THREE ESSAYS ON FAMILY STRUCTURE AND SCHOOL DROPOUT AMONG ADOLESCENTS

BY

LEILA FARIVAR

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Economics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2010

Urbana, Illinois

Doctoral Committee:

Professor Hadi S. Esfahani, Chair
Professor Werner Baer
Professor Anil K. Bera
Assistant Professor Kristine M. Brown

Abstract

This dissertation is a collection of three essays on high school dropout by adolescents and the roll of their siblings in this risky behavior. Dropping out of high school can scar the individual for his entire labor supply period with lower earnings, higher unemployment, and in turn incur considerable social and welfare costs to the economy. The data used in this study is from National Longitudinal Survey of Youth 1997, which has information on about 8,900 individuals aged 12-17 in the year 1997 who are followed up every year since. I use the first 10 rounds of this survey.

The first essay in this dissertation identifies the influence of older siblings on their younger siblings' decision to leave school before graduation. It starts by representing the significant adverse effect of birth order on school completion outcome of the teen. The results indicate that teens with older siblings are more at risk of dropping out of school. Having these initial estimates of the siblings' effect, I look at the behavior of the siblings pairs in the in a subset of 519 pairs from families with two children for which the data is available on both siblings' outcome and characteristics. I address the question of the inherent endogeneity by using family fixed effects models on these pairs, and also by implementing instrumental variables technique. I exploit the older sibling's specific characteristics as an instrumental variable for his outcome in the younger sibling's equation. The sibling specific characteristic used as instruments are: the

older sibling’s gender, the unemployment rate when the older sibling was 16 years old, the mother’s age at his birth, and the intactness of his family when he was 14 years old. The results of the estimation using this set of instruments show positive and significant effect on early school leaving of the younger sisters.

The second essay utilizes a survival model to determine the timing pattern of teenager’s decision to drop out. Preventing and intervening in school dropping out in teens require the knowledge of the timing and pattern of occurrence of this act. I use nonparametric, semiparametric and parametric hazard models to explain the factors affecting the age onset of high school dropping out. Alongside with other socioeconomic factors, this chapter reinforces the roll of siblings on reducing the starting age of this risky behavior in the younger siblings. The results show that a teenager who has an older sibling is more likely to stop schooling at a younger age. In teenage boys who have an older sibling the hazard of dropping out is about 3 times as much depending on the choice of the hazard model for the time to first dropout. The shared frailty among siblings of the same family that contaminates the estimates is measured for the parametric and semiparametric duration models using the expected-maximization algorithm for the clustered data. The impact of shared frailty is ruled out in the case of Cox semiparametric models, but had been evident in some of the parametric models used in this chapter.

The third essay develops a binary choice interaction model with finite number of agents to characterize the peer effect of siblings on the strategic choices of the teenager. This dynamic model incorporates the attractiveness of imitating the behavior of the peers inside the family. The model measures the strategic complementarity between the choice of the teenager and the choice

of his or her siblings, after controlling for shared family fixed effects and utilizing the lagged dependent variable to reduce the unobserved contextual and correlated effect. This model finds significant social interaction effects between siblings, and specifically siblings that are closer in age to one another.

The broad policy implication of this research is to indicate another important channel through which the strategic planning to reduce school dropout rate could be directed. School dropout prevention programs can put more emphasis on the first order or lower order children in multi-kid families to utilize the existing spillover effect on the younger siblings. Also considering this spillover effect, parents' investment in raising a more scholarly firstborn might help them get an additional indirect return to their investment benefiting the other kids in the family.

To My Parents:
Farzaneh Sabri & Alireza Farivar

Acknowledgements

I wish to extend my deepest gratitude to my adviser, Hadi Salehi Esfahani, not only for helping me in my research, but also for being my mentor during my graduate studies at University of Illinois. His profound knowledge, his intuitive insights, his optimism and positive attitude, and his warm and supportive personality are indeed the best combination any PhD student can wish for. I will always keep in mind what he once said to me, that *research* means to search and yet search again, thus not getting the results one seeks is a requisite of this process.

My special thanks to Anil Bera, for his valuable comments and feedback on my research, and also for showing me how to be a good academic teacher. Being his teaching assistant for three years taught me many valuable lessons on how to make rigid statistics formulations pleasantly interesting for students. I would like to sincerely thank my other dissertation committee members Werner Baer and Kristine Brown for reading many drafts of my papers, sometimes in such short notice, and providing me with their valuable comments and guidance.

I could not have accomplished this project without the help and support of my family. I am truly grateful to my parents who have invested so fully and sacrificed so much in providing the best possible education for me. For this and for all their love I am forever thankful. To my own sibling, my brother who has always been there with his arms extended when the times got hard on me, many thanks. And to my fiancé, for his love, patience, and support; thanks for being the audience, the discussant, and the advocate of my research ideas.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Adolescence years are of great importance in a person's life. The adolescence is the passage through childhood into the adulthood that involves personal and social transitions in different aspects of a person's character. The biological, cognitive and social changes during the adolescent years allow the teenager to develop the identity that will serve as a basis for his or her adults life. It is during this time that the adolescent establishes the sense of autonomy and independence and moves away from the childhood dependencies on his parents.

As the adolescent become more independent in making more key decisions about his life, his propensity to engage in risky behavior increase. One explanation for this is the fact that the adolescent is a rather new and inexperienced in decision making, and thus has less knowledge and expertise compared to his parents who to this age were in charge of making the key decisions for him. Also the adolescents have more myopic preferences compared to matured adults, which makes them assign higher discount rates to the futuristic payoffs, and higher values to the present pleasure (Gruber, 2000). The fact the many aspects of juvenile risky behavior and delinquent activities encompass immediate utility and future costs, puts the adolescents largely susceptible to risky behavior.

It is evident that the pattern of individual's involvement in risky behavior is increasing in age, but diminishing as the age passes certain thresholds (Laing, 2009). The adolescent years are when risky behavior such as alcohol and substance abuse, criminal activities, delinquencies such as running away from home, and dropping out of school accelerate.

As the adolescent goes through these transitions, the influence of parents and peers begin to interchange. During childhood, the kids are highly oriented towards their parents, but as they reach the age of adolescence, gradually the conformity towards parents reduces and the adolescent begins conforming more to his peers. Numerous studies have been conducted on peer effects and peer pressure of adolescents. The association between the behavior of the adolescent and his peers. Studies show that the susceptibility to peer influence is at its maximum during early teenage years (Steinberg, 1996).

One aspect of peer influence on the behavior and outcome of the teen, that is often overlooked is the impact of teenaged siblings. Siblings are the most immediate peer group that a teenager is in contact with. In the families with two of more teenaged kids, the cross effect between the teenaged siblings is substantial, as the adolescent is in direct and continues contact with his brothers and sisters. Sibling correlations in delinquent behaviors are often found to be larger than any of the correlations between other peers defined as adolescents' best-friends, between schoolmates living in the same neighborhood, and between pupils in the same grade within a school (Duncan et al. 2001). Despite of the importance of this source of peer effects, there are fewer studies that have looked at the impact and mechanism of this within the household peer effect.

In this dissertation I will focus my attention to the peer effect that comes through a teenager's siblings, or as it can be called the siblings peer effect. The siblings peer effect takes place within the family so it is hardly possible for the teen to avoid. This effect is also relatively more intense as siblings share a lot of common characteristics and background that makes them more receptive to the influence of one other.

Among the different types of risky and delinquent behavior, this dissertation is concerned with adolescents' dropping out of school. Dropping out of high school is a dangerous behavior for the adolescent, with clear economic and personal cost to them as well as social welfare cost to the economy. There has been extensive research interest in teens' dropout phenomenon in the fields of economics, education, and developmental psychology. These studies generally look into the factors that influence a teenager's decision to leave his high school education incomplete.

High school dropout is a widespread and serious problem in the United States, with enormous consequences for students who choose its path. The status dropout rates[1] of young people aged 16 to 24 in the civilian, non-institutionalized population gradually declined between 1980 and 2007, from about 14 percent to a low of 8.7 percent in 2007. Even though the rates are declining,

[1] http://nces.ed.gov/fastfacts/display.asp?id=16

they still represent a large number of people. In the 2005-06 school year, only 73.4% of high school students who were enrolled in public schools graduated with their class. School dropout can have three detrimental consequences: First, the adolescents who drop out of school are more prone to other types of delinquencies and criminal behavior. Secondly, under the current condition of the U.S. economy which is deviating away from the low skilled jobs, and is moving towards the direction of creation of more professional and high skilled job opportunities that require more education, high school dropouts are faced with higher unemployment and lower earnings. And thirdly, dropouts will incur the economy with more social and welfare costs in the long run.

Although the juvenile peer effect, and the phenomenon of teenage dropouts has been explored independently in many studies, fewer researches have examined the links between the two. This study is a bridge between these two bodies of literature, to shed light on the size, timing, and dynamics of siblings peer effect of the school dropout behavior of the adolescents. In my evaluation, I place emphasis on the time pattern of the dropout event and test the hypothesis of reduced age of dropout in teens who are exposed to negative siblings' activities and outcome.   I find that after controlling for individual and family characteristics the adolescents who have one or more dropout older siblings in their families are significantly more likely to become dropouts themselves. More adversely, this dropout in the younger sibling is expected at a younger age.

The data used in the empirical models of this study comes from the National Longitudinal Survey of Youth 1997 (NLSY97) . This dataset is a nationally representative sample of 8984 teenagers who where 12 to 16 years of age at the year 1997. The NLSY97 follows these teens as they age and make the transition from school into the labor force. The rather long time span of NLSY97 enables me to follow the dropouts year by year and keep an ample record of their switch between school enrollment, graduation, and dropout. The design of the NLSY97 is such that it includes all the teenagers in a given sample household who are age eligible to be surveyed. Therefore for any teen included in the survey, there is the complete record of his or her teenaged2 siblings who reside in the same household. I utilize this setup to serve as an age window for the peer like effect within the household. The reason is, much older siblings, even though regarded as role models to the teen, can hardly fit into the "peer" definition. And as for much younger siblings, despite they divert part of parents' supervision and resources, they can barely be considered as an influence on the teen's behavior and decisions. Therefore the maximum four year age window for the participating siblings in the NLSY97 is the criteria upon which I have defined the sibling peer effect in this study. I have considered the age closeness as the intensifier to the siblings peer effect, with the twin having the strongest peer influence.

This dissertation consists of three essays on siblings effect and high school dropout. The first essay (chapter 2) addresses the birth order effect, and estimates the impact of older siblings on the younger sibling's school completion decision. The models in this chapter offer an evidence for the birth order effect. The later-born teens in the family are more likely to be dropouts compared to their earlier-born siblings. The analysis incorporates the use of instrumental variables to identify the effect of older dropout siblings on the dropout event of the younger

[2] Defined, in accordance to NLSY97 survey design, by being born between 1980-1984; thus being 12 to 16 year olds at the year 1997.

siblings. The older siblings characteristics that are not shared between the pair, such as mother’s age at the birth of the older sibling, the unemployment rate when the older sibling was 16, and family intactness when the older sibling was 14 are used as instruments. The policy implication of these findings is that parental investments of the education of their earlier-born children could have spillover effects on the school completion success of their younger children. Also outreach programs that target first-borns in the bigger families could be effective in reducing the teen school dropouts in other kids in the household.

The second essay (chapter 3) utilizes survival models to determine the timing pattern of teenager's decision to drop out of school. I use non-parametric, semi-parametric and parametric hazard models to shed light on some of the key factors affecting the onset of school drop-out among teens. The results suggest that teens who have a dropout older sibling are more likely to dropout at a younger age. After controlling for the family size, and other determinant factors, I find the presence of an older sibling increased the hazard of dropout by 16% in the adolescent respondents of NLSY-1997. This impact when originated from the older sibling that are closer in age to the adolescent is estimated to be 18% for the males and 16% for the females; about three times as much as the effect of much older siblings for both genders. The Cox semi-parametric, and five different fully parametric hazard models all estimate large significant effect for the adverse impact of the presence of a dropout in the family on the younger kids. The finding of this essay show little evidence of the distortion by the shared frailty among siblings in the Cox proportional hazard model. In the parametric models, depending on the choice of the survival time and the cluster error term, the impact of this shared frailty was measurable decrease in the estimated impact of the dropout siblings.

The third essay (chapter 4) builds a theoretic discrete choice model to represent the dynamics of social interaction among teens within a family. The model takes into the account the payoff that a teenager gets from conforming to his siblings behavior, and at the same time, the disutility of choosing the opposite path. This payoff, captured in the model by a social utility term is found to be significant and strong in a sample of households of NLSY97 that have 1 to 4 teenaged siblings. The social interaction is stronger among the sub-sample of families with more than 2 teenaged kids. To justify the existence of siblings' social interaction, the model is purged out of the unobserved contextual and correlated effects. The use of lagged social interaction index reduces the contextual effect. I use family fixed effects to eliminate the correlated effects. The use of lagged siblings outcome and household fixed effects model isolate the targeted peer effects from other bias causing effects. The final model predicts that the effect of weighted average of dropout siblings in a family is an increased odds of dropout vs. graduation by a factor of 1.68, equivalent to 0.24 increase in probability.

Chapter 2

Siblings Effect and School Dropping Out in Teens: Do the Younger Ones Follow the Older Ones?

2.1 Introduction

Adolescence years are crucial times in the life of a person. Many instances of risky behavior such as substance abuse, criminal activities, and delinquencies are start at a young age. Adolescence years are the time when many young individuals start to stray away from the authority and supervision of their parents, and begin making decisions about their own affairs. One of these decisions that has proven to be significantly critical is adolescent's decision to stay in high school and graduate with his class or to drop out.

High school dropout is a widespread and serious problem in the United States, with enormous consequences for students who choose its path. The status dropout rates[3]of young people aged 16 to 24 in the civilian, non-institutionalized population gradually declined between 1980 and 2007, from about 14 percent to a low of 8.7 percent in 2007. Even though the rates are declining, they still represent a large number of people. In the 2005-06 school year, only 73.4% of high school students in enrolled in public schools graduated with their class. The labor force participation

[3] http://nces.ed.gov/fastfacts/display.asp?id=16

rate of dropouts is lower than the high school graduates. In addition, dropouts have higher unemployment rate. In year 2009, dropouts had an unemployment rate of 55.1 percent compared to 28.4 percent for the workers with high school or General Educational Development diploma4.

Dropout is a highly risky decision. Insufficient schooling can have drastic economic consequences for the individual as well as the society. Dropouts are more likely to be unemployed, have lower earnings, engage in criminal activities, be incarcerated, and have poor health. They are more likely to be single parents, on social welfare program, and raise children with lower educational achievements. Dropouts have lower life time earnings. For example, the median income of dropouts aged 18 through 65 was roughly $24,000 in 2007.1 By comparison, the median income of persons ages 18 through 65 who completed their education with a high school credential, or a GED was approximately $40,0005.

Numerous studies in the field of economics, education, and psychology have investigated the determinant factors of the decision of a teenager to drop out of school. Among these determinants, this chapter places its focus on the peer influence. And in particular the type of peer influence effect that acts through the channel of the adolescent's family, specifically through his or her siblings. The question I pose here is whether coming from a household with one or more older siblings makes it more likely for an adolescent to drop out of school before graduation.

[4] http://www.bls.gov/news.release/hsgec.t01.htm

[5] 1 U.S. Department of Commerce, Census Bureau, Current Population Survey (CPS), March 2008.

Although this specific type of peer influence coming from older siblings can be placed under the conventional peer group effect, it could be of more because of the following reasons: (1) This influence takes place within the family so it is hardly possible for the adolescent to avoid it. A typical adolescent lives his everyday life within the family and is surrounded by his siblings, if he or she has any. Because of family binds it is less likely that parents monitor or restrict the exposure of the adolescent to their troublesome older children, as opposed to the parents' attempts to restrict their child's contact with his troublesome friends. There is not much exposure control mechanism that the adolescent's parents can use in this context. (2) Siblings share a lot of common tastes and tendencies and are usually exposed to the same family background and environment. So the transmission of a behavior from one sibling to another is not faced with the resistance that would normally arise due to heterogeneous background among peers. (3) Role modeling is stronger coming from within the family rather than from outside of the family. A kid spends most of his childhood and early teens being exposed to the behavior of his older sibling, thus those siblings are more likely to be picked by the younger kid as somebody to look up to and imitate.

This study uses the NLSY (National Longitudinal Survey of Youth) panel data in the years 1997-2006 to estimate the effect of older siblings on the likelihood of an adolescent dropping out of school. NLSY is a panel dataset consisting of a broad range of questions on the behavior of teenagers and young adults. The previous studies done on the siblings effect have not explored the whole time span of the NLSY-97. In this study I use the 10 rounds of the survey published at present time. In this chapter I seek to interpret the association between sibling structure and the schooling outcome of the adolescent respondents of NLSY utilizing two different identification strategies. I make an attempt to reduce the unobserved heterogeneity between the outcome of the siblings through models of fixed effects and instrumental variables.

The remainder of this chapter proceeds as follows. Section 2.2 reviews the literature on the birth order effect on likelihood of early school leaving in teenagers, as well as the impact of having a dropout older sibling in the family. Section 2.3 describes the data, and section 2.4 lays out the economics models and reports estimates of sibling effects used and their results, while section 2.5 concludes.

2.2 Literature Review

The studies done on the subject of siblings effects can be divided into two parts. First the literature concerning the birth order defined as the position of the teenager in the age hierarchy of siblings in the family, and the allocation of parental resources among the siblings based on their characteristics and outcome. The second group of literature deals with the association among the behavior and educational outcome of siblings in a household.

Parents who have more than one child are faced with the decision of how to allocate their resources among their children. Besides other determinant factors, this decision is also influenced by the birth order of the kids. In the literature models of intra-household allocation try to describe the decision making process of parents when they have more than one child to invest in. Earlier research on this topic goes back to the pioneer work of *Becker (1960), Becker and* Lewis (1973), and Becker and Tomas (1976). They suggest the resources available to parents are allocated optimally to level the tradeoff between the quantity (number of siblings) and quality (educational attainment) of their children. According to these models the families that have more children can invest less on their quality, for instance their education, because they have invested more in quantity. Parental investment in children's education is likely to be dependent on number of siblings in the family, and the sibling's corresponding cognitive endowment. Behrman, Pollak and Taubman (1995) find that parents invest on each child up to the point that the marginal product of investment in that child equals the market rate of return. This competition among children for the parental resources, gives the older kids in the family an advantage as they have less rivals in their younger stages of life. This advantage could also be present for a middle-born child whose immediate younger sibling is born some many years after. Because the long spread of time between him and his much younger immediate sibling, opens up the opportunity for him to absorb more of the parental investments.

