Dropouts	Percent
Observed		0	1	correct
Stay-ins	0	3763	4	99.89%
Dropouts	1	261	1	.38%
			Overall	93.42%

ii) The -2 log likelihood (-2LL) Statistic

For the logistic model that contains only the constant, -2LL is 1938.6355, as shown in Table 19.

Table 19 -2LL for the model containing only the constant

Dependent variable	Y	
Beginning block number	0	Initial Log Likelihood Function
-2 Log Likelihood	1938.6355	

* constant is included in the model

The goodness of fit statistics for the final logistic regression model are presented in Table 20.

Table 20 Goodness of fit Statistics for the final model

-2 Log Likelihood 1715.490

Statistic	Chi-square	d.f	Significance
Model chi-square	223.146	13	.0000
Improvement	4.324	1	.0376

There is a significant difference in -2LL between the final model and the model with only the constant (the model chi-square) at $\alpha = 0.05$ level. This indicates that the variables in the final model explain the data quite well.

Estimated coefficients (b_i's) and related statistics from the final logistic regression model are presented in Table 21 .

Table 21 **Variables in the final model**

Variable	b	S.E	Wald	exp(b)
Urb_rur	.3422	.1645	4.3303	1.4081
Educhd			19.9448	
Educhd(1)	.6306	.1683	14.0455	1.8787
Educhd(2)	-.0284	.1762	.0260	.9720
Educhd(3)	-.0364	.1903	.0366	.9642
Incomehd			12.7630	
Income(1)	.4643	.1325	12.2801	1.5909
Income(2)	.0589	.1276	.2132	1.0607
Mar_st	-.4950	.1882	6.9195	.6095
Tigray	-1.1303	.3067	13.5773	.3229
Addis	-.6852	.2471	7.6904	.5040
Relation	.7020	.1595	19.3689	2.0177
Age			41.1430	
Age(1)	-.5474	.1947	7.9025	.5785
Age(2)	.4064	.1891	4.6195	1.5014
Affar x urb_rur	1.8030	.6595	7.4753	6.0694
Constant	-3.3359	.2453	184.8717	

The odds of dropping out increased by a factor of 1.4081 for students who came from rural areas than those from urban areas.

The odds of dropping out increased by a factor of 1.8787 for students when the head of the family is illiterate as compared to the average effect of all other categories. The odds decreased for students who came from educated families.

When the monthly income of a family is less than 170 birr, the odds of dropping out of a student from such a family increased by a factor of 1.5909 while they increased by a factor of 1.0607 when the student came from a family who earn 170-1000 birr per month.

The odds of dropping out decreased by a factor of 0.6095 when the variable mar_st changes from 0 to 1 (from being married to non-married). They increased by a factor of 2.0177 when relation changes from 0 to 1 (from being common child to other relationship).

The standard age group for primary school students is 7-12. For students within this age group, the odds of dropping out decreased by a factor of 0.5785 as compared to the average effect of other age groups. The odds increased for over-aged (above 12 years) students.

The odds of dropping out of primary schools decreased by factors of 0.5040 and 0.3229 for students from Region 14 and Tigray region, respectively. The odds increased by a factor of 6.0694 for students who reside in rural areas of Affar region as compared to students who live elsewhere in Ethiopia.

Diagnostics

i) The standardized residuals

The standardized residuals for the final regression model are small. They are approximately normally distributed with mean zero and standard deviation 1 since 93.47 percent of them lie between -1 and 1 while 94.15 percent lie between -2 and 2 (Montgomery & Peck, 1992).

ii) The deviance

For the final logistic regression model, we also observed small values for the deviance. Those values that lie between -1 and 1 accounted for 93.43 percent while 95.69 percent lie between -2 and 2.

Small values for standardized residuals and the deviance indicate that the final logistic regression model fit the data quite well. The normal probability plots of the standardized residuals and deviance are given in Fig.1 and Fig. 2, respectively. There are departures from the straight line. This is to be expected, however, since there are cases for which the model does not fit well.

4.3.5 Logistic Regression Results for Secondary Schools

The coding scheme used to represent the categorical variables in the analysis is similar to that used for primary schools except the variable age. The variable age was coded as shown in Table 22 .

Table 22 Deviation coding scheme for variable age

			Parameter coding	
	Value	Freq	(1)	(2)
Variable				
Age				
5-12 years	1	39	-1.000	-1.000
13-18 years	2	1142	1.000	.000
Above 18	3	362	.000	1.000

The final logistic regression model

The forward stepwise selection method resulted in a model with variables: age, family size, oromiya (a dummy coded 1 if a student is from Oromiya region and 0 otherwise), affar (a dummy coded 1 if a student lives in Affar region and 0 otherwise), SNNPR (a dummy coded 1 if a student is from SNNPR and 0 otherwise), and the interaction term addis x urb_rur (coded 1 if a student if from rural areas of Region 14 and 0 otherwise).

Goodness of fit of the model

i) The Classification Table

As shown in Table 23, 90.80 % of the students were correctly classified into either of the two groups, stay-ins and dropouts. It is reasonable to concluded that the final model fits the data well since the percentage of correct classification using the model is as high as 90.80 percent.

Table 23 Classification Table for Y

		Predicted		
		Stay-ins	Dropouts	Percent
Observed		0	1	correct
Stay-ins	0	1398	4	99.71 %
Dropouts	1	138	3	~2.31%
			Overall	90.80%

ii) The -2 Log Likelihood (-2LL) statistic

For the logistic model that contains only the constant, -2LL is 943.4521.

Table 24 -2LL for the model containing only the constant

Dependent variable	Y	
Beginning block number	0	Initial Log Likelihood Function
-2 Log Likelihood	943.4521	

* constant is included in the model

The goodness of fit statistics for the final logistic regression model are presented below.

Table 25 Goodness of fit Statistics for the final model

-2 Log Likelihood 864.843

Statistic	Chi-square	d.f	Significance
Model chi-square	78.609	8	.0000
Improvement	5.432	1	.0198

We conclude that the final logistic regression model fits the data well since the model chi-square, the difference in -2LL between the final model and the model with only the constant, is significant at $\alpha = 0.05$ level.

For the final model, estimated regression coefficients (b_i's) and related statistics are presented in Table 26.

Table 26 **Variables in the final model**

Variable	b	S.E	Wald	exp(b)
Fam_ size			15.7412	
Fam_size(1)	.7092	.2850	6.1908	2.0324
Fam_size(2)	-.0614	.2744	.0501	.9404
Age			31.0382	
Age(1)	-.3697	.2329	2.5191	.6910
Age(2)	.6764	.2382	8.0649	1.9668
Oromiya	.4790	.2025	5.5936	1.6145
Affar	2.1067	.7886	7.2275	8.2211
SNNPR	1.0084	.3296	9.3606	2.7412
Addis x urb_rur	2.7688	.5227	28.0563	15.9392
Constant	-2.7060	.3529	58.7895	

Being in the standard age group for secondary school students (13-18 years) decreased the odds of dropping out by a factor of 0.6910 while being in the age group 19-24 increased the odds by a factor of 1.9668.

The odds of dropping out increased by a factor of 2.0324 for students who came from a family of size 3-5 while they decreased by a factor of .9404 for students who came from large families (size of 6 & above).

The odds increased by factors of 1.6145 , 2.7412, 8.2211 and 15.9392 for students from Oromiya region, SNNPR, Affar region, and rural areas of Region 14, respectively.

Diagnostics

i) Standardized residuals

About 90 percent of the standardized residuals computed for the final logistic regression model lie between -1 and 1 while 91.93 percent of them lie between -2 and 2. This shows that the (standardized) residuals are small, and are approximately normally distributed with mean zero and standard deviation 1.

ii) The deviance

A considerable percentage of the deviance (89.02 percent) lie between -1 and 1 while those which lie between -2 and 2 accounted for 92.78 percent.

Thus it is reasonable to conclude that the final logistic regression model fit the data well based on the values for the standardized residuals and the deviance. Fig.3 and Fig.4 show the normal probability plots for the standardized residuals and deviance, respectively. Here again, there are departures from the straight line. This is because there are cases for which the model doesn't fit well.

4.4 The Cox Regression Model

Let T be a nonnegative random variable, denoting the failure time of a student. The grade level at which a student dropped out of school is considered as the failure time (in years) of a student. It can also be regarded as the survival time of the student in school. Let us suppose that T has a probability density function $f(t)$.

The survival function (SF) , defined as the probability of surviving to at least time t, t >0, is given by

$$S(t) = P[T \geq t] = 1 - F(t)$$

where F(t) is the distribution function of T.

The survival function indicates the probability that a student survives from the time origin (grade 1) to some time beyond t.

The hazard function, $h(t)$, is the probability that a student drops out at time t, given that he or she has survived to that time. It is given by

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t / T \geq t)}{\Delta t}$$

The hazard function is also known as the instantaneous(relative) failure rate at time t.