Another part of this literature is concerned with understanding the effect that siblings have on one another's life-course outcomes. Haurin and Mott (1990) present a theoretical approach to sibling models in the framework of social comparison theory. This theory suggests that individuals adjust their behaviors and attitudes to conform to others who occupy similar social positions or share similar attributes relevant to a particular behavior or attitude. The family provides a context for this social exchange among siblings. Older adolescents are often admired and emulated by younger adolescents because of the greater freedom, privileges, resources, and experience they enjoy. In the family context, older siblings are major role models for younger siblings.

Not only the number of siblings in a person has, but also their configuration could have an impact on many individual outcomes such as wages, labor force participation, and propensity to engage in certain behaviors. Kessler (1991) studies the effect of sibling configuration on determining the future adult wage. He finds that women coming from small families work less than women from large families when they are young, and more that women from large families when they are mature. Butcher and Case (1994) in an study using ??? data, find that woman who are raised with more brothers have higher educational attainment. However some later work by Kastner (1997), and Hause and Cho (1998) do not find this effect significant. On the contrary, Conley (2000) using PSID data finds that having more siblings of the opposite sex reduces one's years of schooling, compared to an individual with more siblings of the same sex. Black et al. (2005) using a dataset on the entire population of Norway, find robust effect of birth order on education. They also find that later-born woman have less earnings and are more likely to have their first birth in their adolescence. However they don't find meaningful effect of the family size on education.

Agrys et al (2006) uses the NLSY-97 to investigate the association between birth order and adolescent behaviors such as smoking, drinking, marijuana use, sexual activity, and crime. Its estimates show that middle-borns and last-borns are much more likely to use substances and be sexually active than their firstborn counterparts. Rose (2006) uses the NLSY-79 and for two cohorts representing the early and later phases of the all-volunteer military enlistment era. This chapter finds that for the early cohort, the likelihood of enlistment decreases with birth order; but there is no significant effect of family size on enlistment for the later.

Gary-Boro et al (2006) use twin birth as an investment for family size in a study on French data and find out that the sex composition of the siblings is important in their educational attainment. They found that on average females with more brothers have lower educations and consequently less earnings compared to males.

The second group of studies highlights the finding that the behavior and choices of the older sibling may have direct influence on the behavior of the younger sibling. There are several explanations for this older to younger transmission.

One explanation is the tendency of younger kids to seek for role models among the people they interact with. And older siblings are often the first available candidates for role models. The other explanation is that if a kid’s sibling is already involved in a risky behavior, the issue of early exposure to that behavior is more important. In the case of school dropout, it is evident in the national data, as well as the data in NLSY that dropout probability increases by age. Therefore an older sibling, who drops out sometime within the common age window, will expose his or her younger siblings to the behavior of dropping out at a much younger age. However it is necessary to control for shared genes and family environment in order to identify these effects.

Ouyang (2004) employs a fixed effect model to difference out these shared characteristics, which are mostly time-invariant. He claims that younger siblings are more likely to adopt smoking at a younger age because of their older siblings' effluence. Another study by Rodgers et al (1992), using data from National Longitudinal Survey of Youth –1979 examines the effect of birth order and siblings sex composition on age at first intercourse. They found that younger siblings tend to have first intercourse earlier than older siblings.

2.3 Data

The models in this chapter use the data from National Longitudinal Survey of Youth-1997 (NLSY97). This data is a national representative sample of 8984 respondents who were between ages of 12 and 16 in year 1997, with an over sampling of racial minorities. The data has been collected annually and to this date eleven rounds of it is released. I will employ ten rounds of this survey covering the period 1997-2006. After the initial round, no new individuals are added to the dataset, therefore the lower and upper bound of the respondent's age increases by one in each round. Retention rate[6] on average is above 80% during these 10 rounds.

[6] Retention Rate is defined as the percentage of base year respondents remaining eligible who participated in given survey year; deceased respondents are included in the calculations. Reason for not participating (non-interview) includes being deceased, not locatable, technical problem, respondents too ill, respondent unavailable, refused to interview, or other. Among these the refusal to interview and being non-locatable are respectively the major reasons.

Table 2.1. Retention rate of the NLSY-97 survey, rounds 1-10

Round	Year	Sample size	Retention Rate
1	1997	8984	-
2	1998	8386	93.3
3	1999	8208	91.4
4	2000	8080	89.9
5	2001	7882	87.7
6	2002	7896	87.9
7	2003	7754	86.3
8	2004	7502	83.5
9	2005	7338	81.7
10	2006	7579	84.1

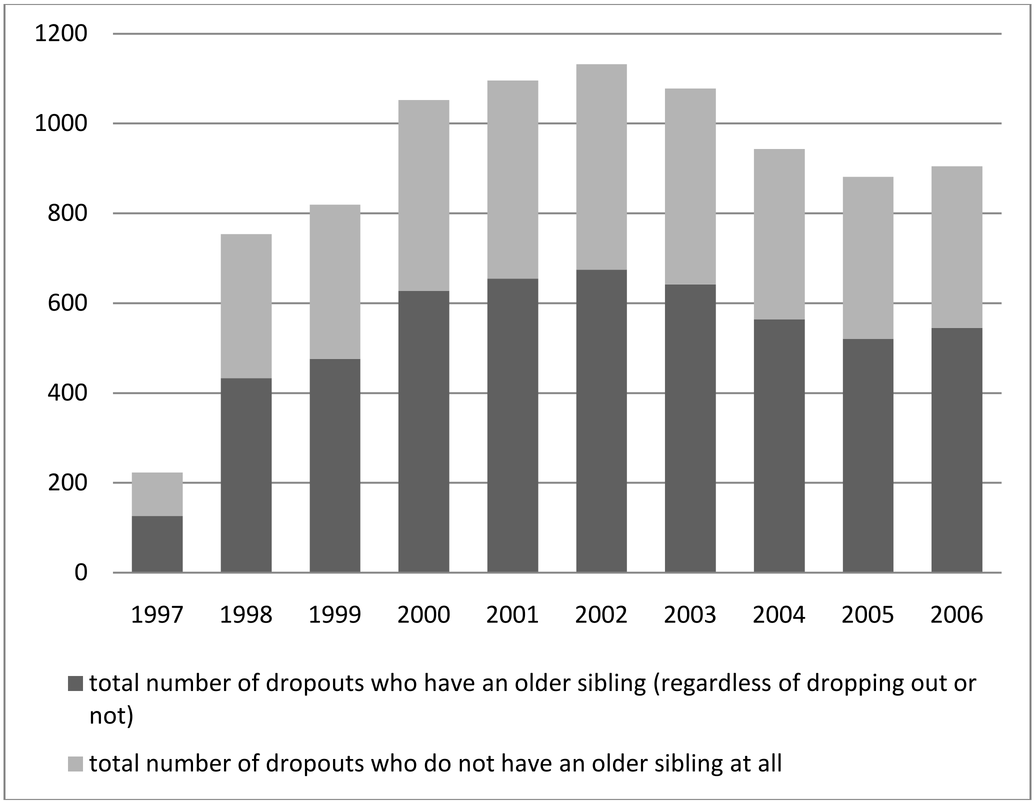
The dependent variables in this chapter are dropping out of school at the time interview each year. Dropping out of school in this chapter is defined as leaving the school before completion and getting any type of completion diploma. The variable “dropout” in this chapter is constructed using the questions asked about the enrolment status of the individual in school at the time of the survey. If the respondent’s status is “Not enrolled”, and have “No high school degree, no GED” he is marked as a dropout at that year. Respondents who are working towards a GED are coded as being enrolled regardless of where that course of study took place, and thus are not considered to be dropout.

Leaving school early could even be illegal depending on the compulsory school laws in different states. The compulsory attendance law[7] requires parents to have their children enrolled in a public or state accredited private or parochial school for a designated period. In the United States, the compulsory education varies by state, beginning at ages five to eight and ending at the

[7] With the exception case of Home-schooling

ages of sixteen to eighteen[8]. A growing number of states have now implemented compulsory education laws that require schooling until the age of 18. Figure 2.1 shows the number of dropouts in each year of the survey. As mentioned, the respondents are all between 12 and 16 years of age in the initial year of the survey. In year 1998 the first wave of respondents passes the age threshold of 16, which is the compulsory age in some states, and thus we see a hike in the number of dropouts.

Figure 2.1. Number of dropouts, with or without older siblings



[8] *Source:* Department of Education, National Center for Educational Statistics, *Digest of Education Statistics, 2004.*

Table 2.2. Dropout rates among the youth and older siblings in NLSY-97

	Among youth of NLSY	Among surveyed older siblings of NLSY
1997	2.5	4.2
1998	7.2	13.2
1999	9.7	14.7
2000	12.6	16.8
2001	13.7	15.8
2002	14.1	15.8
2003	13.9	14.9
2004	12.7	13.2
2005	12.1	12.7
2006	12.0	12.7

In this chapter, the focus is on the effect of the older siblings on school completion decision of the adolescent. NLSY provides the complete roster of the family members of each respondent. The roster includes the relationship, gender, and age of each family member, no matter if he or she resides in the household at the time of the survey, or has moved out. Up to four of these household members recorded in the roster, who qualified for the age criteria of the initial round, i.e. aged 12 to 16 years old in the beginning of the year 1997, were also surveyed. Otherwise the information regarding them is quite limited. For example consider a household with 5 children aged 10, 13, 15, and 19 as of December 31st, 1997. And a 23 year old who has moved out of the household. In this case the 13 and 15 year olds will be surveyed and followed up every year after. But we will not have any information on the characteristics and behavior of the resident 19 year old, or the nonresident 23 year old. This is a shortcoming if specifically the 19 year old is the bad influence dropout kid in the family. However both 19 and 23 year old children will be considered in the counting of older siblings. Table 2.2 shows the dropout rates among the youth of NLYS and their surveyed older siblings in years 1997-2005.

A set of demographic and economic controls are used in the models of this chapter as the covariates. I take the sibship variable constant during the time span of the dataset. This is not a strong assumption as this chapter emphasizes on the information of the older siblings of a respondent, and the sibling configuration of the older siblings does not change much in reality. In addition both resident (in the household) and non-resident siblings are accounted for in generating the sibship variables, so the possibility of an older sibling leaving the household as time goes by, will not change the constructed sibship variables. The violation to this assumption would happen if an older sibling of a respondent dies during this period, or the respondent's family adopts an older kid; both incidences are rare in the data.

Table 2.3 gives the summary statistics of the dependent covariates used in the models. The time-invariant controls are in top panel, and the time-variant controls are in the bottom panel of the table. Apart from demographic variables, there are some controls over characteristics of the parents, such as their highest level of education, the age of mother at the birth of the respondent, and whether or not the teen has lived with both parents till 14 years of age, which captures the intactness of the family. NLSY-97 asks respondents about the degree of monitoring imposed on them by their parents. This categorical variable consists of 4 areas of parental control over the teenager's friends, the families of the teen's friends, school work and teachers, and the parents' inquiry about the whereabouts of the teenager at all times.

The variables used to capture the effect of family income are 3 dummy variables: whether the household income is below 125% of the federal poverty line (poor HH), whether it is over 400%

of the federal poverty line (rich HH), and if there is no household income recorded. The middle income households are the reference case. The data on household income is inquired on each round of the survey. For the cases where there were holes in the stream of reported household income, I filled the gaps with the latest reported income up to that point in time. For example if a household has a reported income of $100,000 for year 1997, missing the data on year 1998, and $120,000 on year 1999, I approximated the family income equal to $100,000 in year 1998.

I use a set of variables to capture the configuration of the sibship in the family. These include the number of resident and nonresident older siblings, the number of older siblings that are within 3 years of age difference with the teenager, and the number of older siblings that have more than 3 years of age difference with the teenager. The summary statistic for the sibship variables is presented in Table 3.2. The survey includes data on up to 4 other teenaged kids in the family of the respondent household. We can then link the characteristics and the behavior of these 4 siblings (given they are present, and older) to the outcome of the primary teenager.

Table 2.3. Summary statistic of the covariates used in the models

	Obs	Mean	s.d.	Min	Max
I. Individual characteristics					
Sex	8984	0.488	0.500	0	1
Black	8984	0.266	0.442	0	1
Hispanic	8984	0.212	0.409	0	1
Other race	8984	0.152	0.359	0	1
Age	8984	18.807	3.229	12	27
II. Family characteristics (shared)					
Father's education (grade level)	7120	12.564	3.212	1	20
Mother's education (grade level)	8290	12.438	2.913	1	20
Low income household	89840	0.276	0.447	0	1
High income household	89840	0.239	0.427	0	1
Missing household income	89840	0.110	0.313	0	1
Urban residence	89840	0.665	0.472	0	1
Household size	79362	3.900	1.727	1	17
One child family	8984	0.091	0.287	0	1
Two child family	8984	0.303	0.459	0	1
Three child family	8984	0.280	0.449	0	1
Four child family	8984	0.166	0.372	0	1
Five plus child family	8984	0.160	0.367	0	1
III. Family characteristics (individual)					
Mother's age at the birth of the respondent	8374	25.482	5.413	10	54
Intact family at 14	7935	0.727	0.445	0	1
Father's monitoring degree	13936	7.640	3.960	0	16
Mother's monitoring degree	18657	9.892	3.317	0	16
IV. Sibship characteristics					
Total siblings	8984	2.176	1.650	0	18
Older siblings	8984	0.946	1.211	0	11
Older brothers	8984	0.480	0.784	0	7
Older sisters	8984	0.466	0.792	0	7
Older siblings more than 3yrs age difference	8984	0.556	1.043	0	11
Older siblings within 3yrs age difference	8984	0.390	0.582	0	5
Older brothers more than 3yrs age difference	8984	0.279	0.642	0	7
Older brothers within 3yrs age difference	8984	0.201	0.432	0	3
Older sisters more than 3yrs age difference	8984	0.277	0.652	0	7
Older sisters within 3yrs age difference	8984	0.189	0.427	0	4
Older siblings that are also in NLSY survey	8984	0.672	0.672	0	4

2.4 Model and Empirical Results

The models in this chapter are divided into five parts: The first set of models assess the Birth Order effect by looking at the hierarchy of the children in a family and its association with their outcome. The second set of models approach the exposure factor that the younger children in a family face when their older siblings get involved in risky behavior. The other three models utilize the data on the pairs of siblings to examine the direct effect of the choices of the siblings on the dropout decision of the younger sibling.

2.4.1 Birth Order Effect

This first set of models in this chapter intends to capture the effect of the birth order on the school completion outcome of the individual. In these models, I want to see whether having one or more older siblings makes it more likely for an individual to drop out of school. In other words the question this model tends to address is whether the second-born or higher order born kids in a family do worse in their school completion outcome compared to their first-born or lower order born siblings. This model do not take into consideration the outcome of the older siblings, and in this sense, it is not seeking to show a causal effect but rather test the widespread belief that the birth order is a determinant of personal or economical success.

This model incorporates the configuration of the older siblings as independent variables in the equation for the outcome of the younger sibling. Consider the following model:

(2.4.1)

$$Y_{it}^{*} = X_{it}\alpha + Z_{it}\beta + O_{i}\gamma + \varepsilon_{it}$$

Where Y* represents the dropout decision of the individual i at the time t, X is a vector of observable individual characteristics, Z a vector of observable family specific characteristics, and O is a vector capturing the configuration of the adolescent's older siblings. Y* is a latent variable corresponding to the dummy variable Y which is the observed decision of the adolescent to either dropout or stay in the high school at time t. Given the assumption that ε is normally distributed[9], Y can be viewed as an indicator for whether Y* is positive. The equation 2.4.1 shows the latent variable model of school dropout decision in the form of a standard univariate probability model:

(2.4.2)

$$\Pr(Y_{it} = 1 \mid X_{it}, Z_{it}, O_i) = \Pr(X_{it}\alpha + Z_{it}\beta + O_i\gamma + \varepsilon_{it} > 0)$$

A positive estimate for the coefficient γ in (2.4.2) indicates that the presence of older siblings in the household will make it more likely for the teenager to drop out of school.

[9] In sections 4.4 and 4.5 I will address the violation of this assumption, along with the strategies to estimate the siblings effect, when the error is not independently normally distributed.