As a preliminary analysis, the survival and hazard functions for each predictor variable were estimated from the observed survival times using the Kaplan-Meier method. This method does not require specific assumptions to be made about the underlying distribution of the survival times, that is, it is non-parametric. We compared the survival times of two or more groups of each predictor variable by plotting the corresponding estimates of their survival functions. The plots indicated that there is a clear cut difference between survival functions of groups of some variables including urb_rur, age, fam_size, income, occupation of the household head, affar, ben_gum, SNNPR, gambella and addis. It was also observed that, even if it was not uniform, the survival functions for groups of variables like sex, educhd, relation and dst_schl were different.

From the plots, we also observed that the estimated survival functions for those students who came from families earning a monthly income of less than 170 birr, engaged in agricultural activities, with family size of 3-5, and families who are illiterate or who completed primary

grades are always less than those for other students. The estimated survival functions for students who reside in rural areas and in Affar, Benshangul/Gumuz, SNNPR, Gambella regions are always less than those for students who live in urban areas, and other regions, respectively. For students who live in Addis Ababa, the estimated survival function is always greater than that for students who live elsewhere in Ethiopia. Students of age-group 7-17 (the standard age group) have a better chance of surviving in school than over-aged and under-aged students.

It is possible that the observed differences may be merely the result of chance variation. Hence the log-rank test (also known as the Mantel-Cox test) was used to test the null hypothesis that, in the population, two or more survival functions are equal against the alternative that the hazard of dropping out at any given time for a student in one group is proportional to the hazard at that time for a student in the other group. For variables whose survival functions cross (sex, educhd, dst_schl, tigray, amhara, somali, harari and dire), however, the general Wilcoxon (Breslow) test was used to test the hypothesis of equality of survival functions. Both tests are based on computing the weighted difference between the observed and expected number of dropouts at each of the time points. For the log-rank test, all the weights are equal to 1, and for the Breslow test, the weights are the number of risks at each time point. Both tests have chi-square distribution with one degree of freedom(Collett, 1994).

The results of the log-rank test, as well as the Breslow test, for testing the null hypothesis that the survival functions are the same for groups of all the predictor variables are presented in Table 27.

Table 27 **Test Statistics for Equality of Survival Functions**

Variable	d.f	Log-rank Statistic	Significance	Breslow Statistic	Significance
Sex+	1	1.30	.2548	6.99	.0082*
Educhd+	3	145.46	.0000*	151.68	.0000*
Urb_rur	1	384.10	.0000*	368.82	.0000*
Age	2	19.53	.0001*	10.43	.0054*
Relation+	1	16.95	.0000*	3.02	.0821
Fam_size	2	38.68	.0000*	20.67	.0000*
Income	2	102.04	.0000*	105.04	.0000*
Occuphd	1	260.18	.0000*	258.27	.0000*
Dst_schl+	2	8.75	.0126*	5.91	.0520
Mar_st	1	0.07	.7962	1.30	.2539
Tigray+	1	0.11	.7420	0.05	.8182
Affar	1	17.64	.0000*	17.76	.0000*
Amhara+	1	5.07	.0243*	0.49	.4819
Oromiya	1	2.63	.1049	2.41	.1206
Somali+	1	1.27	.2604	1.53	.2164
Ben_gum	1	19.41	.0000*	23.41	.0000*
SNNPR	1	89.10	.0000*	71.17	.0000*
Gambella	1	8.67	.0032*	5.98	.0144*
Harari+	1	0.85	.3566	0.18	.6717
Addis	1	45.72	.0000*	61.76	.0000*
Dire+	1	0.00	.9866	0.01	.9185

* Significant at 0.05 level

+ Breslow test was used

The log-rank test shows that the null hypothesis that the survival functions are equal among groups is rejected at the 0.05 level for variables like urb_rur, age, fam_size, income, occuphd, affar, ben_gum, SNNPR, gambella and addis. In other words, it shows that for these variables the assumption that the hazard of dropping out at any given time for a student in one group is proportional to the hazard at that time for a student in the other group (that is, the assumption of proportional hazards) is justifiable.

For variables like relation and dst_schl, the Breslow statistic values are insignificant at the 0.05 level. On the other hand, the log-rank values are significant for these variables. It may be because the Breslow test has very low power when the percentage of censored cases is large like the data under consideration (Prentice & Marek, 1979). In any case, sample based estimates of survival functions may cross even though the corresponding true hazard functions are proportional. Hence, there is no strong evidence against the assumption of proportional hazards for these variables.

Pairwise log-rank tests were done for variables having more than two categories to determine which of the groups are significantly different from each other. All comparisons for educhd are significant except grade 1-6 Vs grade 7-12. All comparisons are significant for age and income variables. Only category 2 Vs 3 (size of 3-5 Vs size of 6 & above), and category 1 Vs 3 (less than 1 km Vs greater than 3 kms) were significant for fam_size and dst_schl, respectively.

A log-rank test for trend across the ordered groups of educhd, income, age, fam_size and dst_schl is summarized in Table 28.

Table 28 **Log-rank Tests for Trend**

Variable	Log-rank statistic	d.f	Significance
Educhd	123.20	1	.0000*
Income	95.23	1	.0000*
Age	9.05	1	.0026*
Fam_size	29.30	1	.0000*
Dst_schl	8.63	1	.0033*

* Significant at 0.05 level

As it can be seen in Table 28, there is a linear trend across the educhd, income, age, fam_size and dst_schl groups.

4.4.1 The Model

Let $X = (X_1, X_2,..., X_p)'$ be a vector of covariates (independent variables) included in the study. When the dependent variable is the cumulative survival function, that is, the proportion of students "surviving" at time t , the model is

$$S(t) = \left(S_o(t)\right)^{\exp(\beta X)}$$

where $S_o(t)$ is the baseline survival function and β is a vector of parameters. The baseline survival function is similar to the constant term in multiple regression in that it is the reference value that is increased or decreased depending on the values of the independent variables and their relationship with the dependent variable. Variables with positive coefficients are associated with decreased survival times while variables with negative coefficients are associated with increased survival times.

The Cox regression model defined above can be written in terms of the hazard function as

$$h(t) = (h_o(t))e^{\beta X} = [h_o(t)]e^{(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)}$$

where $h_o(t)$, known as the baseline hazard function, is the hazard function for an individual for whom the values of all the explanatory variables that make up the vector X are zero.

The Cox model expressed in terms of the hazard function is called the *Cox proportional hazards model*. The reason for the name " proportional hazards" is the fact that for two cases, the ratio of their hazards will be a constant for all time points. The assumption of proportional hazards for two or more groups implies that the cumulative survival curves, and the hazard curves for groups do not cross. The **log-minus-log** (LML) plot, a useful plot for assessing whether the baseline hazard functions are proportional, is used to test the proportional hazards assumption. The lines for the individual categories, for each covariate, are more or less parallel. This implies that the proportional hazards assumption is justifiable for all the variables included in the model.

The coding schemes used to represent the categorical variables included in the analysis are similar to those used in logistic regression model. That is, a dummy-variable coding scheme is used to represent dichotomous variables such as sex, urb_rur, mar_st, relation, occuphd, and the eleven regional variables (tigray, affar, amhara, oromiya, somali, ben_gum, SNNPR, gambella, harari, addis and dire) while a deviation coding scheme is used to represent categorical variables having more than two categories such as educhd, incomehd, fam_size, dst_schl and age. Interaction terms similar to those included in logistic regression model are also included in the analysis.

The β-parameters in the proportional hazards model can be obtained by maximizing the log-likelihood function using the Newton-Raphson procedure. The likelihood function used is not a true likelihood since it does not make direct use of the actual censored and uncensored survival times. For this reason, the estimated β-parameters are usually called the maximum partial likelihood estimates. Details on the method of partial likelihood, the technique used for estimating the model parameters are given in Kalbfleisch & Prentice (1980) and Collett (1994).

A FORWARD STEPWISE selection method was used for model building. In forward selection, variables are considered one at a time for entry into the model. After a variable is added to a model, all variables already in the model are examined for removal. The algorithm stops when no or more variables meet entry or removal criteria, or when the resulting model is identical to a previous one. As in logistic regression, the score statistic was used for entering variables into a model. The likelihood-ratio statistic based on conditional parameter estimates which is due to Lawless & Singhal (1978) was used for variable removal. This statistic was chosen for two reasons: (i) it does not require reestimation of the model unlike the likelihood-ratio statistic based on the maximum partial likelihood estimates, and (ii) in many situations it tends to give similar results to the statistic based on the maximum partial likelihood estimates.

4.4.2 The final proportional hazards model

The likelihood-ratio test for the hypothesis that all parameters are zero is obtained by comparing -2LL for a model in which all coefficients are 0 with -2LL for a model that contains the coefficients of interest. For the initial Cox proportional hazards model, -2LL is 6535.405 (Table 29).

Table 29 **Likelihood statistics for the initial model**

Beginning Block Number 0	Initial Log Likelihood Function
-2 Log Likelihood	6535.405

The value of -2LL for the final proportional hazards model is 6164.556. The difference in -2LL between the initial and final proportional hazards models, labeled as *Change*(-2LL) *from previous Block* in Table 30, is significant at 5 % level of significance- that is the null hypothesis that all the coefficients are zero is rejected. The test of the null hypothesis that all the parameters are zero can also be based on the *Score statistic* (also known as *the global chi-square* or the *overall chi-square*). The score or the overall chi-square statistic for the final proportional hazards model is also significant at 0.05 level. Hence we reject the null hypothesis that all the parameters in the model are zero and conclude that the variables in the final model have explained the data quite well.

Table 30 **Likelihood Statistics for the final proportional hazards model**

-2 Log Likelihood	6164.556		
Statistic	Chi-square	d.f	Significance
Overall (Score)	475.045	18	.0000
Change(-2LL) from Previous Block	370.849	18	.0000

The twelve covariates selected by the forward stepwise selection procedure include urb_rur, educhd, incomehd, fam_size, relation, age, dst_schl, mar_st, somali, ben_gum, SNNPR, and SNNPR x urb_rur. The estimated model parameters for final Cox proportional hazards model are summarized in Table 31.

Table 31 **Statistics for the final proportional hazards model**

Variable	b	S.E	Wald	d.f	Sign.	Exp(b)
Urb_rur	2.1437	.3525	36.980	1	.0000	8.5312
Educhd			40.0539	3	.0000	
Educhd(1)	.6253	.1073	33.9271	1	.0000	1.8687
Educhd(2)	.0587	.1117	.2759	1	.5994	1.0604
Educhd(3)	.1534	.1122	1.8678	1	.1717	.8578
Incomehd			14.3762	2	.0008	
Incomehd(1)	.3444	.0909	14.3700	1	.0002	1.4112
Incomehd(2)	.0436	.0815	.2859	1	.5929	1.0445
Fam_size			12.3042	2	.0021	
Fam_size(1)	.2273	.0989	5.2814	1	.0216	1.2551
Fam_size(2)	-.1629	.1001	2.6486	1	.1036	.8496
Relation	.6102	.1188	26.3883	1	.0000	1.8409
Age			9.9394	2	.0069	
Age(1)	-.5487	.1759	9.7363	1	.0018	.5777
Age (2)	-.4185	.1809	5.3543	1	.0207	.6580
Mar_st	-.6125	.1356	20.3905	1	.0000	.5420
Somali	1.2849	.5856	4.8147	1	.0282	3.6144
Ben_gum	.7597	.3163	5.7685	1	.0163	2.1376
SNNPR	.9376	.2085	20.2219	1	.0000	2.5538
SNNPR x Urb_rur	-.6308	.2752	5.2522	1	.0219	.5322
Dst_schl x Urb_rur			9. 4134	2	.0090	
Dst_schl(1) x Urb_rur	1.2078	.5425	4.9563	1	.0260	3.3462
Dst_schl(2) x Urb_rur	-.2416	.5404	.1999	1	.6548	.7854

families, respectively. On the other hand, the hazard of dropping out decreased by a factor of 0.6784 for students from a high-income earning families.

It was also observed in Table 31 that the hazard of dropping out increases for a student who came from a family of size 3-5 while it decreases for a student who came from a large family (family of size 6 & above).

As it was expected, the hazard of dropping out decreases for students in the standard age group. The estimated risk of dropping out is increased by a factor of 3.3462 for students who live in rural areas with a daily walk of 1-3 kms to school.

The estimated risk of dropping out is 3.6144 times greater for a student from the Somali region as compared to that for a student from other regions. The relative risks for students from the Benshangul/Gumuz and SNNPR regions are 2.1376 and 2.5538, respectively.

4.4.3 Diagnostics

The adequacy of the fitted proportional hazards model was assessed by examining the **martingale residuals**. The Cox-Snell residuals for the ith individual, i = 1, 2, ..., n is given by

$$r_{C_i} = \exp(b'x_i)h_o(t_i),$$

where $h_o(t_i)$ is the estimated cumulative baseline hazard function at time t_i, the observed survival time of that individual, and b is a vector of estimated parameters. The Cox-Snell residual is just the estimated cumulative hazard. The martingale residuals are therefore of the

form

$$r_{M_i} = \begin{cases} -r_{C_i} \text{ for censored observations} \\ 1 - r_{C_i} \text{ for uncensored observations} \end{cases}$$

where r_{C_i} is the Cox-Snell residual for the ith observation.

The martingale residuals are similar to residuals encountered in linear regression analysis in that, in large samples, they are uncorrelated with one another and have an expected value of zero. However, the martingale residuals are not symmetrically distributed about zero.

Plots of martingale residuals against the values of the independent variables or the linear predictor score (X'Beta), which is the sum of the products of the coefficients and the mean-corrected variables for a case, are useful in detecting any violations of the proportional hazards model. Figures 5-9 show plots of the martingale residuals against the values of age, educhd, dst_schl, fam_size and incomhd, respectively. Most of the points in the plots fall in a horizontal band around zero implying that the fitted model is satisfactory. We excluded plots of the martingale residuals against the values of the dichotomous variables because the patterns in such plots make them difficult to interpret. The plot of the martingale residuals against the linear predictor scores (Figure 10) also confirms the conclusion that the fitted model is satisfactory.

It is possible to construct a model that does not force the ratio of the hazard rates for the two groups to be constant over time. That is, a Cox regression model that incorporates time-dependent covariates (predictor variables whose values change with time) can be constructed. This model is less restrictive than the proportional hazards model because it allows us to include

interaction terms, which we think are meaningful, such as age x t_cov, educhd x t_cov, dst_schl x t_cov, incomehd x t_cov and fam_size x t_cov (where t_cov is the survival time) in the analysis. Unfortunately, we have covariate information only at the time of the survey, and this makes it difficult to interpret the results (Trussell & Hammerslough, 1983).

In this chapter, we have attempted to examine the effects of several covariates on dropping out using three statistical techniques: discriminant analysis, logistic regression and Cox proportional hazards model. Using discriminant analysis, we found that variables like urb_rur, age, relation, incomehd, dst_schl, educhd, mar_st, tigray, addis, amhara, gambella & oromiya significantly affect dropping out in primary schools of Ethiopia while variables like age, educhd, urb_rur, relation, mar_st, fam_size, affar, gambella, tigray, amhara and dire affect dropping out in secondary schools significantly.

On the other hand, using the logistic regression model, we found that the variables which significantly affect dropping out in primary schools are urb_rur, educhd, incomehd, mar_st, age, relation, tigray, addis & affar x urb_rur while those in secondary schools are age, fam_size, oromiya, affar, SNNPR, addis x urb_rur. It was also found that there are no significant reductions both in total hit rate estimates and improvement over chance indices when the variables selected by the logistic regression model were entered in the SAS DISCRIM procedure. The discrepancies observed are due to the use of the stepwise selection method. That is, in using the stepwise selection method: (i) the best subset of a given size may not emerge and (ii) only one "good" subset is suggested (Huberty, 1994). It may also be due to the inclusion of interaction terms in the logistic regression model. The logistic regression model, unlike

discriminant analysis, allows us to examine not only the effect of the variables on dropping out but also the effect of each category of the variables on dropping out.

The Cox proportional hazards model has all the advantages of logistic regression over discriminant analysis. In addition it has the advantage of incorporating the grade levels at which students dropout of school. Using this model, we found that educhd, relation, mar_st, urb_rur, SNNPR, incomehd, fam_size, age, dst_schl x urb_rur, SNNPR x urb_rur, ben_gum and somali significantly affect the hazard of dropping out, in decreasing order. They are ordered in terms of the loss incurred in chi-square when the term is removed.

CHAPTER V

5. Discussion and Conclusion

The main objective of this study was to examine the effects of various external factors- factors that are not directly related to the school system- on dropping out of school. It was also of interest to make regional as well as gender comparisons. We, therefore, discuss the general and specific conclusions that can be made based on the findings of the study.

Factors that significantly affect dropping out in primary schools differ from those in secondary schools. Urban/rural residence, educational level of the household head, monthly income of the household head, marital status of parents, age of the student and relation to household head are the major factors that contribute to dropping out in primary schools. Dropouts in primary schools are more likely to come from rural areas, from parents with no or low educational level and from parents who earn a monthly income of 170 Birr or less. These findings are consistent with the findings of the other studies about school dropouts in developing countries and to some extent in developed countries.

Primary school dropouts are also more likely to be overaged (above age 12). Age at entrance has much to do with this. A child in Ethiopia is expected to be 7 years old on entry to grade 1 but, according to the data obtained from the Ministry of Education, about 46 percent of the students who were registered at grade 1 in the 1995/96 academic year were above age 10. By the time they finish the primary cycle, they will be above age 15. This might lower their aspirations to learn and eventually lead them to drop out. Students at the age of late adolescent may quit school

because of lack of motivation which is caused by the educational system that does not prepare them for the adult world but only to mere learning.

It was also observed that students who are not common children to the household head and his/her spouse have a higher risk of dropping out. The odds of dropping out of primary schools in Tigray region and Region 14 are significantly lower than those in other regions. The factor by which the odds of dropping out increased for students in primary schools of rural areas of Affar region is alarming.