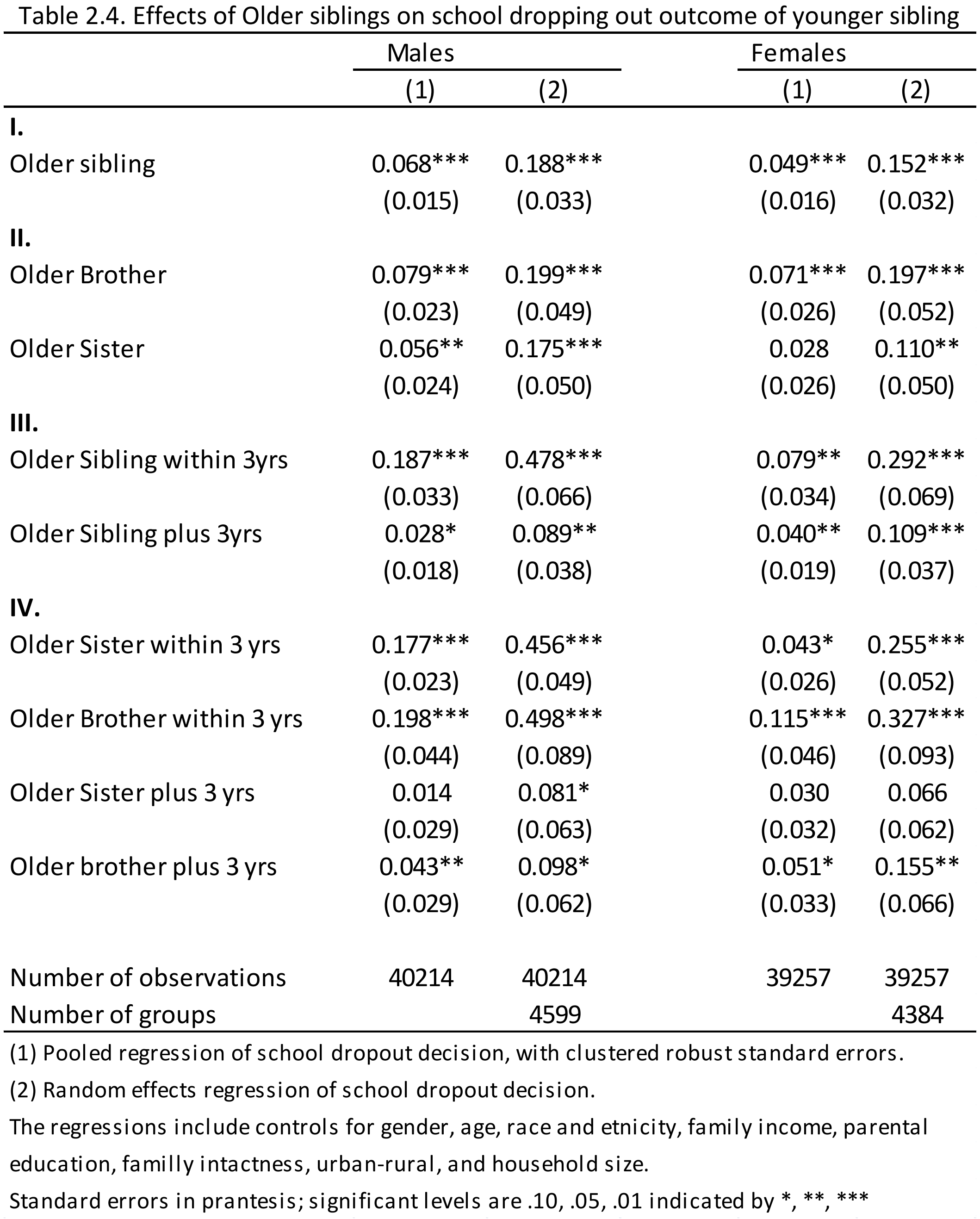


Table 2.4. Effects of Older siblings on school dropping out outcome of younger sibling

	Males		Females	
	(1)	(2)	(1)	(2)
I.				
Older sibling	0.068***	0.188***	0.049***	0.152***
	(0.015)	(0.033)	(0.016)	(0.032)
II.				
Older Brother	0.079***	0.199***	0.071***	0.197***
	(0.023)	(0.049)	(0.026)	(0.052)
Older Sister	0.056**	0.175***	0.028	0.110**
	(0.024)	(0.050)	(0.026)	(0.050)
III.				
Older Sibling within 3yrs	0.187***	0.478***	0.079**	0.292***
	(0.033)	(0.066)	(0.034)	(0.069)
Older Sibling plus 3yrs	0.028*	0.089**	0.040**	0.109***
	(0.018)	(0.038)	(0.019)	(0.037)
IV.				
Older Sister within 3 yrs	0.177***	0.456***	0.043*	0.255***
	(0.023)	(0.049)	(0.026)	(0.052)
Older Brother within 3 yrs	0.198***	0.498***	0.115***	0.327***
	(0.044)	(0.089)	(0.046)	(0.093)
Older Sister plus 3 yrs	0.014	0.081*	0.030	0.066
	(0.029)	(0.063)	(0.032)	(0.062)
Older brother plus 3 yrs	0.043**	0.098*	0.051*	0.155**
	(0.029)	(0.062)	(0.033)	(0.066)
Number of observations	40214	40214	39257	39257
Number of groups		4599		4384

(1) Pooled regression of school dropout decision, with clustered robust standard errors.

(2) Random effects regression of school dropout decision.

The regressions include controls for gender, age, race and etnicity, family income, parental education, familly intactness, urban-rural, and household size.

Standard errors in prantesis; significant levels are .10, .05, .01 indicated by *, **, ***

Table 2.4 reports the results of estimation of the sibship coefficient γ for different configurations of older siblings, separately estimated for teenage males and females. Among the controls used but not reported in table 2.4, are two measures of household income compared to the poverty line, number of people living in the household, and the degree of parental monitoring on the activities of the adolescent. Other controls used in the regressions are age and race dummies, parents' education, a measure of family intactness, mother's age at the birth of the teenage, and year fixed effects.

The model (I) captures the effect of being a second of higher order born on the probability of dropping out of school. The estimates of gamma in the random effects panel regression suggest that the boys who have an older sibling have about 19% more chance of being a dropout. The teenage girls with older siblings have a 15% higher probability of dropping out. Model (II) adds the gender impact to the birth order model. The data suggests that having an older brother in the family increases the chance of dropout for both genders. Having an older sister has significant but not as large an effect. Model (III) encloses evidence that the older siblings closer in age to the adolescent have a much larger impact than the much older siblings. A middle or last born teenage boy has 48% more chance of dropout if he has a close-in-age older sibling. Similar teenage girl is faced with 29% more probability of a dropout. And finally model (IV) combines the effects of age and gender of the older siblings. The effect of a close-in-age older brother or sister on the probability of a teenage boy's school dropout is 50% and 46% respectively. The same effect for a teenage girl is 32% and 26% respectively. Overall, all the univariate probability models in table 2.4 associate a higher probability of dropout to the middle-born and last-born adolescents in a family.

Another extension to the model is to investigate the effect of older siblings in different family sizes. I divide the household in the data into two-kid, three-kid, four-kid, and five or more kid families and estimated the effect of being a middle or last born child in a family of a certain size.

Table 2.5. Effect of birth order in various family sizes on dropout outcome of later-borns

	Two-kid families			Three-kid families		
	All	Males	Females	All	Males	Females
Second-born	0.225***	0.408***	-0.036	0.117*	0.063	0.191*
	(0.051)	(0.066)	(0.081)	(0.056)	(0.079)	(0.082)
Third-born				0.341***	0.449***	0.230*
				(0.061)	(0.083)	(0.094)
Observations	23,485	11,893	11,592	22,293	11,387	10,760

	Four-kid families			Five and more-kid families		
	All	Males	Females	All	Males	Females
Second-born	-0.004	0.165	-0.100	0.212**	0.202	0.311**
	(0.068)	(0.092)	(0.108)	(0.076)	(0.104)	(0.115)
Third-born	-0.122	-0.196*	0.029	0.183*	0.049	0.379**
	(0.075)	(0.099)	(0.118)	(0.081)	(0.113)	(0.118)
Fourth-born	0.062	0.024	0.111	0.327***	0.129	0.509***
	(0.086)	(0.115)	(0.135)	(0.081)	(0.112)	(0.119)
Fifth-born				0.347***	0.295**	0.441***
				(0.073)	(0.102)	(0.108)
Observations	13,402	7,103	6,179	12,686	6,056	6,630

The results are derived from pooled logit regression models with cluster standard errors.
Standard errors in prantesis; significant levels are .10, .05, .01 indicated by *, **, ***
The regressions include controls for gender, age, race and etnicity, family income, parental education, famillly intactness, urban-rural, and household size.

Table 2.5 reports the results of estimation of dropout decision of the teenagers in different family size and birth order specifications. The estimates are from a pooled logistic regression model with clustered standard errors. The same set of covariates in table 2.4 are also included in the regression model and reported in table A.1 of the appendix. The estimated results from these regressions, although not significant in some case, reiterate the large effects of being later-born on the increased risk of school dropout. In almost all family sizes, the birth order effect of dropping out increases monotonically for the later-borns.

2.4.2 The Exposure Factor

The models discussed in this section, the existence of a birth order effect is evident. Although it is plausible to think this effect is in partly caused by the older siblings, without taking the older siblings' behavior into account, this claim would not have a strong foundation. The model in this section opens the possibility to include the older siblings' behavior into the picture. The conventional way of incorporating the information on the outcome of the older sibling to that of his younger one, is through models of siblings pairs, which are vastly used in labor and household studies. I will adopt this methodology in the sections 2.4.4, and 2.4.5. Using the pair layout in a three or more kid families will often entail the loss of information on the behavior of other siblings who are not included in the primary pairing of the data. In this section, I will define a factor that tries to capture the collective behavior of the older siblings

While conducting the initial rounds of interviews, NLSY interviewed the siblings of any respondent of the survey, as long as those siblings had the age qualification of the survey (being between 12 and 17 years old in the year 1997). Using the household roster information in NLSY, it is possible to link each adolescent respondent of the survey to up to 4 of his or her adolescent siblings. These siblings could be older or younger than the adolescent respondent. I define the variable rtoDO as the ratio of the dropout older siblings of one adolescent respondent:

(2.4.3)

$$rtoDO = \frac{\sum_{j=1}^{4} Y_j}{k}$$

Where Y is the dropout outcome of the older sibling j, and k is the number of older siblings of respondent i who are interviewed as a part of the survey. This ratio ranges from zero to 1, with zero being a respondent with no dropout older siblings, and 1 being an adolescent respondent whose older siblings are all dropouts. Adding this ratio to the univariate probability models of section 2.4.1 will provide a more informative measure of sibling effect, and also provides a relative quality measure of the type of the older siblings one has. Another interpretation of this ratio is the "exposure" measure it provides. If we assume dropout siblings having negative influence on their younger siblings, and the scholarly older siblings having a positive effect on their younger siblings, this ratio can also be interpreted as an exposure factor to the dropout behavior. Table 2.6 shows the results of estimation of the model with the inclusion of the ratio of dropout older siblings.

Table 2.6. Probit coefficients of school dropping out, considering the ratio of older dropout siblings

	Males		Females	
	(1)	(2)	(1)	(2)
I.				
Older Sibling	0.033**	0.174***	0.039***	0.201***
	(0.017)	(0.036)	(0.018)	(0.037)
Ratio of dropout older siblings	0.898***	0.901***	0.866***	1.005***
	(0.087)	(0.110)	(0.088)	(0.109)
II.				
Older Brother	0.042**	0.185***	0.055**	0.237***
	(0.025)	(0.052)	(0.027)	(0.056)
Older sister	0.024	0.164***	0.025	0.165***
	(0.025)	(0.053)	(0.028)	(0.054)
Ratio of dropout older siblings	0.898***	0.900***	0.863***	1.004***
	(0.087)	(0.110)	(0.088)	(0.109)
III.				
Older Sibling within 3yrs	0.116***	0.413***	0.013	0.250***
	(0.034)	(0.067)	(0.036)	(0.070)
Older Sibling plus 3yrs	0.003	0.080*	0.048***	0.183***
	(0.021)	(0.042)	(0.021)	(0.042)
Ratio of dropout older siblings	0.848***	0.873***	0.879***	1.000***
	(0.087)	(0.110)	(0.088)	(0.109)
IV.				
Older Brother within 3 yrs	0.126***	0.433***	0.044	0.285***
	(0.044)	(0.089)	(0.048)	(0.095)
Older brother plus 3 yrs	0.013	0.075	0.054	0.219
	(0.032)	(0.067)	(0.034)	(0.069)
Older Sister within 3 yrs	0.107***	0.393***	-0.017	0.213***
	(0.044)	(0.091)	(0.048)	(0.095)
Older Sister plus 3 yrs	-0.007	0.074	0.044	0.145**
	(0.031)	(0.066)	(0.035)	(0.066)
Ratio of dropout older siblings	0.848***	0.873***	0.875***	0.999***
	(0.088)	(0.110)	(0.089)	(0.109)

The regressions include controls for gender, age, race and etnicity, family income, parental education, familly intactness, urban-rural, and household size.

Standard errors in prantesis; significant levels are .10, .05, .01 indicated by *, **, ***

Similar to the results of section 2.4.1, the estimated γ is positive and significant for males, suggesting that having an older sibling will make the younger brother more likely to drop out of school. This effect is significant for both genders if we consider the quality of the older sibling, which is captured by the percentage dropout variable. The coefficient on the *rtoDO* variable is strongly positive and significant here.

2.4.3 Models with Integrated Behavior of Older Siblings

The models in this section incorporate the information on the behavior of the older sibling, as part of sibling pair analysis. As discussed in section 4.2, NLSY-97 provides information on up to 4 older siblings who are also aged 12-17 years old in 1997. The sibling data that could be used in this framework is therefore an age-selected subset of the siblings of NLSY-97. Out of the 8984 respondent of NLSY, 4035 of them have at least one older sibling interviewed, 839 of them have at least 2 older siblings interviewed, 134 of them have at least three older siblings interviewed, and only 30 of them have four older siblings interviewed. I used a subset of the data that includes the information of the individuals with at least one sibling interviewed given that sibling is an older sibling. For the cases of multiple siblings pairs per family, I included only the pair that is closer in age to each other.

Consider the following model:

(2.4.4)

$$Y_{1it} = X_{it}\beta + Z_{1it}\alpha + u_i + \varepsilon_{1it}$$

$$Y_{2it} = X_{it}\beta + Z_{2it}\alpha + Y_{1it}\gamma + u_i + \varepsilon_{1it}$$

Where Y1 is the older sibling dropout decision and Y2 is the younger sibling dropout decision. X is a vector of observable shared, or family specific characteristics, and Z indicate the vector of sibling-specific characteristics. The term *u* is the time-invariant unobservable shared characteristics between siblings.

I make an assumption that the school completion behavior of the older siblings is not affected by those of younger siblings. In other words the variables related to older siblings' choice are exogenous to the second equation. It is possible to identify γ from second equation alone if there are appropriate controls for unobserved family heterogeneity (*u*). Proper controls for *u* eliminate the possible bias caused by $Corr(Y_1, u)$. Table 2.7 shows the results of estimating this model. The estimated γ is positive and significant in the model, indicating a strong association between having dropout older siblings and the possibility of a dropout for the adolescent.

Table 2.7. School Dropout, considering the schooling decision of the sibling

	All		Males		Femlaes	
	(1)	(2)	(1)	(2)	(1)	(2)
Dropout oldesr sibling	1.533***	1.734***	1.380***	1.634***	1.768***	1.872***
	(0.191)	(0.230)	(0.273)	(0.344)	(0.303)	(0.323)
Age difference between siblings	-0.056	-0.050	0.029	0.062	-0.121	-0.124
	(0.090)	(0.093)	(0.122)	(0.140)	(0.130)	(0.126)
Same-sex siblings pairs	0.082	0.126	0.156	0.201	0.026	0.051
	(0.171)	(0.191)	(0.243)	(0.287)	(0.252)	(0.262)
Age first dropped out	0.170***	0.135***	0.159***	0.105**	0.178***	0.165***
	(0.018)	(0.022)	(0.024)	(0.034)	(0.026)	(0.030)
Black	-0.130	-0.087	0.187	0.360	-0.639	-0.638
	(0.231)	(0.264)	(0.303)	(0.379)	(0.373)	(0.394)
Hispanic	-0.007	0.035	-0.072	0.072	0.032	0.023
	(0.233)	(0.284)	(0.323)	(0.438)	(0.319)	(0.374)
Other race	0.020	-0.031	0.274	0.207	-0.222	-0.239
	(0.243)	(0.296)	(0.342)	(0.461)	(0.340)	(0.392)
Low income household	0.201	0.298	0.416	0.569	0.076	0.098
	(0.185)	(0.214)	(0.252)	(0.321)	(0.271)	(0.293)
High income household	-0.369	-0.366	-0.153	-0.064	-0.599	-0.627
	(0.219)	(0.253)	(0.319)	(0.363)	(0.351)	(0.371)
Urban residence	-0.572**	-0.633**	-0.698**	-0.837**	-0.487	-0.501
	(0.179)	(0.212)	(0.251)	(0.318)	(0.268)	(0.292)
Father's education	-0.016	-0.037	-0.040	-0.072	0.002	-0.007
	(0.044)	(0.045)	(0.058)	(0.065)	(0.069)	(0.065)
Mother's education	0.008	0.031	-0.003	0.025	0.045	0.056
	(0.037)	(0.042)	(0.053)	(0.064)	(0.057)	(0.059)
Intact family	-0.085	-0.047	0.101	0.138	-0.246	-0.207
	(0.202)	(0.223)	(0.286)	(0.352)	(0.291)	(0.289)
Mother's age at the birth of respondant	-0.010	-0.008	0.001	0.008	-0.020	-0.020
	(0.017)	(0.018)	(0.024)	(0.027)	(0.025)	(0.025)
Household size	0.124*	0.146*	0.102	0.130	0.178*	0.187*
	(0.054)	(0.058)	(0.071)	(0.088)	(0.083)	(0.078)
Father's monitoring	-0.022	-0.025	-0.058	-0.068	0.020	0.023
	(0.021)	(0.025)	(0.032)	(0.037)	(0.026)	(0.034)
Mother's monitoring	-0.084***	-0.094***	-0.043	-0.051	-0.132***	-0.138***
	(0.022)	(0.024)	(0.033)	(0.037)	(0.029)	(0.033)
Number of older siblings	0.082	0.096	0.085	0.118	0.117	0.119
	(0.083)	(0.080)	(0.122)	(0.120)	(0.111)	(0.108)
Number of observations	5,238	5,238	2,783	2,871	2,367	2,367
Number of groups		1,628		881		747

(1) Pooled logit regression, clustered robust standard errors

(2) Random effects logit regression

Standard errors in prantesis; significant levels are .10, .05, .01 indicated by *, **, ***

Included but not reported in the regression are dummy variables for years of survey, and the dummy variables for the missing values on parents education, mothers age and family intactness.

As mentioned in section 2.3 of this chapter, the information on the older siblings in NLSY97 is not fully inclusive. Only the siblings that are in a specific age window are included in the survey. This source of selection in the data will distort the values estimated for the effect of dropout older sibling variable. To make a robustness check of the results in table 2.7, I consider a subsample of respondents who have only one sibling who is older, and who has been interviewed in the survey. Table 2.8 compares the results of estimation of the model in the full sample , with the subsample of the younger siblings whose only older sibling is also interview in the NLSY survey. The full result of estimation for the subsample is in table results of this subsample specification are reported in table A.2 of the appendix.

Table 2.8. School Dropout, considering the schooling choice of the older sibling^

	All		Males		Females	
	(1)	(2)	(1)	(2)	(1)	(2)
I. Full sample						
Dropout oldesr sibling	1.533***	1.734***	1.380***	1.634***	1.768***	1.872***
	(0.191)	(0.230)	(0.273)	(0.344)	(0.303)	(0.323)
Number of observations	5,238	5,238	2,783	2,871	2,367	2,367
Number of groups		1,628		881		747
II. Subsample of two-kid families*						
Dropout oldesr sibling	0.788*	1.336***	0.615*	0.886*	1.042*	2.043**
	(0.332)	(0.349)	(0.400)	(0.453)	(0.595)	(0.625)
Observations	4,494	4,645	2,297	2,367	2,009	2,278
		519		258		248

^ The subsample includes the respondents who have only one older sibling who is also surveyed, and this surved sibling is their only sibling.

(1) Pooled logit regression, clustered robust standard errors

(2) Random effects logit regression

Standard errors in prantesis; significant levels are .10, .05, .01 indicated by *, **, ***

Included but not reported in the regression are dummy variables for years of survey, and the dummy variables for the missing values on parents education, mothers age and family intactness.