The risk of dropping out of school is not increased for students who came from non-married (single, divorced, separated or widowed) parents. This conclusion confirms the result obtained by Tadesse(1975).

On the other hand, factors that significantly affect dropping out in secondary schools include age of the student and the size of the household. Like primary school dropouts, dropouts from secondary schools tend to be overaged.

Contrary to the general literature, we found that students who came from families of size 3-5 have a higher risk of dropping out than those who came from large families (size of 6 & above). It can be argued that students from relatively smaller families carry a heavy burden of responsibilities than those from large families where the heavy load can be shared by many.

The odds of dropping out from secondary schools appear to be higher in Oromiya region, Affar region, Southern Nations and Nationalities People Region (SNNPR), and rural areas of Region 14.

Unlike other studies in this area, we observed that sex is not among the major factors that significantly contribute to dropping out except that the survival function for female students from grades 1 to 9 was higher than that for male students while it was less than that for male students beyond grade 10. It may be because some factors such as early marriage, child bearing responsibility, and psycho-social problems such as pre-marital sex, teenage pregnancy, teenage abortions and disinterest in studies force adolescent girls to quit school.

The Cox proportional hazards model suggested that the survival times of students in the school system are shorter for students who came from rural areas, from parents with no or low educational level, from low and medium- income earning families and from families of size 3-5. This shows how serious the dropout problem is because the lower the grade level at which the students dropout, the greater the probability that the students leave the school system before they have acquired the necessary knowledge that could help improve their life and that of their society.

Students drop out of school for various reasons. About 17 percent of the dropouts responded that they left school because they were tied up with work while 16 percent dropped out of school because it is too costy. The majority of the dropouts reported that they discontinued going to school for reasons other than being tied up with work, can not afford, the school is too far and poor quality. According to Tekeste (1990), the educational crisis (dropping out is one of the

indicators) in Ethiopia is observed due to the reasons that the expansion of education sector beyond the economic development of the country, and the irrelevance of the curriculum that does not respond to the needs and interests of individuals and society. Some studies suggested that parents withdraw their children from school because the academic year interferes with the farming seasons (Brimer & Pauli, 1971; Tadesse,1975; USAID, 1993 & Yelfign, et al., 1995).

The dropout problem is unlikely ever to go away. All our effort is to minimize its incidence since education is undoubtedly a basic requisite for responsible citizenship for maintenance of our way of life, and for successful entry into today's complicated working world. The first step towards effective interventions, that is , identifying the major factors that significantly contribute to dropping out of school, has been taken in this study. Based on the findings of the study, some recommendations that may help in the development of policies that deals with the dropout problem are made:

i) Dropping out of school is more common among students who came from families with no or low educational level. It is because parents who have not been to school or , despite having been in school, remain illiterate and innumerate fail to understand in what way most of what their children do will be of benefit to them. Schools should, therefore, work with such parents to make them understand the aims and activities of the school and the benefits of education in general;

ii) Schools should provide guidance and counseling services that

- assist students, who have to work to add to the meagre income of the family, to budget their time properly;
- provide guidance, to adolescent boys and girls whose probability of dropping out were found to be high, to perceive alternatives to being dropouts.

iii) Further research efforts should focus on developing comprehensive models that incorporate both internal and external factors. Such efforts will improve our understanding of the problem and help us identify potential dropouts at an early age and this in turn helps us to initiate early interventions for disadvantaged students who generally have a high probability of dropping out.

References

Austin, A.W. (1975). Preventing Students from Dropping Out. Jossey-Bass, London.

Brimer, M.A, & Pauli, L. (1971). Wastage in Education: A world Problem. UNESCO, Paris: IBE, Geneva.

Broadfoot, P., & Shutherland, M.B. (1987). Educating for Changing Sex Roles. *Comparative Education*, Vol. 23, No.1.

Collett, D. (1994). Modelling Survival Data in Medical Research. Chapman & Hall, London.

Constanza, M.C., & Afifi, A.A (1979). Comparison of stopping rules in forward stepwise discriminant analysis. *Journal of American Statistical Association*, Vol. 74.

Coombs, P.H.(1968). The World Educational Crisis: A system Analysis. Oxford University Press.

Coombs, P.H.(1985). The World Crisis in Education: From the View of the Eighties. New York University Press.

Cox, D.R.(1972). Regression Models and Life Tables (with discussion). *Journal of Royal Statistical Society*, Series B, Vol 34.

Crenbach, L. (1962). Educational Psychology. New York: Harcourt, Brace and World.

Desta Asayehgn (1977). Students' Alienation: A Study of High School Students in Ethiopia. London: University of Microfilms International.

Douglas, J.W.B; Ross, J.M. & Simpson, H.R. (1968). All Our Future: A longitudinal study of secondary education. London.

Douglas, W.A (1977). Glad to be Out? Edinburgh, Scottish Council for Research in Education.

Draper, N.R, & Smith, H. (1981). Applied Regression Analysis, 2nd ed., Wiley, New York.

Edigar, M. (1987). School Dropouts, Absenteeism and Tardiness. US Department of Education. Office of Educational Research and Improvement (OERI).

Edington, E.D. (1970). Disadvantaged Rural Youth. Review of Educational research, Vol. 40, No.1.

Ekstrom, R.B.; Goertz, M.E.; Pollack, J.M., & Rock, D.A. (1986). Who drops out of high school and why? Findings from a National Study. *Teachers College Record*, Vol. 87, No.3.

Elliott, D.S; Voss, H.L. & Wending, A. (1966). Capable Dropouts and the Social Melieu of the High School. *The Journal of Educational Research*, Vol. 60, No.4.

Fafunwa, A.B. Education in the Mother Tongue: A Nigerian experiment-the six-year (Yoruba medium) primary education project at the University of IFE, Nigeria.

Fine, M. (1986). Why Urban Adolescents Drop into and out of Public High school. *Teachers College Record*, Vol. 87, No. 3.

Frey, S.H. (1970). Adolescent Behaviour in School. Chicago, Rand-McNally and Comp.

Halpern, R. (1986). Effects of Early Childhood Intervention on Primary school Progress in Latin America. *Comparative Education Review*, Vol. 30, No.2.

Hand, D.J (1981). Discrimination and Classification. Wiley, New York.

Hathaway, S.R, et al. (1963). Adolescent Personality and Behavior. University of Minnesota, U.S.A.

Hauck, W.W., & Donner, A. (1977). Wald's test as applied to hypothesis in logit analysis. *Journal of the American Statistical Association*, Vol. 72.

Hosmer, D.W., & Lomesow, S (1989). Applied Logistic Regression. Wiley, New York.

Huberty, C.J (1994). Applied Discriminant Analysis. Wiley, New York.

Hughe, J.M.(1973). The Slow learner in your class. Hazell, Watson and Viney Ltd, London.

Jennings, D.E.(1986). Judging inference adequacy in logistic regression. *Journal of the American Statistical Association*, Vol. 81.

Kablaoui, B. N. & Pautler, A.J. (1991). The Effects of Part-time Work Experience on High School Students. *Journal of Career Development*, Vol.17, No.3.

Kalbfleisch, J.D., & Prentice, R.L (1980). The Statistical Analysis of Failure Time Data, Wiley, New York.

Kolstad, A.J., & Owings, J.A. (1986). High school dropouts who change their minds about school. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Lachenbruch, P.A (!975). Discriminant Analysis. Hafner Press, New York.

Lawless, J.F., & Singhal, K.(1978).Efficient screening of nonnormal regression models. *Biometrics*, Vol. 34.

Lawless, J.F (1982). Statistical Models and Methods for Lifetime Data, Wiley, New York.

Lehtonen, R., & Pahkinen, J.E. (1995). Practical Methods for Design and Analysis of Complex Surveys. Wiley, Chichester, England.

Levy, M.B. (1971). Determinants of Primary School Dropouts in Developing Countries. *Comparative Education Review*, Vol. 15, No.1.

Lichter, S.O, et al. (1962). The Dropout. The Free Press of Glencoe, New York.

Lockheed, M., and Verspoor, A.M., with Bloch, D., Englebert, P., Fuller, B.. King, E., Middleton, J., Paques, V., Rodd, A., Romain, R., & Welmond, M.(1991). Improving Primary Education in Developing countries. Oxford University Press, New York.

Meerman, J.(1980). Paying for Human Development. In Implementing Programs of Human Development, ed. P.T.Knight. World Bank Staff Working Paper No. 403. Washington, D.C.

Mialaret, G. (1979). The Child's Right to Education. Paris: UNESCO.

MoE (1969). Department of Program, Planning and Research. Addis Ababa.

MoE (1997). Education Statistics, Annual Abstract, Ministry of Education, Addis Ababa.

Montgomery, D.C., & Peck, E.A., (1992). Introduction to Linear Regression Analysis, Wiley, New York.

Morrison, D.F (1976). Multivariate Statistical Methods, 2nd ed., McGraw-Hill, New York.

Nag, M., et al. (1977). The Economic Value of Children in Two Peasant Societies, Proceeding of the International Population Conference. Leige, Belgium: International Union for the Scientific Study of Population.

Neill, S.B. (1979). Keeping Students in School: Problems and solutions. AASA critical issues report. Arlington, VA: American Association of School Administrators.

Norusis, M.J. SPSS Inc. (1990). SPSS/PC + Advanced Statistics 4.0 for the IBM PC/XT/AT and PS/2. SPSS Inc: Chicago.

Perraton, H. (1983). Secondary education at a distance. International Extension College, London.

PHRD: Policy for Human Resources Development (1996). Household Demand for Schooling, Addis Ababa.

Prentice, R.L., & Marek, P. (1979). A qualitative discrepancy between censored data rank tests. *Biometrics*, Vol. 35.

Rao, C.R (1973). Linear Statistical Inference and Its Applications, 2nd ed., Wiley, New York.

Rosenbaum, P.R. (1986). Dropping Out of High School in the United States: An Observational Study. *Journal of Educational Statistics*, Vol. 11, No.3

Rumberger, R.W. (1983). Dropping out of high school: The influence of race, sex, and family background. *American Educational Research Journal*, Vol. 20.

Rumberger, R.W. (1987). High School Dropouts: A Review of Issues and evidence. *Review of Educational Research*, Vol. 57, No. 2.

Russell, K. (1968). Stay in School. *American Education*, Vol.4, No.6.

Steinberg, L., Blinde, P.L., & Chan, K.S.(1984). Dropping out among language minority youth. *Review of Educational Research*, Vol. 54.

Tadesse Mengesha (1975). Primary School Dropouts in Rural Ethiopia: Planning and Policy Implications. Ann Arbor; Michigan: University of Microfilms International.

Tekeste Negash (1990). The Crisis of Ethiopian Education; Some Implications for Nation Building. Uppsala University, Sweden.

Trussell, J., & Hammerslough, C. (1983). A hazards model analysis of the covariates of infant and child mortality in Sri Lanka, *Demography*, Vol. 20, No.1

UNESCO (1971). Social Background of Students and Their Change. UNESCO, Geneva.

UNESCO (1975). Women, Education and Equality: A Decade of Experiment. Paris.

UNESCO (1988). Family Life and Sex Education. UNESCO, Thailand.

USAID (1993). The Demand for Primary Schooling in Rural Ethiopia: A Research Study, Addis Ababa.

Wehlage, G.G, & Rutter, R.A(1986). Dropping out: How much do schools contribute to the problem. *Teachers College Record*, Vol. 87, No. 3.

World Education Report, 1995.

Ymen, C.H. (1967). Child Guidance Clinic. McGraw Hill Book Company, New York.

Yelfign W., Zewdu D., Alemnesh H., & Anbesu B. (1995). Study on Primary School Female Participation and Performance in Cheha District, Ministry of Education, Addis Ababa.