As expected, the estimates of the sibling effect is smaller in the subsample, which is consistent with the fact that there might be older dropout siblings for the respondents of the full sample that are not accounted for in the data. For example consider the case where there are three siblings in the family, a dropout 19 year old, a dropout 16 year old, and a 14 year old. The oldest sibling is not surveyed as a part of NLSY, and therefore although we know he exists we don't know if he is a dropout or a graduate. The sibling pair of the 16 and 14 year old will end up in our sample. Therefore the effect the model estimates for the 16 year old, is in fact a mix of both the visible and the hidden dropout older siblings. The constructed subsample in this section is free of these hidden siblings effect, but yet estimates positive and significant siblings effect.

2.4.4 Fixed effects estimation

The dataset used in this chapter is three dimensional with respect to individuals, the family they belong to, and the time. There are 8984 teens in this dataset who belong to 6819 unique households, each observed over the period 1997 to 2006. There are 1862 households that include more than one adolescent respondent. Because the structure of the dataset is hierarchical or multilevel, the linear fixed effects component corresponding to these levels can be eliminated using the fixed effects model, and thus result in better estimated of the coefficients of the model.

Consider the model 4.5:

(2.4.5)

$$Y_{it} = X_{it}\beta + Z_i\eta + W_{jt}\gamma + Q_j\rho + \alpha_i + \varphi_j + \mu_t + \varepsilon_{it}$$

i=1,…,8984 is the indicator for individual, t=1,…,10 indicates time, and j=1,…, 6819 is the family/household identification. Y_{it} is the dependent variable X_{it} is a vector of observable and time-variant individual characteristics. Z_i is a vector of time-invariant vector of individual characteristics like gender, race, etc. W_{jt} is a vector of family level characteristics like household income, residence, etc. Q_j is a vector of time-invariant family variables such as parents' age difference, parents' education, etc. ε_{it} is the independent and identical error. The linear fixed effects errors, which are the unobserved heterogeneities in this model, are α_i for individual component of the error, φ_j for family component of the error, and μ_t for the time component of the error.

If these fixed individual and family components of the error are assumed uncorrelated with each other, a random effects model can produce consistent estimates for the parameters. However it is often unacceptable to convey that individual level and family level heterogeneities are not correlated with one another. Fixed effects regression model is one way of dealing with these correlated error terms. If one takes a time difference within each unique individual-family combination, all time-invariant terms will be eliminated from the model. This action will eliminate the problem f unobserved individual and family effects, but at the same time, the coefficients for *Zi* and *Qj* will be dropped out of the estimation too. Since the focal point of this model here is to capture the influence of the behavior of the older siblings on the younger siblings in the family, and this variable is time-variant, the abovementioned problem is not an issue here. Table 2.9 reports the results of the fixed effects models on the individual based, and the family based approach. The complete results of the model are in table A.3 of the appendix.

Table 2.9. Fixed effects estimation of the effect of dropout older sibling, on the dropping out outcome of the teen

	Pooled, clustered s.e.	Individual Fixed Effects	Family Fixed Effect
All			
Dropout Older sibling	1.106***	0.665***	0.156
	(0.110)	(0.136)	(0.115)
Number of observations	20,702	5,709	7,034
Number of groups		631	546
Males			
Dropout Older sibling	0.977***	0.417*	0.103
	(0.151)	(0.190)	(0.171)
Number of observations	10,751	3,145	3,619
Number of groups		353	325
Females			
Dropout Older sibling	1.247***	0.991***	0.830***
	(0.160)	(0.201)	(0.184)
Number of observations	9,951	2,564	2,917
Number of groups		278	257

The regressions include controls for gender, age, race and etnicity, family income, parental education, familly intactness, urban-rural, and household size.

Standard errors in prantesis; significant levels are .10, .05, .01 indicated by *, **, ***

The sibling effect is reduced in magnitude as the unobserved shared effects between siblings are purged out of them model. After the deletion of the unobserved heterogeneity from the models, the sibling effect is still significant and positive for the adolescent females in the data.

2.4.5 Instrumental Variable Models

The last set of models in this chapter is devoted to the practice of instrumental variables. It is evident that even after controlling for as many background characteristics of the two siblings in the models discussed in 2.4.1, 2.4.2, and 2.4.3, the effect of older sibling’s dropout are still under the criticism of non- causality. The fixed effect models discussion in section 2.4.4 offer a remedy for this problem, but in the process they eliminated those characteristics of the siblings that are fixed over the individual or the household.

The instrumental-variables approach provides a consistent estimation given that one can find valid instruments that can explain the endogenous variable (in this model, the dropout status of the older sibling) and at the same time are orthogonal to the residual term. What I propose as instruments for the older sibling's dropout decision are whether he has lived in an intact family when he was 14 year old, the age of sibling's mother at this birth, and the unemployment rate when the sibling was 16 year old. These sibling variables will be valid instruments if they do not directly influence the (younger) teen's own dropout outcome after controlling for the individual and shared family characteristics.

Although this assumption is debatable, in the lack of better instruments in the data I employ them to have a means of validating the results of the models in the previous section. Consider the structural models of the siblings dropout decision, in which Y_2* is the dependent variable in the structural equation and the Y_1 is and endogenous vector of covariates.

(2.4.6)

$$Y_{1i} = X_1\beta + Y_2\gamma + u_i$$

$$Y_{1i} = X_1\pi_1 + X_2\pi_2 + v_i$$

X_2 has exclusive impact on Y_2 and not Y_1. The variable Y_1* is latent, but the binary outcome of dropout (Y_1) is observed to be equal to 1 if $Y_1* > 0$ and $Y_1 = 0$ if $Y_1* < 0$.

To estimate the model in (2.4.6) and alternative, less structural approach is to use the two-stage least-squares (2SLS). This method ignores the binary structure of the dropout variable, although the estimates will still be consistent, the error would be heteroskedastic. In the estimation of the model I will use heteroskedastic robust standard errors which will reflect this problem in the inferences by inflating the standard errors appropriately. The results of the estimation of this model are reported in the upper panel of table 2.10.

Table 2.10. Instrumental-variable estimates of older sibling effect on teen's school dropout outcome

	All		Males		Females	
	1st stage	2nd stage	1st stage	2nd stage	1st stage	2nd stage
I. IV model with heteroskedasticity robust standard error						
Dropout older sibling		0.836***		-0.135		0.771***
		(0.278)		(0.262)		(0.274)
Family intactness at age 14 (older sibling)	-0.016***		-0.014**		-0.016**	
	(0.005)		(0.007)		(0.007)	
Mom's age difference (older sibling)	-0.0005*		-0.001***		0.0001	
	(0.000)		(0.000)		(0.000)	
Unemployment rate at 16 (older sibling)	0.009		-0.014		0.036***	
	(0.007)		(0.010)		(0.011)	
Observations		20,702		10,751		9,951
II. Panel data IV model						
Dropout older sibling		0.817*		0.080*		0.821*
		(0.491)		(0.470)		(0.503)
Family intactness at age 14 (older sibling)	-0.018*		-.018		-.017	
	(.009)		(.0123)		(.015)	
Mom's age difference (older sibling)	-0.0006		-.001*		-.000	
	(.000)		(.000)		(.000)	
Unemployment rate at 16 (older sibling)	0.016		-.001		.037*	
	(.014)		(.019)		(.022)	
Observations		20,702		10,751		9,951
Number of s		2,308		1,208		1,100

The regressions include controls for gender, age, race and etnicity, family income, parental education, family intactness, urban-rural, and household size.
Standard errors in prantesis; significant levels are .10, .05, .01 indicated by *, **, ***

The instrumental-variable model yields positive and significant impact of the dropout older sibling on the dropout chances of the younger sibling. The effect is insignificant for the younger boys.

The lower panel in table 2.10 shows the estimation results for a panel data model of instrumental-variable. The estimated coefficients are not as strongly significant but represent similar results. The sibling effect for boys is not significant and the effect for the girls is slightly significant. The results of these two models suggest the presence of endogeneity of the older sibling outcome to some extent. However the younger sisters in the family still seem to be adversely influenced by the dropout decision of their older siblings.

2.5 Discussion and Conclusion

This chapter used the data on the adolescent cohort of National Longitudinal Survey of Youth 1997 to study the effect of older siblings in the family on the dropout outcome of their younger siblings. The models in this chapter control for a wide range of observed individual and family factors that have been suggested to have influence on the schooling outcome of the teen. The first set of models act as an introduction to the topic of birth order and older siblings. They assess the impact of birth order on the schooling outcome of the younger kids in the family holding sibship size and siblings sex composition constant. These models suggest that having an older sibling increases the chance of dropout in the adolescent respondents of NLSY by 18 and 15 percent for younger boys and girls, respectively. If that older sibling is closer in age to the later born kid, this probability would increase to 47 and 29 percent.

The next advance in the model was to capture the behavior of the older sibling, using the data from those of the adolescent’s siblings who were also independently included in the survey. I first define an *exposure factor* which is the percentage of dropout older siblings for each individual. This factor is particularly informative in the families than have more than one dropout kids. The next set of model were the structural models of the dropout outcome of the younger kid. These models were first estimated under the rather strong assumption of the exogeneity of the older sibling's dropout decision. The result of this model suggests positive and significant siblings effect on both younger brothers and sisters. The robustness of the result was checked using a subsample of 519 two-child families with full information on both siblings.

The last part of this chapter incorporated the possibility that the dropout outcome of both siblings might be determined simultaneously, due to the factor that influence both siblings and are unobservable to the researcher. The family fixed effects models used in section 4.4 difference out these shared unobservable effects, and the instrumental-variables models of section 4.5 use exogenous variations in the older sibling outcome. In both models the significant of the sibling effects is reduced, but is not entirely eliminated. For the case of young female teens, the sibling effects in both models is positive and significant.

In summary the empirical analyses in this chapter provides evidence that having older siblings in the family increases the risk that the younger child in the family would drop out of high school. However, the fixed effects and instrumental-variables models in section 4.4 and 4.5 which treat the endogenous nature of siblings' outcome still do not rule out the sibling effects all together.

Possible explanation for these results could come through the theories of exposure and peer imitation. A younger kid in a family that has an older sibling involved in risky behavior is exposed to that behavior much earlier than he normally would. This could work either through breaking the taboo of the delinquent behavior for the younger kid and encourage him to try that behavior himself. Or considering the fact that younger adolescents find role models in their elders in the family, a dropout role model would be an encouragement for the adolescent to leave school. The policy implication of these findings is that parental investments of the education of their earlier-born children could have spillover effects on the school completion success of their younger children. Also outreach programs that target first-borns in the bigger families could be effective in reducing the teen school dropouts in other kids in the household.

Chapter 3

Family Structure and the Timing of Early School Leaving in Teenagers

3.1 Introduction

The link between completing high school education on the labor market productivity and the subsequent income of individuals has long been established in the field of labor economics. Although the rates of high school dropout is lower in the US and other developed countries, the adverse impact of this problem is enhanced when considering the increasing shift of the blue-collar occupations to the cheaper labor force in the developing countries. Thus dropping out of school is a phenomenon not only engaging the families and educators, but also attracting the policy makers in large. In May 2009 president Obama announced a $900 million grant available to schools and states to reduce the high school dropout rates by investing in the improvement of chronically troubled schools. Although the role the schools play in keeping the kids from becoming dropouts is fundamental, the factors outside of the school portrait an important role in shaping this risk as well.

The peer influence that adolescents receive from their classmates and playmates has been evidently shown in various studies. What should not be downplayed in this matter is the share of

This peer effect influence that work through the kid peers inside his home: his siblings. The impact of this sort of peer effect, or in other words peer exposure, can hardly be avoided by the kid, or contained by intervention of parents, as sometimes in the remedy for the classmates peer effects. This exposure is first hand and around the clock in nature.

In this chapter I will focus on the role the older siblings of an adolescent play in the time pattern of his schooling decision. This study assesses the risk of school dropping out in families that have an older child already dropped out of school. Using the data on two or more siblings in one household enables us to contain the unobservable family factors that influence the schooling behavior of both siblings.

The data used in this study is the NLSY panel data of 1997-2006. A series of parametric and non parametric duration models are used to estimate the effect of older siblings on the reduction of the age of the school dropping out in a teenager. I will closely study the timing of initiation of early school leaving in the adolescent respondents of NLSY, and will link the two concepts of birth order and the age at which the teenager decides to leave the school.

The existence of the clusters of siblings in different families in the dataset of NLSY97, creates the kind of unobservable heterogeneity that is shared between observations that belong to the same family/cluster. The lack of independence among observation in the data cause by this cluster effect, or *frailty*, can lead to biased estimates of the coefficients. I will apply a methodology of proportional hazard models with shared frailty to alleviate this problem and to also estimate the magnitude of this shared dependency in the model.

The remainder of the chapter proceeds as follows. Section 3.2 reviews the literature on the birth order effect on the age onset of early school leaving in adolescents, as well as the role of the behavior of the older siblings on the age onset of the dropping out. Section 3.3 describes the data, and the constructed variables of survival timing. Section 3.4 discusses the empirical models of survival analysis of school dropping out, and addresses the problem of heterogeneity in the hazard models using the shared frailty estimation, while section 3.5 concludes.

3.2 Literature Review

Preventing and intervening in early school leaving in teens require the knowledge of the timing and pattern of occurrence of this act. The age at which adolescent first engages in any type of delinquent or risky behavior could be predictive of later problems with these issues; with earlier acts placing individuals at greater risk for later sever consequences. Knowledge of the age at which the teens are more at risk will make the prevention efforts targeted at them more effective.

In medical studies, there is substantial literature on the age of onset of drug and alcohol abuse in young teens . For instance a one year delay in the initiation of drinking could result in 5% to 9% decrease in alcohol dependency at older age. (Grant et al, 2001). The age of onset of alcohol and marijuana use in teenagers can be increased by proactive monitoring by parents, and reduced when the teen is exposed to other teens that use substance (Kosterman et al, 2000). Hanna et al (2001) in a study of adolescents aged 12–16 surveyed in the Third National Health and Nutrition Examination Survey (NHANES III) show that the use of tobacco and illicit drugsat an early age (age 13 to 16) is likely to increase the probability of later dependency on alcohol and drugs, as well as school problems, early sexual experiences and pregnancy.

There is also a branch of literature in Criminology and Psychology on the subject of age of initiation of delinquencies in adolescents. Developmental theories of crime have focused on analyzing whether the age of onset of criminal behavior has a causal impact on subsequent risky behavior, and whether the determinants of onset vary with age. Nagin & Farrington (1992) do not find evidence to support the causal effect of early onset of delinquencies on later criminal involvements as opposed to simply marking a starting point. However, they find that covariates affecting the the onset of delinquencies and criminal behavior change with the age of the adolescent. Another channel through which the age of initiation has been studies in this literature is its impact not only on the occurrence of the later crimes, but also on its severity. Tolan & Tomas (1995) using the National Youth Survey-1976, show that the early (before age 12) onset of minor delinquencies could result in higher risk of involvement in more intense offences such as robbery and assault in later ages.

In a young adult’s life, dropping out of school is a risky behavior that could have serious long lasting consequences for the individual, as well as the society. Rumberger (1987) shows that the dropouts are more likely to be unemployed later in their lifetime, and are often hired with lower pay and less advancements. The dropouts are more likely to engage in criminal behavior and have higher probability of incarcerations which would induce even higher social costs (Grossman & Kaestner, 1997). Stearns and Glennie (2006) using the data on North Carolina’s public school children find that students at ninth grade have the highest rate of dropout. They also find that students aged 16 and younger are less likely to dropout due to the prospect of joining the labor force, and are unemployed after dropping out.

Dropping out of school could have very persistent effect. Card and Lemieux (2001) using the NLSY-79 show that completed educational attainment is highly correlated with enrollment behavior during ages 16-24. They follow individuals aged 14-16 in year 1979, and find out that 75% of those who dropped out when they passed 16 years of age, never went back to school in the following 10 years. And the majority of the 25% who returned, obtained at most 1 additional year of schooling.

3.3 Data

The data used in this study is from the National Longitudinal Survey of Youth (NLSY-97). This surveys questions 8984 respondents who were between ages of 12 and 16 as of Dec 31, 1996. It has followed them every year since. I will employ 10 rounds of this survey covering the period 1997-2006. After the initial round, no new individuals are added to the dataset, therefore the lower and upper bound of the respondent ages increase by one in each round. Retention rate[10] from year to year is on average above 80% during these 10 rounds.

The dependent variables considered here is dropping out of school. I define dropping out of school as leaving the school before completion and getting a high school or General Education Development degree. In NLSY's survey questionnaire there are fields addressing the respondent's enrollment status in school as of the survey date. Table 3.1 shows how the information on the school enrolment status of the teen is coded in the NLSY questionnaire.

Table 3.1. The enrollment status of the youth respondent of NLSY-1997

Code	Description
1	Not enrolled, No high school degree, No GED
2	Not enrolled, GED
3	Not enrolled, High school degree
4	Not enrolled, some college
5	Not enrolled, 2-year college graduate
6	Not enrolled, 4-year college graduate
7	Not enrolled, Graduate degree
8	Enrolled in grades 1-12, Not a high school graduate
9	Enrolled in a 2-year college
10	Enrolled in a 4-year college
11	Enrolled in a graduate program
-3	Refused to answer
-5	Valid skip

[10] Retention Rate is defined as the percentage of base year respondents remaining eligible who participated in given survey year; deceased respondents are included in the calculations. Reason for not participating (non-interview) includes being deceased, not locatable, technical problem, respondents too ill, respondent unavailable, refused to interview, or other. Among these the refusal to interview and being non-locatable are respectively the major reasons.

A “dropout” in this chapter is defined as an individual that reported “Not enrolled”, and have “No high school degree, no GED”. Respondents who are working towards a GED are coded as being enrolled regardless of where that course of study took place.

Table 3.2. The sample example to illustrate the calculation of the age of first dropout

Youth ID	Dropout in round 1?	Dropout in round 2?	Dropout in round 3?	Dropout in round 4?	Dropout in round 5?	Dropout in round 6?	Dropout in round 7?	Dropout in round 8?	Dropout in round 9?	Dropout in round 10?	**dropout time**
1001	1	0	1	1	0	1	1	0	0	0	**1**
1002	0	0	0	0	0	0	0	0	0	0	**0**
1003	0	0	1	0	0	0	0	1			**3**
1004	0	0	0	0	0	0	0	0	0	0	**0**
1005	1	1	1	1	0	1	1	1	1	1	**1**
1006	0	0	0	0	0	0	0	0	0	0	**0**
1007	0	0	0	0	1	0			0	0	**5**
1008	0	1	1	1	1	1	1	1	0	0	**2**
1009	0	0	0	1	0	1	1	1	0		**4**
1010	0	0	0	0	0	0	0	0	0	0	**0**
1011	1	1	1	0	1	1	1	0	0		**1**
1012	0	0	0	0	0	0	0	0			**0**
1013											**NA**
1014	0	0	0	0	0	1	1	1	1	1	**6**

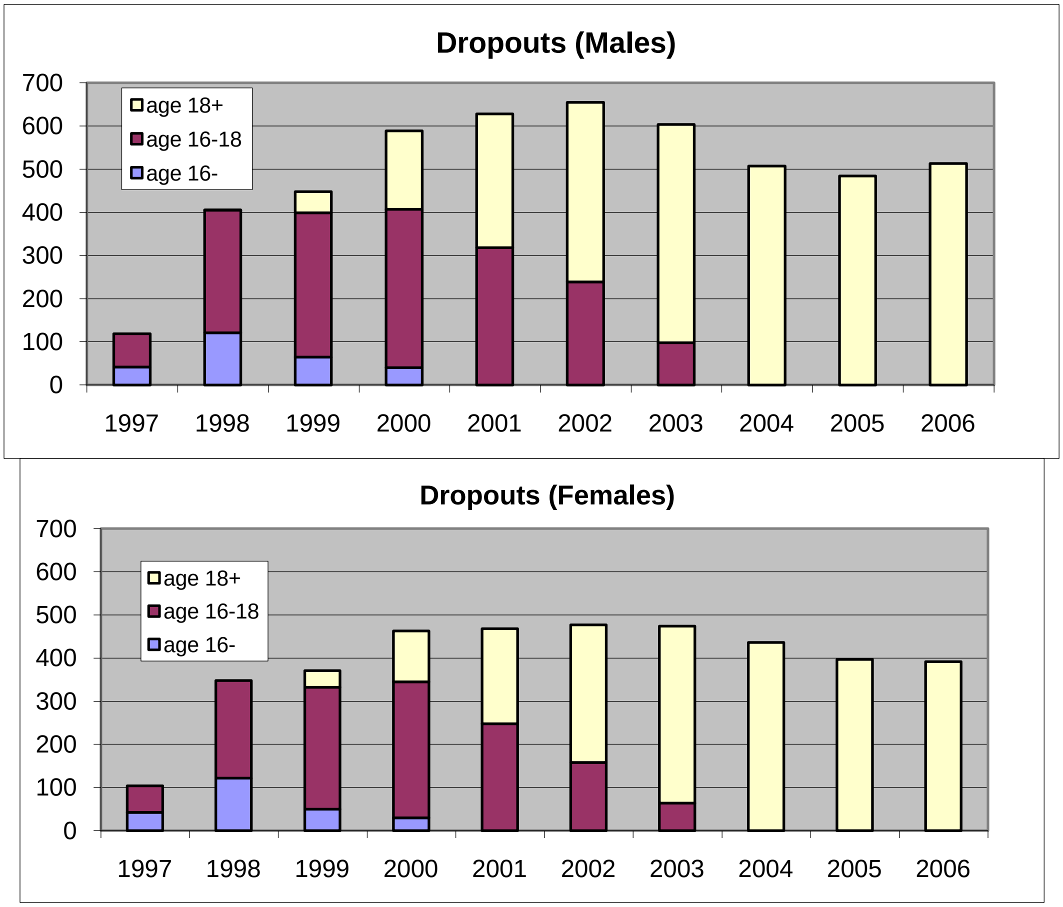
The duration models analyze the time to the occurrence of the dropout event. The main variable thus would indicate the starting point of dropping out of school. The panel feature of the NLSY dataset makes it possible to construct this duration at the point of the first failure, which is the earliest round of the survey in which the teen is considered a dropout. This can be extracted from comparing the teenagers answers to the enrollment status question in consecutive rounds of the NLSY-97. In order to construct this variable, the following assumptions are made: If an individual drops out at some point in time, but goes back to school sometime after, the failure is considered on the first dropout and will not reset, as seen in the illustrative Table 3.2[11]. The missing values (gaps in Table 3.2) are assumed to follow the immediate non-missing value to their left (e.g. row 1005 in Table 3.2). In the case that an entry is missing for all rounds, the duration value will be marked as missing as well. I marked all-zero entries as being right-censored (e.g. row 1004).

Dropping out of school could be illegal depending on the compulsory school laws in different states. The compulsory attendance law requires parents to have their children in a public or state accredited private or parochial school for a designated period. In the United States, where my data is collected from, the compulsory education varies by state, beginning at ages five to eight and ending at the ages of sixteen to eighteen[12]. A growing number of states have now compulsory education until the age of 18. Figure 2.1 shows the number of dropouts in each year of the survey. As mentioned, the respondents are all between 12 and 16 years of age in the initial year of the survey. In year 1998 the first wave of respondents passes the age threshold of 16, which is the smallest compulsory age in some states, and thus we see a hike in the number of dropouts.

[11] I also considered another measure for age of initiation which marks the starting point of dropout only if the teenager has shown the behavior in any two consecutive rounds. The result of the hazard analysis was robust to this modification, and thus not reported.

[12] *Source:* Department of Education, National Center for Educational Statistics, *Digest of Education Statistics, 2004.*

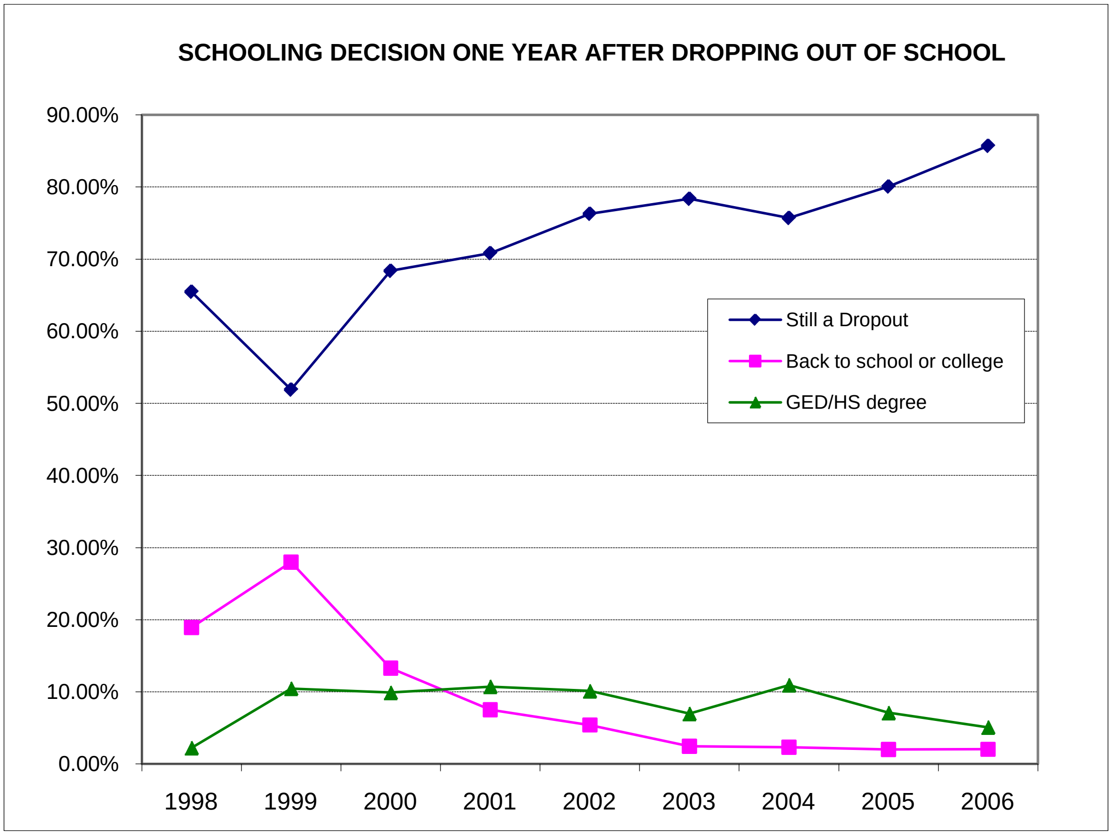
Figure 3.1. Number of Dropouts by Age, 1997-2006


Dropouts (Males)

Dropouts (Females)

In the dataset used in this chapter, the NLSY-97, at least 70% of the individuals who dropped out the year before stay dropout in the following year (Figure 3.2). On average less than 8% of the people go back to school in the year after they drop out. And less than 10% of the people in the dataset graduate from high school or get a General Educational Development (GED) degree the year following their dropout. This pattern is more pronounced for the people who dropped out at older ages. Those younger dropouts who break the spell of dropping out sooner, are more likely to go back to school. The older dropouts that leave the school later in the time span of the survey, are more likely to seek an equivalent high school degree.

Figure 3.2. The Schooling choice of the dropouts, one year after the event



In this chapter, the focus is on the effect of the older siblings on the age at which a teenager drops out of school for the first time. NLSY provides the complete roster of the family members of each respondent. The roster includes the relationship, the gender, and the age of each family member, no matter if he or she resides in the household at the time of the survey, or has moved out. Up to four of these household members accounted for in the roster who qualified for the age criteria of the initial round (i.e. aged 12 to 16 years old in the beginning of the year 1997) were also surveyed. Unless the older siblings in the family are aged within this window, the information regarding them is only limited to the roster information ad thus not useful for this study. As an example, consider a household with 5 children aged 10, 13, 15, and 19 as of

December 31st, 1997. And a 23 year old who has moved out of the household. In this case the 13 and 15 year olds will be surveyed and followed up every year after. But we will not have any information on the characteristics and behavior of the resident 19 year old, or the nonresident 23 year old. This is a shortcoming if specifically the 19 year old is the bad influence dropout kid in the family. However both 19 and 23 year old children will be considered in the counting of older siblings. Table 3.3 shows the breakdown of the surveyed and not surveyed dropouts in the dataset.

Table 3.3. Number of dropouts in the 10 round of NLSY 1997, by older sibling specification

year	Total dropouts	dropouts with an older sibling	dropouts with an older sibling who is surveryed	dropouts with an older sibling who is surveryed and who dropped out
1997	223	126	96	15
1998	754	433	368	59
1999	819	476	400	82
2000	1052	628	528	130
2001	1096	655	548	143
2002	1132	675	597	157
2003	1078	642	565	144
2004	943	564	500	125
2005	881	521	468	118
2006	905	545	478	121
Total	8883	5265	4548	1094

The last column of table 3.2.3 is the subsample of the NLSY used where the analysis in this chapter involves the schooling behavior of the older siblings. I consider the sibship variable constant during the time span of the dataset. This is not a strong assumption as this chapter emphasizes on the information of the older siblings of a respondent. As both resident (in the household) and non-resident siblings are accounted for in generating the sibship variables, the count could be assumed not to change much, as time passes. The violation to this assumption would be if an older sibling of a respondent dies during this period, or the respondent’s family adopts an older kid, both incidences are rare in the data.

A set of demographic and socioeconomic control variables are used in the different duration models of this chapter. Table 3.4 gives the summary statistics of the dependent covariates used in the models. Apart from demographic variables, there are some controls over characteristics of the parents, such as their highest level of education, whether or not they are dropouts themselves, age of mother at the birth of the respondent, and whether or not the teen has lived with both parents till 14 years of age.

The variables used to capture the effect of family income are 3 dummy variables: whether the household income is below 125% of the federal poverty line (Poor HH), whether it is over 400% of the federal poverty line (Rich HH), and if there is no household income recorded. The middle income households are the reference case. The data on household income is inquired on each round of the survey. For the cases where there were holes in the stream of reported household income, I filled the gaps in the data using the last reported income up to that point in time. For example if a household has a reported income of $100,000 for year 1997, missing the data on year 1998, and $120,000 on year 1999, I approximated the family income equal to $100,000 in year 1998.

Table 3.4. Summary statistics of the duration model independent variables

Variable	Obs	Mean	Std. Dev.	Min	Max
A. Demographic and socioeconomic variables					
Female	8984	0.488	0.500	0	1
Age	8984	14.307	1.475	12	18
Black	8984	0.266	0.442	0	1
Hispanic	8984	0.211	0.408	0	1
Other race	8984	0.152	0.359	0	1
Highest grade of the biologic parents education	8984	12.417	4.167	0	20
Father is a dropout	7120	0.229	0.420	0	1
Mother is a dropout	8290	0.236	0.425	0	1
Mother's age at birth of the teen	8374	25.482	5.413	10	54
Intact family at age 14	7935	0.727	0.445	0	1
Low income household	8984	0.207	0.405	0	1
High income household	8984	0.155	0.362	0	1
Missing income info	8984	0.270	0.444	0	1
Household size	8984	4.549	1.540	1	16
Urban residence	8604	0.764	0.425	0	1
One child family	8984	0.091	0.287	0	1
Two child family	8984	0.303	0.459	0	1
Three child family	8984	0.280	0.449	0	1
Four child family	8984	0.166	0.372	0	1
B. Sibship configuration variables					
Older sibling	8984	0.946	1.211	0	11
Older brother	8984	0.480	0.784	0	7
Older sister	8984	0.466	0.792	0	7
Older sibling within 3 years	8984	0.390	0.582	0	5
Older sibling older than 3 years	8984	0.556	1.043	0	11
Older brother within 3 years	8984	0.201	0.432	0	3
Older brother older than 3 years	8984	0.279	0.642	0	7
Older sister within 3 years	8984	0.189	0.427	0	4
Older sister older than 3 years	8984	0.277	0.652	0	7
Have any dropout older sibling	4035	0.153	0.360	0	1

I use a series of variables to capture the set up of the sibship in the family. These include the number of resident and nonresident older siblings, the number of older siblings that are within 3 years of age difference with the respondent, and the number of older siblings that have more than 3 years of age difference with the respondent. The survey includes data on up to 4 other teen aged kids in the family of the respondent household. I link the characteristics and the behavior of these 4 siblings, given they are present in the survey and are older, to the outcome of the primary teenager.

3.4 Models and Empirical Results

The models in this chapter are built to analyze the time to occurrence of the school dropping out event in the youth respondents of NLSY97. The impact of the older siblings are addressed in two ways: (1) Birth order effect: Assessing the effect of an older sibling in on the school completion outcome of a teenager, regardless of the outcome of the older sibling. (2) Exposure Effect: Targeting the effect of the school completion outcome of the older sibling on that of the younger one. The first approach has the advantage of absorbing the information on all of the older siblings of the teen respondents, as well as the advantage of using a bigger pool of the data. The second approach factors in the specific dropout behavior of the older siblings, and in this sense introduces more effective information to the model. The disadvantage of this approach is the loss of the older siblings outside the age window of the survey.

The results in the first chapter suggested the existence of a significant inverse relationship between the age of onset of dropping out in adolescent and the likelihood of the teens staying a dropout at present. The continuing step in this analysis addressed in this chapter is to determine the factors that might have an impact on the age of start of school dropping out, and in particular testing to see if presence of older siblings in the household, and their behavior, has any effect on changing the age of onset of dropping out. Models of duration analysis are a good choice to help answer these questions.

There are three approaches to the estimation of the duration models: (1) Nonparametric methods, which make no an assumption on the distribution of the duration variable, nor on how the covariates change the survival experience. I will use Kaplan and Meier (1945) method of estimation of survival curves, and use the qualitative covariates such as the format of sibship to test for any significant differences in the survival curves in the different categories of those covariates. (2) Semiparametric modeling, which do not require an assumption on the shape of the duration variable, by concentrating on the order on which the failure occurs, rather than the distribution of the failure time. But at the same time, the effect of the other covariates can be parameterized. I use Cox's (1972) Proportional Hazard Model to analyze the effect of different individual and family covariates on the hazard rate. (3) Parametric modeling, which assumes a functional form for the duration variable, and estimates the time to failure using the functional form adopted.

3.4.1 Nonparametric Models

The analysis of survival data can take different forms depending on what the researcher would be willing to assume about the data on the event survival. Nonparametric analysis, by not imposing any restrictions on the survival procedure, allows the data to speak for itself. This approach will not allow for modeling the effect of the covariates that might have an impact on the survival schedule. However it is possible to compare the estimates of the nonparametric survival function at different values of some qualitative covariates.