ANNEXES

Fig. 1 Normal P-P Plot of Standardized Resi
for Primary Schools

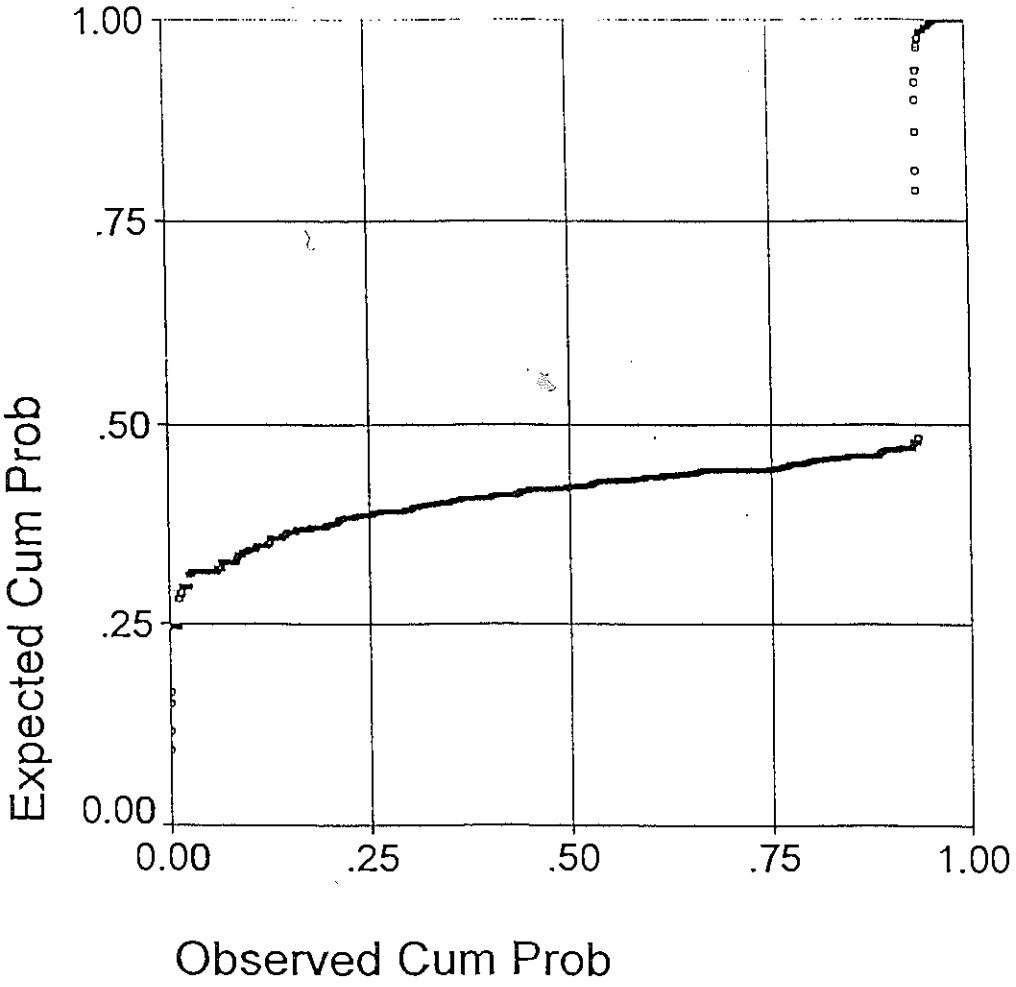


Fig.2 Normal P-P Plot of Deviance for Primary Schools

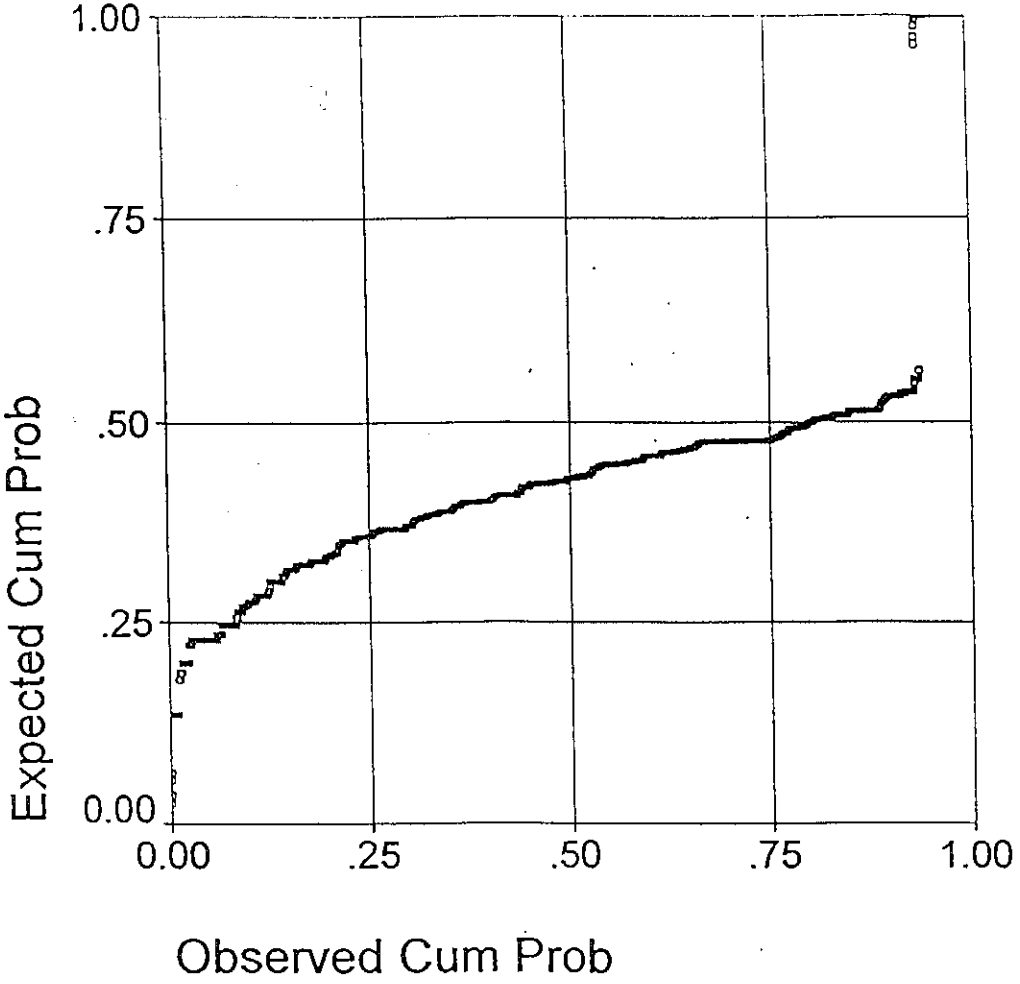


Fig. 3 Normal P-P Plot of Standardized Res for Secondary Schools

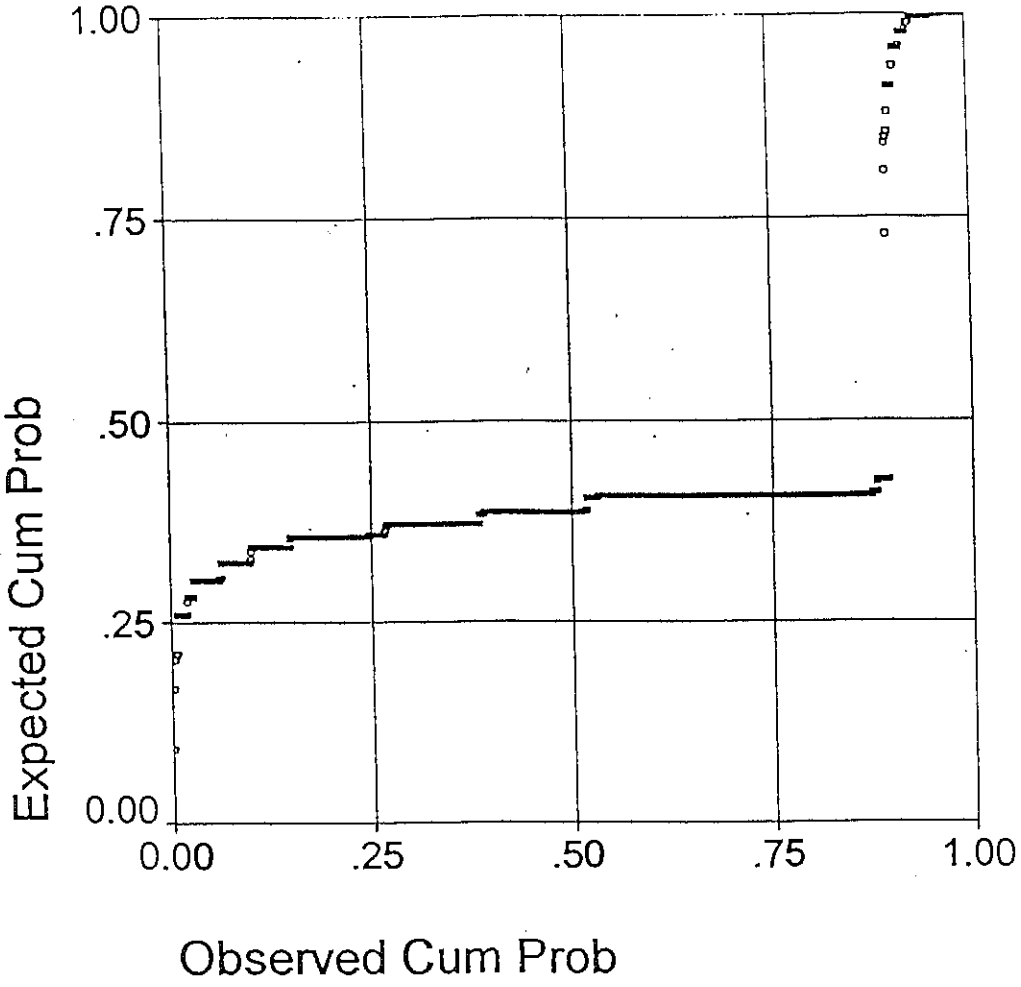