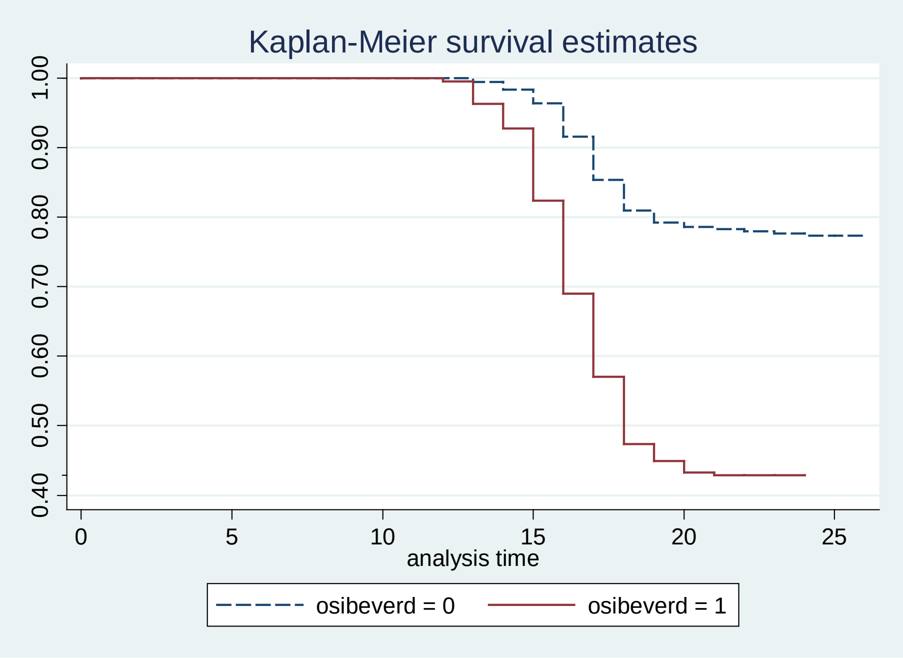
The first nonparametric model in this section is the Kaplan-Meier (1958) estimator. It calculates the survivor function S_t which is the probability of survival until the point t at time, as

(3.4.1)

$$\widehat{S_t} = \prod_{j\,|t_j<t} \frac{(n_j - d_j)}{n_j}$$

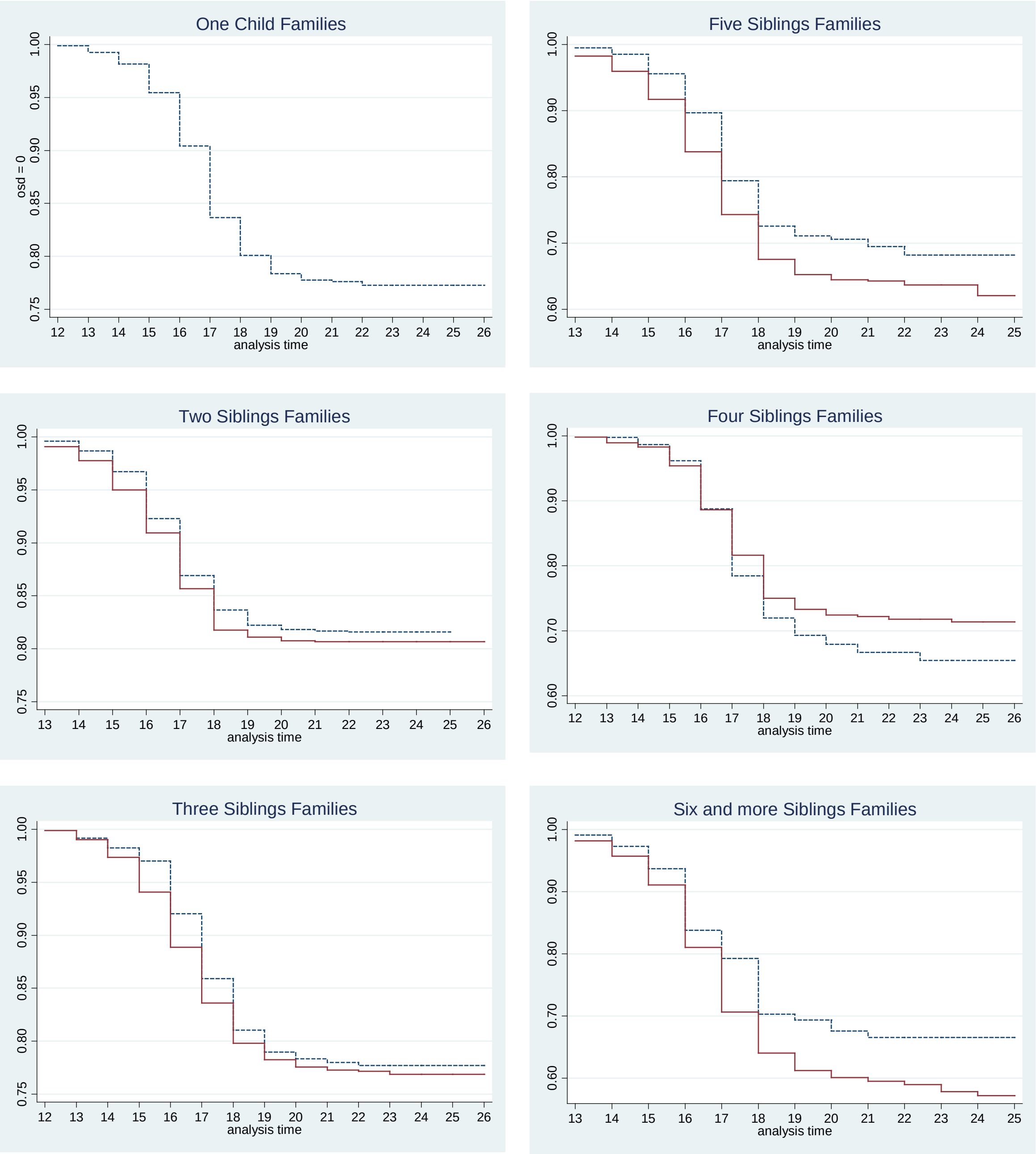
where n_j is the number of individuals at risk at time t_j , and d_j is the number of failure at time t_j. The survivor function S_t in (3.4.1) is calculated numerically using the actual realizations of the event. The estimation of the survivor function for school dropout in our data is shown in the figure 3.3, in which the deviation between the survival curve of the adolescents who have dropout older siblings, and that of the ones who do not have such sibling is apparent. The original numerical results of Kaplan-Meier estimation are in table B.1 of the appendix.

Figure 3.3. Kaplan-Meier estimates of school dropping out for teens with or without dropout older siblings.



Kaplan-Meier survival curves by different sibship sizes are depicted in Figure 3.4 of the appendix. The dotted lines indicate the survival for the teens who do not have an older sibling, and the solid line indicate the survival curve of the teens with at least one older sibling. The sample sizes used in deriving these curves are 814, 2720, 2520, 1491, 722 and 717.

Figure 3.4. Kaplan-Meier survival curves, by different sibship sizes



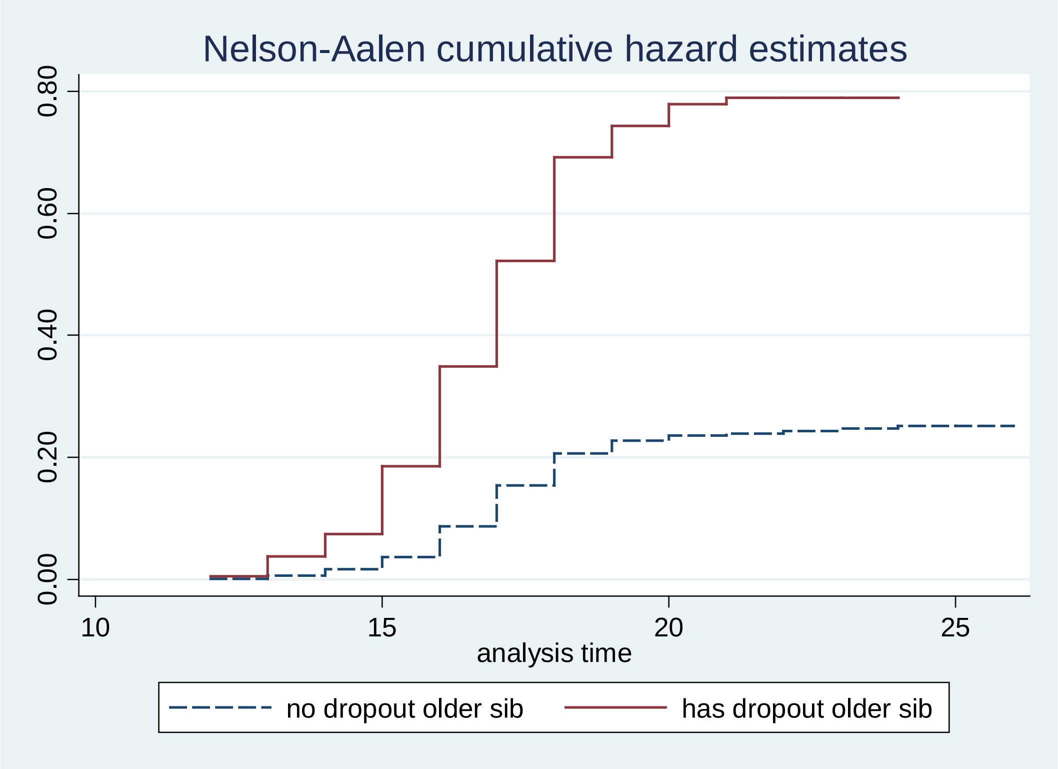
The cumulative hazard function can be estimated using Nelson-Aalen estimator (1972, 1978), as

(3.4.2)

$$\widehat{H_t} = \sum_{j\,|t_j<t} \frac{d_j}{n_j}$$

where n_j is the number of individuals at risk at time t_j , and d_j is the number of failure at time t_j . The result of the estimation of the hazard of school dropout is illustrated in figure 3.5 The original numerical results are in table B.2 in the appendix.

Figure 3.5. Nelson-Aalen curves for the cumulative hazard estimated of school dropout



3.4.2 Semiparametric Regression Models

Let T be the random variable capturing the time when a delinquency (failure) happens, with and t be its realization. Define f(t) as the density and F(t) as the distribution function of T. The Hazard rate is defined as :

(3.4.3)

$$h(t) = \frac{f(t)}{1 - F(t)}$$

which is the rate at which failure happens at time t, given it had not have happened before.

Consider a hazard function in the form of equation 3.4.4:

(3.4.4)

$$h\left(\mathrm{t}|\ \mathrm{x}_j\right) = h_0(t)\ \mathrm{r}(\ x_j\,\beta_{\mathrm{x}})$$

The hazard function in (3.4.2) consists of two multiplicative terms. $h_o(t)$ is the baseline hazard function of an unspecified form, which characterizes how the hazard changes depending on the duration time variable. The function r(.) characterizes how the hazard function changes as a result of a change in the covariate vector x. The baseline hazard term can be left unestimated since it will cancel from the calculations when we consider the proportion of hazard in different values of the covariates. Thus the model makes no assumption about the shape of the baseline hazard function over time. The equation (3.4.5) shows the proportion of the hazard in two different set of values of covariates x_j and x_k defined as the hazard ratio:

(3.4.5)

$$HR\left(t, \mathrm{x}_j, \mathrm{x}_k\right) = \frac{h\left(\mathrm{t}|\ \mathrm{x}_j\right)}{h(\mathrm{t}|\ \mathrm{x}_k)} = \frac{\mathrm{r}(\ x_j \beta_\mathrm{x})}{\mathrm{r}(\ x_k \beta_\mathrm{x})}$$

The hazard ration only depends on the function r(.) and the values of the covariates, and not the shape of the baseline hazard function. Cox (1972) proposed this model for the estimation of the hazard ratio and used exponential functional form for the r(.) function:

(3.4.6)

$$h\left(\mathrm{t}|\ \mathrm{x}_j\right) = h_0(t)\ \exp(\ x_j \beta_\mathrm{x})$$

The Cox model is estimated using partial likelihood estimation. Similar to the hazard function itself, the likelihood fuction of the proportional hazard model can be factored into two parts. One that depends on the β_x and one that depends on the baseline hazard function. The partial likelihood method discards the second factor and maximizes the likelihood based on the first factor. The resulting estimators from this method are unbiased and normally distributed; however they are not fully efficient because some information is lost by ignoring the nature of the baseline hazard function. But the loss of efficiency is often small enough to be tolerated (Efron 1977).

As the analysis of the Cox semiparametric model is base on the unchanging nature of the baseline hazard function, it is essential to verify this assumption for the empirical model. I check the assumption of proportionality of the hazard function based on the analysis of the scaled Schoenfeld (1982) residuals. In this method the residuals of the Cox model are estimated as the difference between the covariate values of the failed observations, and the weighted average of the covariate values over all observations at risk of failure at each point of time. To provide a better diagnosis, these differences are then scaled by their estimated variance. If the hazard is not proportional, these residuals will show a time pattern in them. The hypothesis tested here is whether there is a non zero slope in the regression of the scaled Schoenfeld residuals when a time trend is fit to them. In the initial estimations of the Cox proportional hazard model for this data, the proportionality test did not verify the assumption of hazard proportionality in the model including all covariates linearly.

One way to fix this problem is by stratifying the model based on the covariates that cause the non-proportionality. In Stratified Cox estimation, the assumption that everyone faces the same baseline hazard is relaxed in favor of:

$$h(\mathrm{t}|\ \mathrm{x}_j) = h_{01}(t)\ \exp(\ x_j \beta_\mathrm{x}) \qquad \text{, if j is in group 1}$$
$$h(\mathrm{t}|\ \mathrm{x}_j) = h_{02}(t)\ \exp(\ x_j \beta_\mathrm{x}) \qquad \text{, if j is in group 2}$$

The baseline hazards are allowed to differ by group, but the coefficient beta is constrained to be the same.

Based on the pattern of residuals in the model, the candidate variables that might have caused the non proportionality are: Gender, Age, Race, and Ethnicity of the respondents. After stratifying the model based on these variables, the resulting model passes the proportionality test.

The first set of models in this chapter aim to capture the impact of a teens birth order, and his or her sibship configuration. The covariates capturing the configurations of the older siblings are used in the proportional hazard model. The result of the effect of siblings on the hazard rate of dropping out of school in the stratified model in reported in Table 3.5.

Table 3.5. The effect of siblings on the age of first school dropping out

	All Sample				Males				Female			
Model 1												
Have an older sibling	0.161***				0.207***				0.092			
	(0.156)				(0.074)				(0.087)			
Model 2												
Have an older brother		0.181***				0.241***				0.096		
		(0.056)				(0.074)				(0.087)		
Have an older sister		0.114**				0.113*				0.116*		
		(0.056)				(0.075)				(0.089)		
Model 3												
Older sibling within 3 years			0.190***				0.175***				0.161**	
			(0.045)				(0.066)				(0.081)	
Older sibling older than 3 years			0.054**				0.025				0.066	
			(0.024)				(0.038)				(0.043)	
Model 4												
Older brother within 3 years				0.186***				0.267***				0.169*
				(0.071)				(0.075)				(0.092)
Older brother older than 3 years				0.042				0.033				0.062
				(0.041)				(0.051)				(0.068)
Older sister within 3 years				0.149**				0.170**				0.118*
				(0.061)				(0.077)				(0.097)
Older sister older than 3 years				0.066				0.075				0.052
				(0.041)				(0.056)				(0.062)
observation	5817	5817	5817	5817	3007	3007	3007	3007	2810	2810	2810	2810
failure	1410	1410	1410	1410	817	817	817	817	593	593	593	593
Proportionality test (prob>chi2)	.103	.156	.159	.176	.576	.592	.517	.405	.109	.110	.116	.102

Cox model is stratified using gender, age and race dummies

Standard errors in prantheses. ***p<.01, **p<.05, *p<.10

Included covariates in the Cox model: parents education, family intactness, family income, mother's age difference with respondent, household size, and urban/rural.

The result of models (1)-(4) in table 3.5 suggest that the teens who have higher birth orders in a family are significantly faced with a greater hazard of dropping out of school compared to the teens who are of the lower birth orders. As the results show, this impact is more pronounced in young boys, compared to young girls. The hazard of dropping out of school is higher for teens who have an older sibling closer to their age (within 3 years age difference) compared to those who have much older sibling. The adverse effect of older brothers seems to outweigh that of the older sisters. The impacts of other covariates included in the model but not reported in the table are in Table B.3 of the appendix.

The second set of hazard models in this chapter aims to capture the effect of younger siblings' exposure to the event of dropout in their older siblings. In these models, the information on the behavior of the older sibling is incorporated in the analysis. As discussed in section 3.2, NLSY97 provides information on up to 4 older siblings who are also aged 12-16 years old in 1997. This design of the survey, will put some restrictions on the information available to use in the analysis. The data that could be used in this framework is a selected subset of the NLSY97 dataset. These are the families who have 2 to 5 age-eligible teenagers present in the household at the time of the initial round of the survey. For instance an older sibling who is 18 year old and leaving in the household will not included in the survey because only the children within 12 to 16 years old in the year 1997 were eligible to be interviewed.

Table 3.6. shows the results of estimating this model. The variable of interest in these set of models is Dropout Older Sibling. This variable is equal to 1 if at least 1 out of the 4 older siblings interviewed has been a dropout in any time within the 10 year time span of the dataset.

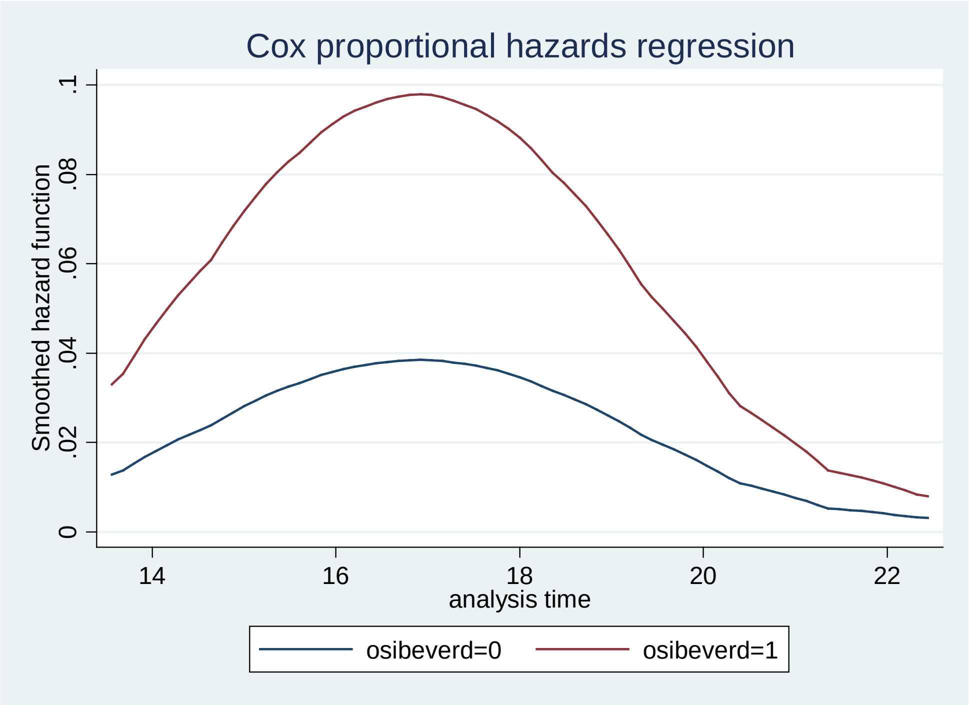
Table 3.6. Cox Proportional Hazard Model of the Age of School Dropping Out

	All Sample	Male	Female
Has an older sibling	0.017	-0.069	0.131
	(0.099)	(0.129)	(.157)
Has a Dropout older sibling	0.936***	0.847***	1.033***
	(0.094)	(0.127)	(0.143)
Parents highest degree of schooling	-0.008	-0.014	0.004
	(0.019)	(0.024)	(0.029)
Father is a dropout	0.014	0.116	-0.112
	(0.113)	(.148)	(0.177)
Mother is a dropout	-0.0130	-0.202	-0.035
	(0.121)	(0.160)	(0.189)
Mother's age at birth of the teen	0.004	0.003	0.005
	(0.007)	(0.009)	(0.012)
Intact family at age 14	-0.136	-0.034	-0.277*
	(0.087)	(0.118)	(0.130)
Low income household	0.750***	0.687***	0.849***
	(0.104)	(0.136)	(0.163)
High income household	-0.523**	-0.448	-0.625*
	(0.191)	(0.241)	(0.315)
Missing income info	0.257*	0.159	0.390*
	(0.108)	(0.142)	(0.167)
Household size	0.050*	0.078*	0.013
	(0.024)	(0.032)	(0.037)
Observations	2630	1365	1265
Failure	686	365	291
Proportionality test	0.21	0.864	0.024

Cox model is stratified using variables gender, age and race dummies
Standard errors in prantheses. ***p<.01, **p<.05, *p<10

Although the use of the proportional hazard method factors out the estimation of the baseline hazard function, it is of interest to investigate the pattern of survival in the phenomenon of school dropping out in the adolescents. $h_o(t)$ can be estimated as the derivative of the cumulative hazard function after being smoothed for the inherent discontinuities in it. In figure 3.6, the hazard functions are numerically calculated and graphed for the two groups of adolescents who have at least one older dropout sibling in their family, and the adolescents who do not.

Figure 3.6. Estimated hazard functions: teens with versus without older dropout siblings



3.4.3 Parametric Models

The nonparametric and semiparametric models discussed in sections 3.4.1 and 3.4.2 had the advantage of imposing zero or minimal restrictions on the form of the survival experience. However if the nature of the survival experience is known form previous research, the parametric models have certain advantages. Their fitted values can directly provide the estimated survival time, and they utilize the full maximum likelihood estimation which has more efficiency compared to the partial likelihood estimation. I used five different parametric models to estimate the hazard of dropping out of school. The first three functional forms, Exponential, Weibull, and Gompertz model the hazard of dropout. The other two functions, Log-normal and Log-logistic model the failure time. The results of the nonparametric estimation using the maximum likelihood method are in table 3.7.

Table 3.7. The parametric estimates of school dropping out in youth of NLSY-97

	(1)	(2)	(3)	(4)	(5)
Has an older sibling	0.068	0.088	0.091	-0.011	-0.018
	(0.096)	(0.096)	(0.096)	(0.023)	(0.024)
Has a Dropout older sibling	0.838***	1.034***	1.023***	-0.255***	-0.266***
	(0.092)	(0.092)	(0.092)	(0.025)	(0.025)
Parents highest degree of schooling	-0.006	-0.007	-0.007	0.001	0.001
	(0.018)	(0.018)	(0.018)	(0.004)	(0.005)
Father is a dropout	0.002	0.012	0.011	-0.001	-0.003
	(0.110)	(0.109)	(0.109)	(0.027)	(0.028)
Mother is a dropout	-0.160	-0.211	-0.210	0.030	0.039
	(0.118)	(0.118)	(0.118)	(0.029)	(0.030)
Mother's age at birth of the teen	0.003	0.003	0.003	-0.001	-0.001
	(0.007)	(0.007)	(0.007)	(0.002)	(0.002)
Intact family at age 14	-0.084	-0.109	-0.108	0.025	0.031
	(0.085)	(0.085)	(0.085)	(0.021)	(0.022)
Low income household	0.682***	0.806***	0.802***	-0.193***	-0.205***
	(0.102)	(0.102)	(0.102)	(0.025)	(0.026)
High income household	-0.536**	-0.556**	-0.556**	0.108**	0.127**
	(0.190)	(0.190)	(0.190)	(0.037)	(0.043)
Missing income info	0.270*	0.294**	0.294**	-0.058*	-0.068**
	(0.105)	(0.105)	(0.105)	(0.025)	(0.026)
Household size	0.035	0.049*	0.049*	-0.010	-0.010
	(0.023)	(0.023)	(0.023)	(0.006)	(0.006)
Urban residence	0.012	0.025	0.025	-0.001	-0.004
	(0.096)	(0.095)	(0.095)	(0.023)	(0.024)
Female	-0.296***	-0.359***	-0.356***	0.079***	0.086***
	(0.078)	(0.078)	(0.078)	(0.019)	(0.019)
Age	0.004	-0.110***	-0.124***	0.029***	0.027***
	(0.030)	(0.031)	(0.031)	(0.007)	(0.007)
Black	0.335***	0.365***	0.368***	-0.084***	-0.091***
	(0.094)	(0.094)	(0.094)	(0.023)	(0.024)
Hispanic	0.151	0.168	0.170	-0.040	-0.040
	(0.123)	(0.123)	(0.123)	(0.030)	(0.031)
Other race	0.154	0.169	0.169	-0.040	-0.043
	(0.129)	(0.129)	(0.129)	(0.032)	(0.032)
Constant	-5.026***	-12.893***	-6.146***	3.056***	3.083***
	(0.558)	(0.676)	(0.570)	(0.137)	(0.139)
Observations	2,630	2,630	2,630	2,630	2,630
No. of failures	686	686	686	686	686

Standard errors in parentheses. *** p<0.001, ** p<0.01, * p<0.05

(1) Exponential regression -- log relative-hazard form
(2) Weibull regression -- log relative-hazard form
(3) Gompertz regression -- log relative-hazard form
(4)Lognormal regression -- accelerated failure-time form
(5)Loglogistic regression -- accelerated failure-time form

It is important to recognize that the use of parametric models are appropriate when the researcher has an idea of what the baseline hazard looks like so to impose that idea into a constraint to the model in order to (1) obtain the most efficient estimates for the coefficients and (2) obtain a better estimate of the baseline hazard based on that idea. If no such insight into the shape of the hazard is available, the Cox model that imposes no –and consequently no 'wrong'-distributional form on the baseline hazard- might be a better choice of model. The non-parametric estimate of the hazard in figure 3.6 shows an increasing trend in the hazard that has larger speed at the beginning and then slows down. Considering this pattern, the Weibull distribution that pertains a monotone increasing hazard rate might be a good choice for the parametric model. Also the log-normal, and log-logistic models that show an increase at first followed by a decrease as duration time goes by, could also apply to the model in this chapter. According to this model, having a dropout older sibling will reduce the time to failure (dropout) of the teen by 25% which is by about 3 months earlier than the dropout age onset of a teen who do not have such older sibling.

3.4.4 Sibship Frailty in Hazard Models

The siblings in one family often share many unobservable characteristics that influence their behavior and socioeconomic outcome. Their dropout survival experience is not an exception to this shared unobservable effect. In the estimation of the semiparametric and parametric models in sections 3.4.1 and 3.4.2. it was implicitly assumed that the duration models are fully specified, meaning that the unaccounted for information contained in the residual are randomly distributed. However when dealing with the data that entails observation on somehow related individuals, like siblings in a family, it is plausible to assume that the unobservable characteristics of one siblings is to some extent related to those of the other sibling. The problem of these shared unobservable effects have been outlined in the first chapter of this study in the context of linear regression models. I will address this problem in the framework of duration models.

Models of shared frailty can address the effect of the shared unobservable among siblings of the same household on the hazard rate of one. The frailty[13] models are designed for the duration data that has groups or clusters of individuals. These models enable us to account for the individual heterogeneity when it is shared among the members of a cluster. The frailty is defined as a latent multiplicative effect that is assumed to have unit mean and finite variance. Consider the cluster specific frailty of $\mathrm{w_i}$, as a multiplicative term in the general form of hazard function:

(3.4.7)

$$h(\mathrm{t_{ij}}\,|\,w_i, \mathrm{x}_j) = w_i h_0(\mathrm{t_{ij}}) \exp(\, x_{ij}\, \beta_{\mathrm{x}})$$

Index i represents the groups or clusters, and index j represents the observations within the groups. The frailty W_i is shared within all the members of the ith group. W_i is assumed to have mean equal to 1 and a finite variance θ. If the frailty factor is larger (smaller) than 1 for a cluster, then the individuals in that cluster have a greater (smaller) hazard of the dropping out of school.

3.4.4.1 Parametric Hazard Models with Shared Frailty

There are usually two functional forms used to specify the frailty factor: gamma distribution and inverse gaussian distribution. For the simplicity of calculations, consider the gamma distribution [14]with parameter *a* and scale parameter of *1/a*, and define the frailty as:

(3.4.8)

$$f(w_i) = w_i^{a-1} \exp(-aw_i\,) a^a / \Gamma(a)$$

[13] The term Frailty was first used by Vaupel et al. in a study of dynamics of mortality.

[14] In addition to the Gamma distribution for the frailty factor, I also use the Inverse Gaussian distribution with mean 1 and finite variance. Tables 4.5, and A.4 include both choices for the frailty distribution.

By design, w_i has mean equal to 1 and variance equal to *θ=1/a.* If the variance is zero, then the observations are independent and shared heterogeneity is of no impact. But if the variance is significantly different than zero, then the frailty factor which the hazard is multiplied by, will impact the survival process to account for the shared heterogeneity in the clusters.

Consider the general form of Weibull proportional hazard model in (4.9), with the shape parameter of p.

(3.4.9)

$$h(\mathrm{t}|\ \mathrm{x}) = \ \exp(\ x\beta_{\mathrm{x}})\mathrm{pt}^{\mathrm{p}-1}$$

In the absence of heterogeneity, p and β will be estimated from the data on the survival time and the x covariates, as in section 3.4.3. By introducing frailty into the model the hazard function will be changed into:

(3.4.10)

$$h(\mathrm{t}|\mathrm{x}) = w\ \exp(\ x\beta_{\mathrm{x}})\mathrm{pt}^{\mathrm{p}-1}$$

where w is a random unobservable factor with mean 1 and variance θ. The hazard function in (3.4.10) can be estimated using the EM algorithm for clustered data as in Guo and Rodriguez (1992). The results of the estimation for the Weibull hazard model, as well as the Exponential hazard model are presented in table 3.8. The shared frailty estimation for the other types of distributions are reported in table B.4 of the appendix.

Table 3.8. Parametric estimation of the shared failty models of dropout hazard

	Exponential Hazard			Weibull Hazard		
	Standard	Gamma Freilty	Inv. Gaussian Frailty	Standard	Gamma Freilty	Inv. Gaussian Frailty
Have a dropout older sibling	0.838	0.838	0.838	1.034	0.145	0.262
	9.13	9.13	9.13	11.21	0.82	1.63
Have an older sibling	0.068	0.068	0.068	0.088	0.262	0.214
	0.71	0.71	0.71	0.91	2.15	1.77
Parents highest grade of education	-0.006	-0.006	-0.006	-0.007	-0.012	-0.013
	-0.33	-0.33	-0.33	-0.39	-0.48	-0.53
Father is a dropout	0.002	0.002	0.002	0.012	0.014	-0.009
	0.02	0.02	0.02	0.11	0.1	-0.06
Mother is a dropout	-0.160	-0.160	-0.160	-0.211	-0.167	-0.178
	-1.36	-1.36	-1.36	-1.79	-1.07	-1.15
Mother's age at birth of the teen	0.003	0.003	0.003	0.003	0.005	0.005
	0.42	0.42	0.42	0.5	0.51	0.53
Intact family at age 14	-0.084	-0.084	-0.084	-0.109	-0.028	-0.014
	-0.98	-0.98	-0.98	-1.28	-0.25	-0.13
Low income household	0.682	0.682	0.682	0.806	1.157	1.138
	6.69	6.69	6.69	7.91	7.24	7.28
High income household	-0.536	-0.536	-0.536	-0.556	-0.802	-0.752
	-2.82	-2.82	-2.82	-2.92	-3.4	-3.14
Missing income info	0.270	0.270	0.270	0.294	0.316	0.318
	2.56	2.56	2.56	2.79	2.08	2.1
Household size	0.035	0.035	0.035	0.049	0.061	0.078
	1.51	1.51	1.51	2.11	1.62	2.04
Urban residence	0.012	0.012	0.012	0.025	-0.024	0.005
	0.13	0.13	0.13	0.26	-0.17	0.03
Female	-0.296	-0.296	-0.296	-0.359	-0.425	-0.437
	-3.82	-3.82	-3.820	-4.61	-4.38	-4.490
Age	0.004	0.004	0.004	-0.110	-0.202	-0.208
	0.15	0.15	0.15	-3.59	-4.86	-5.05
Black	0.335	0.335	0.335	0.365	0.570	0.601
	3.54	3.54	3.54	3.86	3.85	4.1
Hispanic	0.151	0.151	0.151	0.168	0.323	0.277
	1.230	1.230	1.230	1.360	1.750	1.530
Other race	0.154	0.154	0.154	0.169	0.161	0.22
	1.2	1.2	1.2	1.32	0.81	1.13
_cons	-5.026	-5.026	-5.026	-12.893	-14.602	-14.761
	-9	-9	-9	-19.08	-16.59	-16.61
Frailty variance (theta)		0.000	0.000		2.118	3.536
		-0.08	-0.03		4.18	4.69

Values of test statistic is reported below each estimate.

In the Weibull model, the variance of the frailty factor is significantly bigger than 1 in both gamma, and inverse Gaussian choice of frailty. This implies the existence of shared heterogeneity among siblings of the same family which has inflated the estimates of the effect of dropout older siblings. The corrected estimates for this shared frailty are reported in the second and third columns of the table 3.8. However when the distribution of hazard is assumed to follow an exponential distribution, the estimates for the variance of the frailty is zero, which implies that the unobserved heterogeneity with this choice of distribution is not significant. The answer to the question of which of these two models are more appropriate to use, requires a prior understanding on the behavior of the duration procedure to match it with the best possible parametric function. For example, for a homogeneous population with a hazard function that increases in time a monotone fashion- which is consistent with how the dropout hazard increases by age, the monotone Weibull function seems to be a good choice.

3.4.4.2 Cox Proportional Hazard Model with Shared Heterogeneities

The modeling of heterogeneity in the context of shared frailty are better done in the fully parametric models, as discussed in section 3.4.1.1, where the contribution of the multiplicative shared error can be estimated in both baseline, and factored parts of the hazard function. However by imposing the assumption that the shared frailty only has impact on the exponential part of the Cox model, it is possible to estimate how it will affect the coefficients of the model. The Cox model with shared frailty is equivalent to a random effects Cox model, Shared frailty are used to model within group correlation. Observations within a group are correlated because they share the same frailty, and the extent of correlation is measured by θ. In the context of school dropping out, it is plausible to expect that the survival data on the siblings within a household are correlated because some families would inherently be more frail to dropping out than others. Consider the Cox proportional hazard model in equation 3.4.6. The shared frailty factor is modeled as a latent random effect multiplied to the hazard function for family(group) i:

(3.4.11)

$$h\left(\mathrm{t}|\ \mathrm{x}_j\right) = h_0(t)w_i \ \mathrm{exp}(\ x_j\beta_{\mathrm{x}})$$

where $\mathrm{w_j}$ is the group level frailty with mean 1 and variance θ. For $\mathrm{v_i} = \log(\mathrm{w_i})$, the hazard can be expressed as

(3.4.12)

$$h\left(\mathrm{t}|\ \mathrm{x}_j\right) = h_0(t)\ \mathrm{exp}(\ x_j\beta_{\mathrm{x}} + \mathrm{v_i})$$

which is similar to a standard linear random effects model. If θ is equal to zero, the Cox shared frailty model reduces to the standard Cox model. The results of the estimation of the dropout hazard are reported in table 3.9. The estimated variance of the frailty factor is not significantly different than zero. Therefore, although the estimation of the coefficients are slightly changed in the adjusted model, the presence of heterogeneity in the household level seem to be not of crucial importance in the framework of Cox proportional hazard modeling of this data.

It is worthy to mention here another method of correction for the heterogeneity caused by clustered nature of the data. This method is to adjust the standard errors of the estimates by clustering on the household id. This way the precision of the point estimates are adjusted for the possible correlation between the observations without making any parametric assumption on the nature of these shared unobservables that impact the survival time. This method however will not correct the estimates for this error. The result of the estimation using this method is also reported in the table 3.9.

Table 3.9. Cox proportional hazard estimation of dropout, with family shared frailty

	Standard	Robust s.e.	Shared Frailty
Have a dropout older sibling	0.973	0.973	0.937
	(0.091)	(0.096)	(0.094)
Have an older sibling	0.054	0.054	0.058
	(0.096)	(0.092)	(0.098)
Parents highest grade of education	-0.008	-0.008	-0.007
	(0.018)	(0.018)	(0.019)
Father is a dropout	0.023	0.023	0.022
	(0.109)	(0.104)	(0.112)
Mother is a dropout	-0.179	-0.179	-0.171
	(0.118)	(0.119)	(0.121)
Mother's age at birth of the teen	0.004	0.004	0.004
	(0.007)	(0.007)	(0.007)
Intact family at age 14	-0.112	-0.112	-0.107
	(0.085)	(0.084)	(0.088)
Low income household	0.742	0.742	0.764
	(0.102)	(0.108)	(0.105)
High income household	-0.536	-0.536	-0.542
	(0.190)	(0.190)	(0.192)
Missing income info	0.275	0.275	0.275
	(0.105)	(0.109)	(0.108)
Household size	0.041	0.041	0.043
	(0.023)	(0.023)	(0.024)
Urban residence	0.014	0.014	0.010
	(0.096)	(0.096)	(0.099)
Female	-0.327	-0.327	-0.330
	(0.078)	(0.078)	(0.079)
Age	0.005	0.005	-0.001
	(0.030)	(0.029)	(0.031)
Black	0.307	0.307	0.318
	(0.095)	(0.095)	(0.098)
Hispanic	0.128	0.128	0.135
	(0.123)	(0.123)	(0.127)
Other race	0.167	0.167	0.165
	(0.128)	(0.128)	(0.133)
Obs	2,630	2,630	2,630
Frailty variance (theta)			0.099
			(0.105)

Standard errors in prantheses.

3.5 Discussion and Conclusion

The main goal of this chapter was to learn about the timing of the decision of an adolescent to leave the school before graduation, as well as the role of some of the factors that have an impact on this timing. Using the duration analysis for the age of school dropout in teens, I looked into the models of how of an adolescent's risk of dropping out of school is shaped by one or more incidence of school dropping out in his older siblings. I argue that the presence of older siblings in a family will put the adolescent in a greater risk of dropping out at a younger age. Using Cox proportional hazard model, and after controlling for the family size, and other determinant factors, I find the presence of an older sibling increased the hazard of dropout by 16% in the adolescent respondents of NLSY97. This impact when originated from the older siblings that are closer in age to the adolescent, is estimated to be 18% for the males and 16% for the females; about three times as much as the effect of much older siblings for both genders.