Fig.4 Normal P-P Plot of Deviance for Secondary Schools

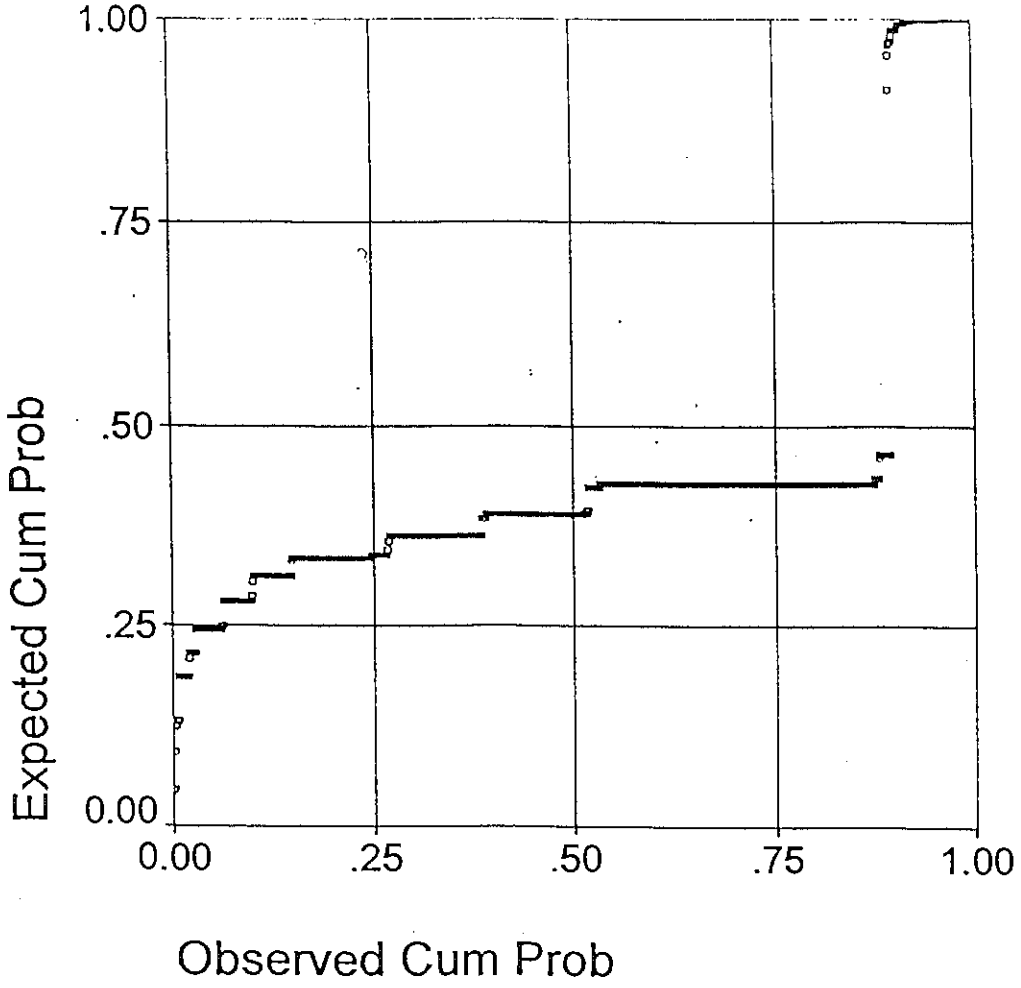


Fig.5 Plot of martingale residuals against age

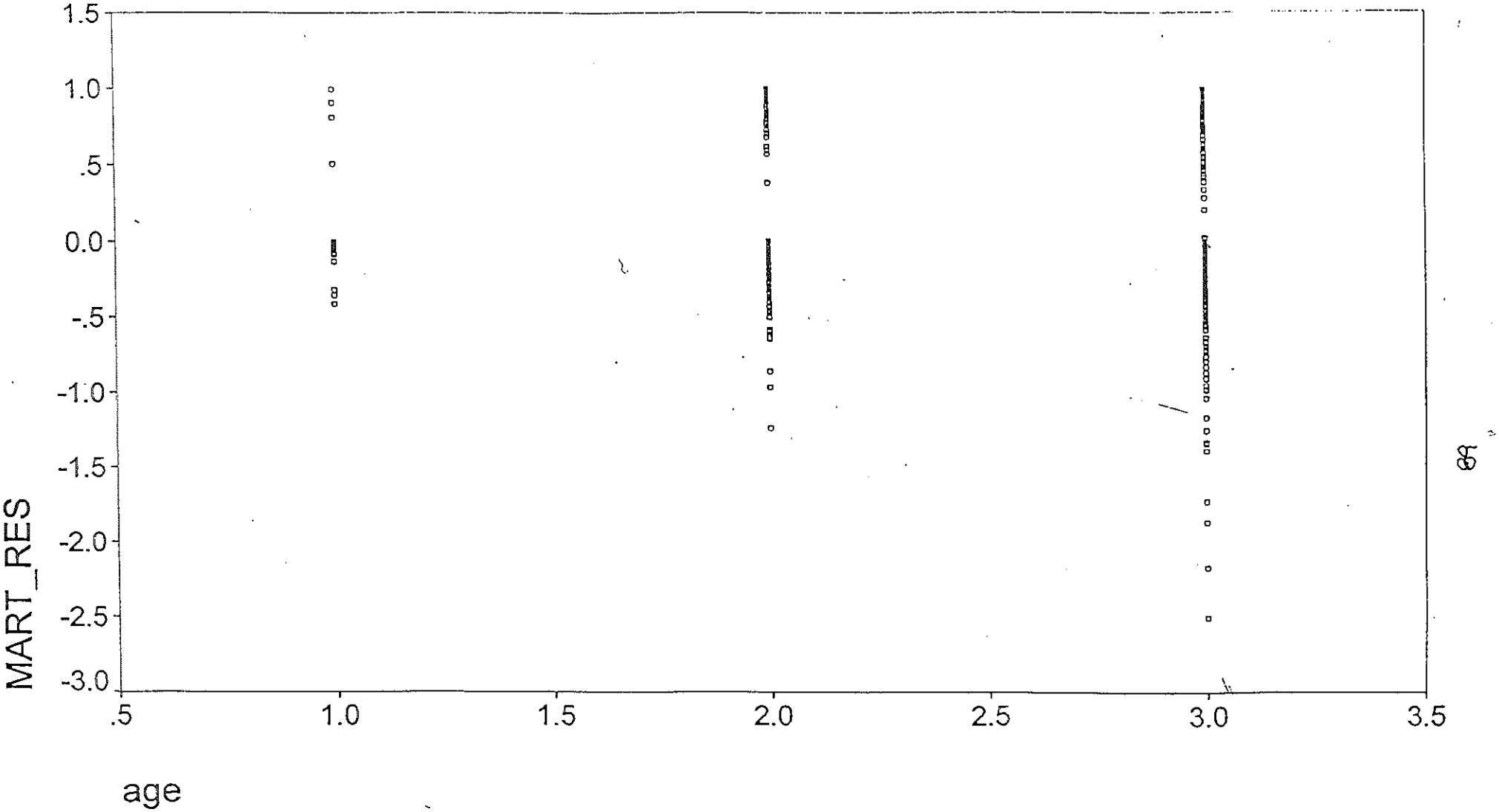


Fig. 6 Plot of martingale residuals against educational level c

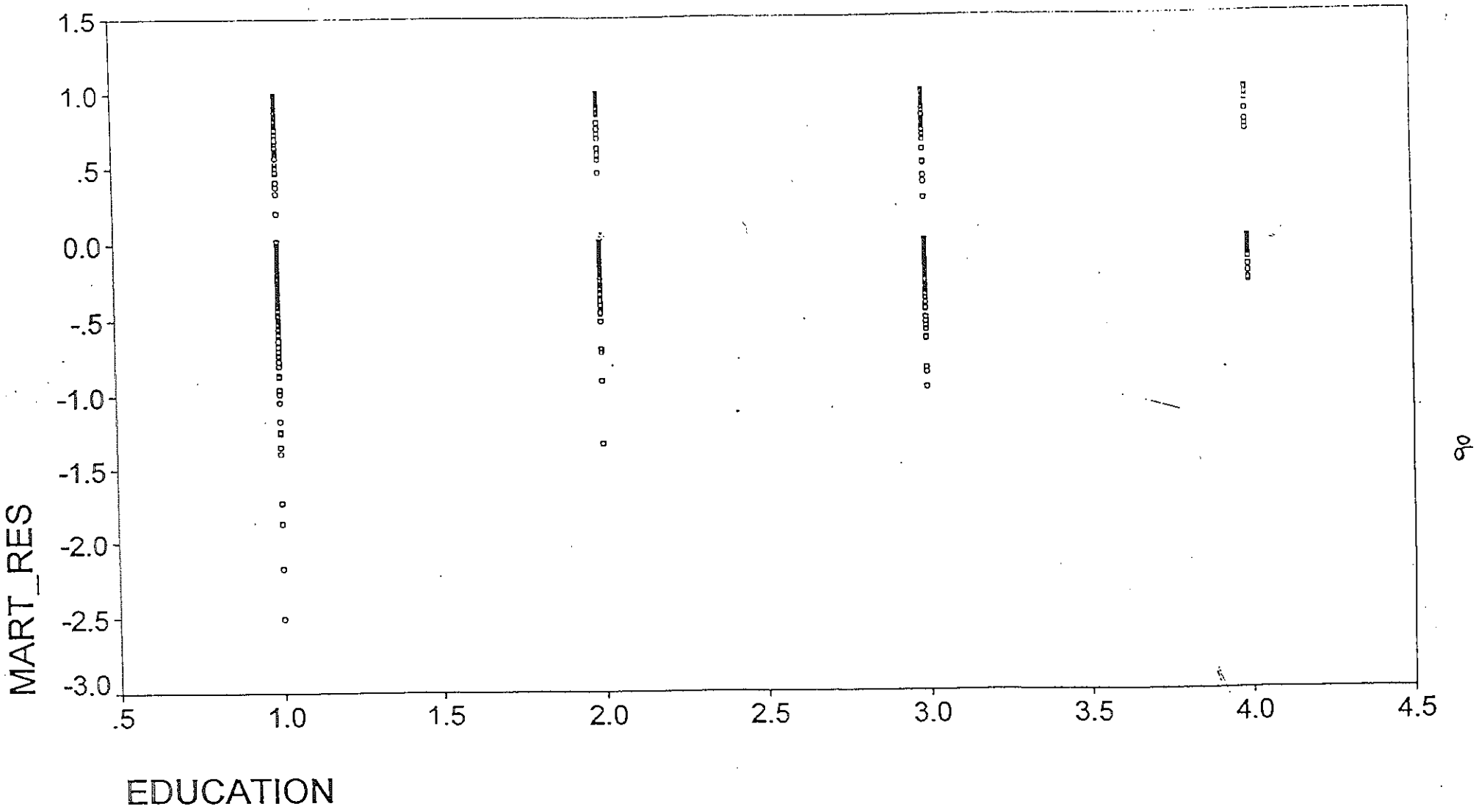


Fig. 7 Plot of martingale residuals against distance to schoc

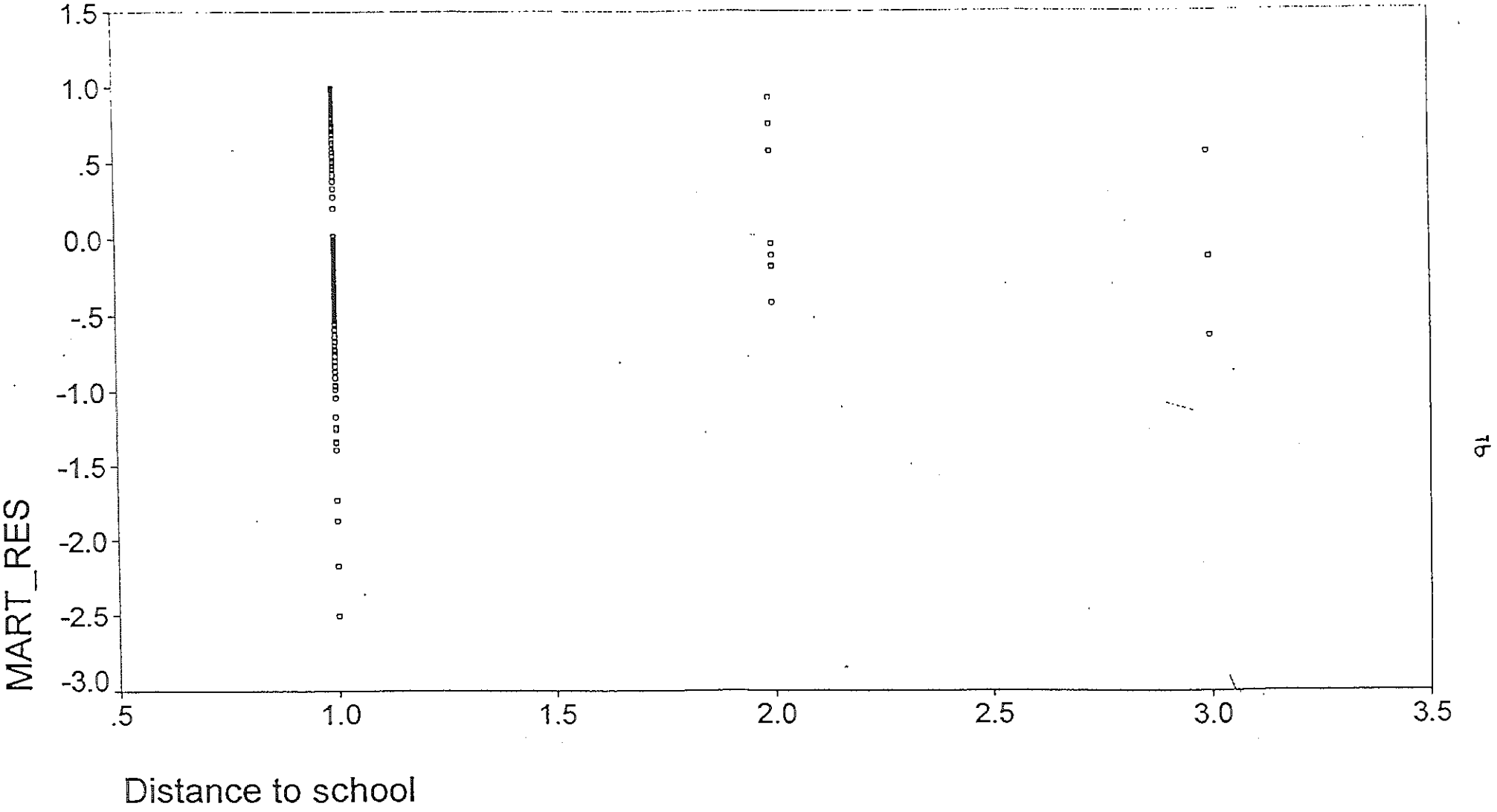


Fig. 8 Plot of martinagle residuals against family size

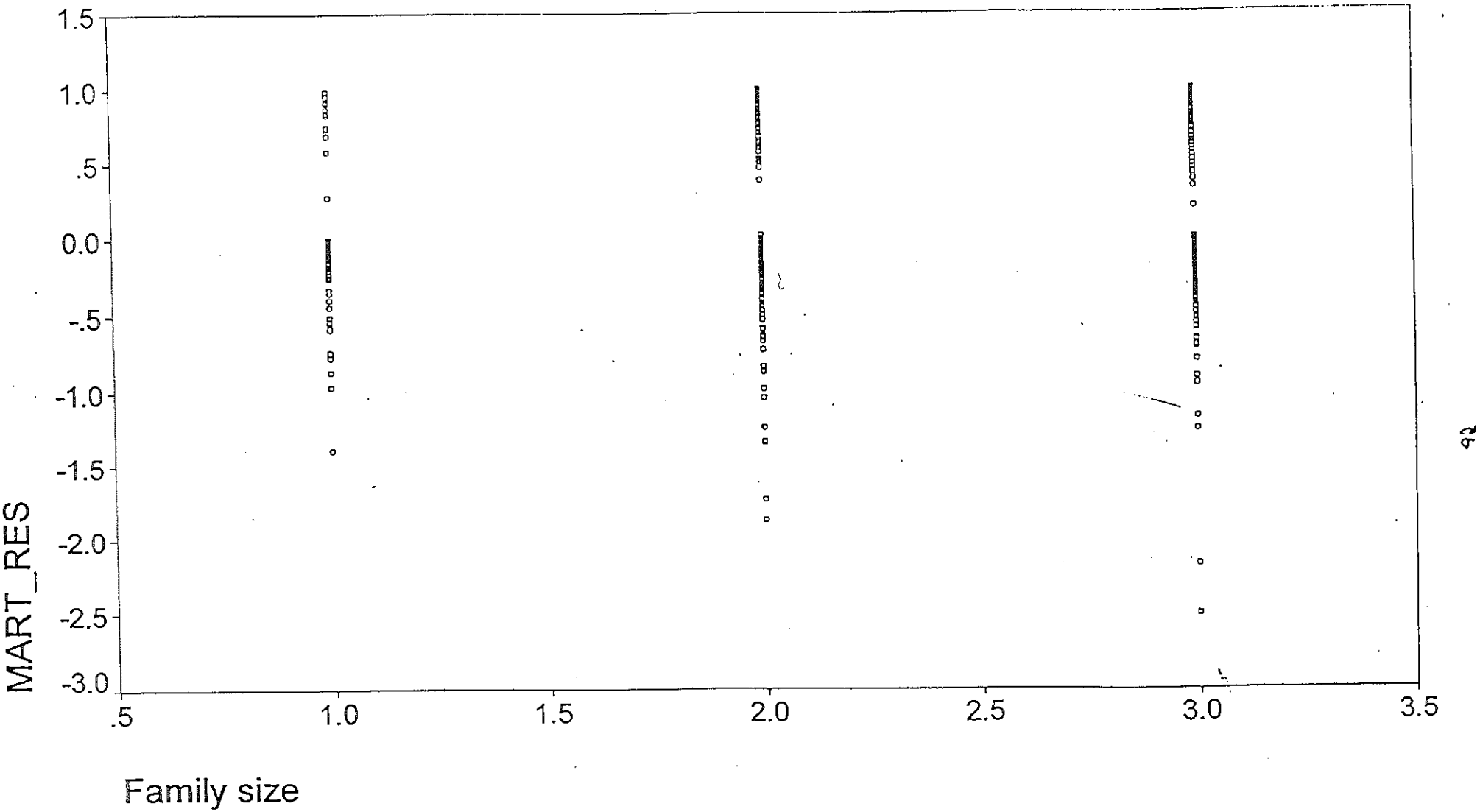


Fig. 9 Plot of martingale residuals against income

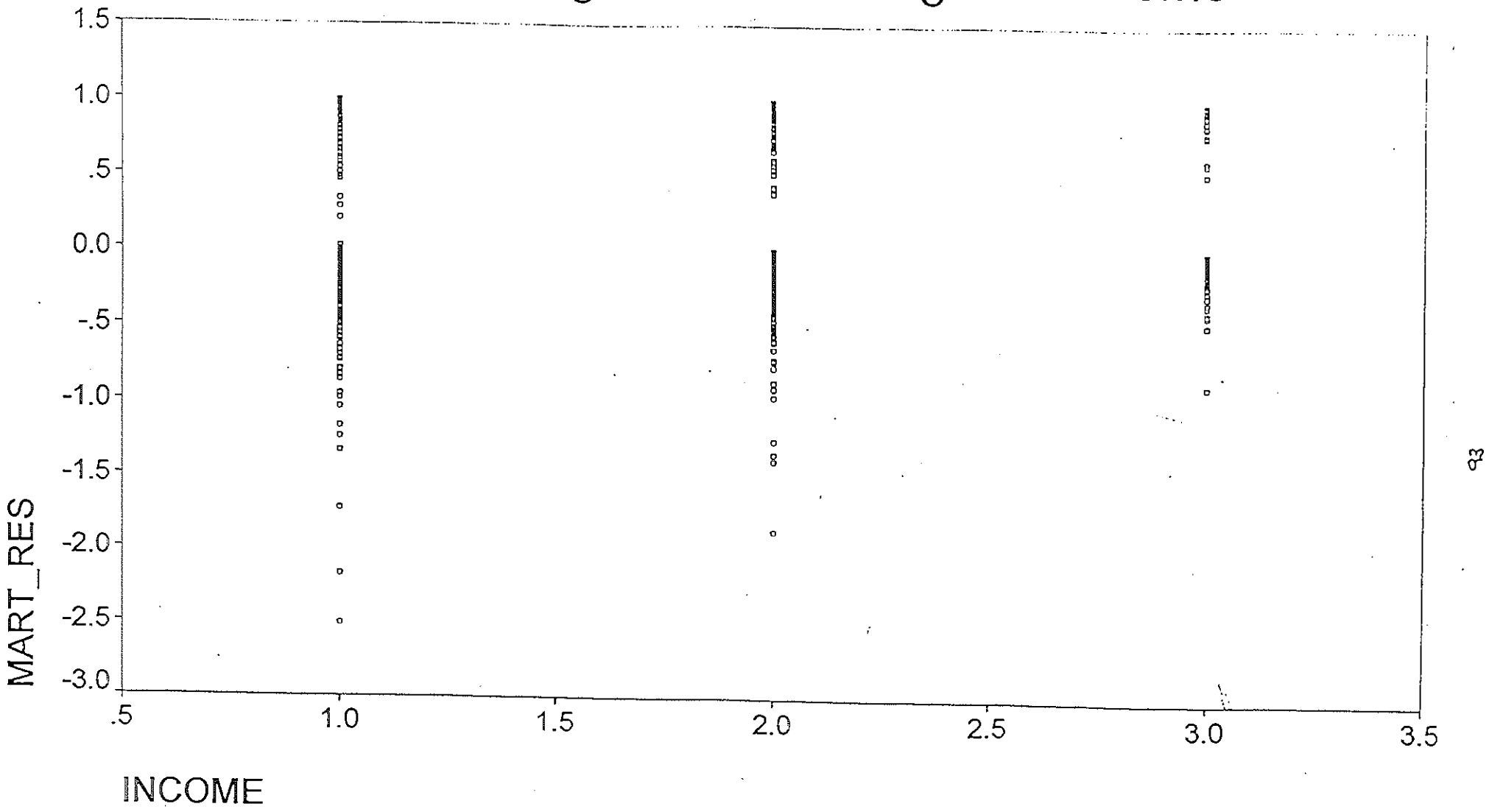


Fig. 10 Plot of martingale residuals against X'Beta