The next step in the analysis was to capture not only the presence of the older siblings in the household, but also their schooling outcome. In NLSY-1997 there is information available on some of the siblings of an adolescent. Not all the older siblings of an adolescent respondent in the NLSY are questioned as a part of the survey. The siblings surveyed are the ones that were closer in age to the respondent, because the much older (or much younger) siblings did not fit into the age requirement (being 12 to 16 years of age) of the survey. I used nonparametric, semiparametric and parametric models to investigate the contribution of having a dropout as sibling to the hazard of the adolescents' dropping out of school themselves. The Cox semiparametric, and five different fully parametric hazard models all estimate large significant effect for the adverse impact of the presence of a dropout in the family on the younger kids. However the effect of shared frailty is ignored in these estimates.

The final attempt of this chapter was to take account of the shared frailty in the siblings data used in the survival analysis. I followed the algorithm of Guo and Godrigues (1992) for incorporating the shared frailty effect in the models for the hazard of the younger sibling. The finding of this chapter show little evidence of the impact of shared frailty in the Cox proportional hazard model. However in the parametric models, depending on the choice of the survival time distribution, and the assumed form of the cluster error term, the shared family frailty was estimated to be from 2 to 3.3 in variance. The impact of this significant frailty was a considerable drop in the estimated impact of the dropout siblings.

Chapter 4

A Discrete Choice Model of the Siblings Influence in their Dropout Decision

4.1 Introduction

The effect of the siblings on the educational attainment of one another can fit into the economic models of social interaction. A young person is constantly subject to the influences of the peers around him in his classroom, in his neighborhood, and within his family. Social interactions are particularly strong in the adolescent years of life as the young individual is actively seeking life role models in the people whom he is in contact with.

People often show the tendency to conform their behavior to the behavior of the ones around them. This seems to be the case in the schooling decision making as well. Using a model of social distance, Akerlof (1997) argued that individuals receive utility from behaving like an "average" person in a reference group. He shows that the individuals tend to make decisions about schooling based on their tendency to correspond to the educational attainment of the significant people around them, as the deviation from the average behavior of their close contacts will induce a cost to the individual. Becker (1996) and Durlauf (1999) indicate that the probability that an adolescent adopts a certain behavior depends on the prevalence of that behavior among their peers.

The school based and neighborhood based peer influence on the young adults have been well documented in the literature. The school-based peer influences on adolescent's risky choices have been well documented (Evans et al. 1992, Gavira and Raphael 2001, Clark and Loheac 2003). By contrast, few studies have investigated the dynamic interactions among peers within the family: the siblings. This chapter will focus on the peer effects and social interactions that rise from within the family through the adolescent siblings. The social interaction among the siblings is more predominant compared to the adolescent's interactions with his friends or classmates, and are less possible for the adolescent to avoid or retract. This chapter will develop a binary choice interaction model with finite number of agents to characterize the peer effect of siblings on the strategic choices of the teenager. This dynamic model incorporates the attractiveness of imitating the behavior of one's siblings. The model measures the strategic complementarity between the choice of the teenager and the current and previous choices of his or her siblings. An empirical application of the social interaction model in school dropout outcome of the teenagers is also presented using the siblings data from National Longitudinal survey of Youth.

The chapter is organized as follows: Section 4.2 overviews the existing literature on the teenage group interactions. Section 4.3 presents the discrete choice model of social interaction between siblings. Section 4.4 explains the reflection problem in the estimation of social interactions.

Section 4.5 used the data to test the existence and the magnitude of the siblings' complementarities in dropping out of school, in correspondence to the model. The last section contains the concluding remarks.

4.2 Literature Review

The analysis of social influence and peer pressure in determining adolescents' educational outcomes goes back to the work of Pollak (1976), who examines the behavioral consequences of preferences depending directly on others' behavior in model of habit-formation and learning. Zimmer and Toma (2000) analysis indicates that peer effects are a significant determinant of educational achievement. The effects of peers appear to be greater for low-ability students than for high-ability students.

Brock and Durlauf (2000) provides an analysis of aggregate behavioral outcomes when individual utility exhibits social interaction effects. They study generalized logistic models of individual choice which incorporate terms reflecting the desire of individuals to conform to the behavior of others in an environment of non-cooperative decision making. Duncan et al (2001) find that the sibling correlations in delinquent behaviors are larger than any of the correlations between peers defined as adolescents' best-friends, between schoolmates living in the same neighborhood, and between pupils in the same grade within a school. Their data suggests that family-based factors are several times more powerful than neighborhood and school contexts in affecting adolescents' achievement and behavior.

Gaviria and Raphael (2001) analyze school-based peer effects in the individual discrete choice behavior of tenth-graders. They find strong evidence of peer-group effects on the dropout behavior of the individuals, where their testing for endogenity of this effect is rejected. Their index for the peer effect is the percentage of the peers that are dropouts. Sacerdote's (2001) study using data on college roommates in the dormitory finds that the academic performance of one's roommate is positively correlated with the student's academic performance. Hoxby (2000) identifies the effects of peers in the classroom using sources of variation that are credibly idiosyncratic, such as changes in the gender and racial composition of a grade in a school in adjacent years. She finds that students are affected by the achievement level of their peers: a one point increase in the reading scores of the peers raises a student's own score between 0.15 and 0.4 points, depending on the specification. Ammermueller and Pischke (2009) find sizable estimates of peer effects for fourth graders in six European countries. They use school fixed effects to control for the nonrandom assignment of students across schools, and make an assumption that once the students select the school they attend they are randomly assigned to different classes within one school.

Kooreman and Soetevent (2007) investigate peer effects in substance abuse at the school-class level using Dutch National School Youth Survey data for the year 2000, using school fixed effect as a way to identify the effect. Their estimated peer effects without the school fixed effects are large and statistically significant, larger for boys (0.72) than for girls (0.58), and larger within genders than across genders. Including the school level fixed effects dramatically reduces the magnitudes of the estimates to 0.13 and 0.11 respectively, with the latter statistically insignificant.

4.3 Model

This section outlines the basic model of discrete choice of the teenagers between dropping out of or remaining enrolled in school. The model follows the framework suggested by Brock and Darlauf (2001, 2003) and Soetevent and Kooreman (2007) to account for the social interactions in small groups when the choice variable is a binary option. This model will allow for the determination of the existence or lack of conformity in the decision making of the teenagers that are under the influence of their sibling-peers.

Consider a large group of teenagers (agents) organized into small and non-overlapping groups (households). Each agent is in interaction with his siblings in the household. The size of each group in the model is therefore the size of sibship in each household *(N).* At each point of time, the teenager in family *i* makes a binary decision to either dropout, or to stay in school. The binary choice set of the individual is Y_i={-1,1} where 1 indicates the dropout decision and -1 indicate the enrolled status (no dropout)[15]. The agents in each group have interaction with one another so that their strategy to dropout or not depends on their own behavior as well as the behavior of the other agents/siblings in their group/family. The strategy profile of agent *i* therefore consists of (Y_i, Y_{-i}) which is the combination of his own choice, plus the choices of every other person in his group.

[15] The choice of (-1,1) would give qualitatively the same results as the conventional (0,1) choice as long as the model have an intercept term. The advantage of using this notation is that it makes it possible to capture the effect of the “good” peer at the same time.

Each agent makes his choice between dropout or schooling so to maximize his payoff function of $V(Y_i, X_i, Y_{-i}, \varepsilon_i)$: Y →R , where X_i represents the individual characteristics of agent i, and ε_i is the unobservable individual-specific random part of the payoff. Following Soetevent and Kooreman (2007) the payoff function is laid out as having three additive terms:

(4.3.1)

$$V(Yi, Xi, Y\text{-}i, \mathcal{E}i) = U(Yi, Xi) + S(Yi, Xi, Y\text{-}i) + \mathcal{E}i$$

The term U(Yi, Xi) is the Private Utility that the teenager derives from his choice of Yi. This term also assumed to depend on Xi , the exogenous socio-economic characteristics of the agent. The private utility term can get different functional forms; For simplicity in this chapter it is assumed to be a linear function of the teenagers characteristics.

(4.3.2)

$$U (1,Xi) = \beta 1Xi$$

$$U(-1,Xi) = \beta\text{-}1Xi$$

The term S(Yi, Xi, Y-i) is the Social Utility associated with the interaction of the agent's siblings/peers outcome with the self outcome. This term is a function of the self and peers decisions, as well as the observed individual's characteristics that might explain individual's taste to follow for detest the peer pressure. Consider the Social Utility term to be defined as

(3.3)

$$S_i = \mathrm{S}(Y_i, X_i, Y_{-i}) = \frac{\gamma}{(\mathrm{N}-1)} Y_i \sum_{\mathrm{j} \neq \mathrm{i}} Y_j$$

This functional form, associates the social utility of the teenager to the aggregate outcome of his siblings. The coefficient γ determines the level of complementarity between the behavior of the teenager and his siblings/peers. If the γ is positive, the agent will get a positive value for his social utility if he conforms his behavior with the prevalent behavior of the group. That is if most of his siblings in the family are dropouts, and he drops out himself, the term S will be positive. Similarly if most of his siblings are enrolled or graduates, resulting in a negative value for the sum, the term S will be positive if he adopts the no-dropout (-1) strategy as well. The positive γ is an indication of the strategic complementarity between the choices of the siblings and thus point to the existence of the dynamic peer effect. On the contrary with a negative value for the γ, the social utility term is only positive when the agent chooses to go against the majority outcome of his peer group.

Consider the latent variable v* which is the difference in total payoff of the teenager when he chooses to dropout versus when he chooses not to dropout.

(4.3.4)

$$v^* = V(1) - V(-1)$$

$$= (\beta_1 - \beta_{-1})X + \frac{\gamma}{N-1} \sum Y_j + E(1) - E(-1)$$

$$= (\beta_1 - \beta_{-1})X + \gamma\, \bar{Y} + E(1) - E(-1)$$

The optimal payoff maximizing strategy of each agent is to choose their action based on the following criteria:

(4.3.5)

$$Y_i = \begin{cases} 1 & if\ v_i^* > 0 \\ -1 & if\ v_i^* \leq 0 \end{cases}$$

The pure Nash equilibrium profile happens if an only if all the agents in the group choose their action according to (3.5). Bjorn & Vuong (1983) and Kooreman (1994) prove that this Nash equilibrium exists, and that for the small group size, the number of the possible equilibriums are bounded. For more detail on the model refer to Soetevent and Kooreman (2007). In section 5 I will test this model with the data on siblings on NLSY to estimate the coefficient γ, and determine the degree of conformity among the siblings in one family.

It is important to note that the social utility model of equation (4.3.3) treats peer effect coming from each of the agents in the group as identical to another. The results of the models in the first two chapters of this dissertation make clear distinctions among the sibling/peers of a teenager. The siblings who are closer in age to the agent have a larger influence compared to the siblings that are farther apart in age. Furthermore the sibling effect is varied whether coming from a brother or a sister depending on the own sex of the teenager. The uniform effect across the agents as modeled by Soetevet & Kooreman will miss these differential effects. I will extend the model to employ a linear structure which will incorporate appropriate weighing of the aggregate siblings outcome according to the age closeness of each pair. Asides from the age difference of the sibling and the teen, there are some other weight factors that could be considered, such as the effect of the siblings being a full, half, step, or adoptive child of the family, which will make the effect non-uniform.

(4.3.6)

$$S_i = \mathrm{S}(Y_i, X_i, X_j, Y_{-i}) = \frac{\gamma}{(\mathrm{N}-1)} Y_i \sum_{\mathrm{j}\neq\mathrm{i}} w_j \left(X_i, X_j\right). Y_j$$

In (4.3.6) the social function is now defined to be dependent on both the characteristics of the teenager and those of his siblings. The weight function of (4.3.7) is the age-closeness factor weighing the outcome of the siblings up or down.

(4.3.7)

$$w_j = 1/|A_j - A_i|$$

where A_i and A_j represents the Age of the individual and his sibling. For the age difference of 1 year there is no weight assigned, and for the case of twin siblings an arbitrary weight equal to 2 is considered.

Another way to interpret these weights is to consider the discounting effect they have on the effect of dropout siblings. Considering the five age difference schedules allowed by the NLSY survey, the equivalent discount rate corresponding to the age differences of 2, 3, 4, and 5 years in

(4.3.6) are 0.4, 0.44, 0.41, and 0.38. Therefore the weighting schedule in (4.3.6) is approximately equivalent of a discount rate of 40% for the years the siblings are apart in their age.

Another extension to the model that will be used in the empirical estimation is to differentiate the siblings peer outcome by gender. That is evaluating the social utility equation of (4.3.5) separately for sisters and brothers of an individual.

4.4 Identification Strategy

It is important to address the possibility of the endogeneity of the siblings' behavior in the models of peer influence. The behavior of the teenager and his siblings might simultaneously be influenced but the shared cofactors that exist within their group/family and are unobservable to the researcher. For example being born to parents who devalue high school graduation would influence all the siblings in the family, and thus the dropout of the teenager in this family is more likely to be the result of the parents' attitude rather than siblings' dropouts. The challenge

is to isolate peer effects from this correlated effects due to the correlation between peer composition and the omitted individual or institutional characteristics that can affect student outcomes. To control for this non-random selection into families, I will also estimate a version of the model that includes family specific fixed effects.

Another main issue that should be considered in the estimation of the social interaction effects is that the behavior of the individual in a group could be moving conjointly with the average behavior of the members of that group, and not determined as a result of the behavior of the peers in the group. This problem unveiled by Manski (1992, 2000) as the *reflection problem*, which can flaw the empirical estimates of the peer effect if the estimation technique fails to account for this group based endogeneity.

Specifically, individual behavior may be similar to the behavior of the social group he belongs to because of these three different reasons, which are defined by Manski (1993, pp. 532—533):

> "*Endogenous effects*, wherein the propensity of an individual to behave in some way varies with the prevalence of that behavior in the group;
>
> *Exogenous (contextual) effects*, wherein the propensity of an individual to behave in some way varies with the distribution of background characteristics in the group;
>
> *Correlated effects*, wherein individuals in the same group tend to behave similarly because they have similar individual characteristics or face similar institutional environments."

Manski notes that both endogenous and exogenous effects reflect social interactions, whereas correlated effects are rather a statistical, non-social, phenomenon. While endogenous and exogenous effects both reflect social interactions, the policy implications differ sharply. Manski cites the example of a tutoring program, which is provided for some students in a school. The achievement of non-tutored students in the same school will improve with endogenous effects, but not with exogenous effects (as the tutoring program does not change the reference group's characteristics).

A standard regression of individual behavior on group means cannot distinguish between endogenous and exogenous effects, and only in some situations can both be distinguished from correlated effects. This identification difficulty, which Manski calls the reflection problem, arises because group behavior is by definition the aggregation of individual behavior.

One solution to overcome the reflection problem is to link the behavior of the individual with the lagged behavior of the social group, in place of the contemporaneous value of the group average behavior (Lee, 2007). This resolves the identification problem if one can determine the appropriate lag length. (Manski, 2000, p. 129). I will use this strategy to alleviate the reflection problem in the model.

4.5 Empirical results

This section aims to use the model laid out in section 4 to establish whether an adolescent's propensity to dropout is affected by the prevalence of the dropout behavior among the individual's siblings. The relevant group of social interaction is defined as the sibship in the family the individual belongs to. Each individual in the sample has between 1 and 4 siblings. It is important to note that the structure of NLSY's survey that has an age condition of 12 to 16 for its respondents will introduce some type of age selection into the sample. In other words the sample is consisted of families that are bound to have 1 to 4 kids in a time span of 4 to 5 years.

4.5.1 Data description

The sample used in this section is derived from the National Longitudinal Survey of Youth (NLSY97). This is a nationally represented survey of about nine thousand teenagers aged 12 to 16 in the year 1997, which were surveyed annually since then. For each teenager in the dataset, his or her siblings who were also aged within the 12 to 16 years window were also included in the survey. The other siblings in the family are accounted for in the family roster, but are not interviewed, thus lacking the information on their behavior. To overcome this information disparity among siblings in a family, I will use a subset of the NLSY97 where all the siblings in the family are age-eligible to be interviewed as part of NLSY in the year 1997. This restriction will reduce the sample size to 1113 individuals. These individuals have between minimum one and maximum four siblings who are in the survey and whose characteristics and behavior including their dropout decisions is recorded in each round of the interviews. The summary statistics of the variables used in the model are presented in the table 4.1.

Table 4.1. Summary Statistics of the Independent variables in the Sample

	Mean	Std. Dev.	Min	Max
Individual characteristics				
sex	0.482	0.500	0	1
age	19.296	2.986	13	26
black	0.215	0.411	0	1
hispanic	0.170	0.375	0	1
other race	0.137	0.344	0	1
Family characteristics				
low income hh	0.216	0.411	0	1
high income hh	0.313	0.464	0	1
missing hh income	0.119	0.324	0	1
urban	0.640	0.480	0	1
dropout dad	0.387	0.487	0	1
dropout mom	0.283	0.450	0	1
intact family	0.633	0.482	0	1
mom-child age diff	25.570	5.342	13	54
Sibship characteristics				
sisters	0.603	0.623	0	3
brothers	0.628	0.608	0	4
older soblings	0.659	0.674	0	4
younger siblings	0.572	0.619	0	3
total siblings	1.231	0.515	1	4

Sample size is 10017, consisting of 1113 individuals in 9 rounds of the survey 1998-2006.

School dropout variable is constructed such that, at each point of time, the individuals who are either enrolled in school, or have a high school or a G.E.D. degree are considered "Not a dropout". The dropout rate by age in this subsample is presented in table 4.2.

Table 4.2. Dropout rates in the sample by sibship, gender, and age

	Mean	Std. Dev.	Sample size
Dropout rates among			
all teens	8.79%	0.283	8749
siblings of the teen	7.72%	0.251	10017*
sisters of the teen	7.05%	0.248	5355
brothers of the teen	9.07%	0.280	5697
Dropout rate by age			
less than 16 yro	3.48%	0.183	1062
between 16 & 18 yro	9.77%	0.297	2713
more than 18 yro	9.39%	0.292	4974
Dropout rate by gender			
males	10.29%	0.304	4459
females	7.23%	0.259	4290

* Missing information on dropout status of the sibling is counted as zero.

In this chapter I will use different measures of peer effects of school dropout The first measure is the percentage of the dropout siblings, that is shown on the top panel of table 4.2. The other set of measures in correspondence with the weighing schedule of equation (4.3.5) are shown in the table 4.3.

Table 4.3. Dropout among sibling peers of the teenagers in the sample

	Obs	Mean	Std. Dev.	Min	Max
Percentage of dropouts in the household					
All siblings					
simple	10017	7.724%	0.251	0	1
age-weighted	10017	5.216%	0.199	0	2
Sisters					
simple	5355	7.05%	0.248	0	1
age-weighted	5355	5.04%	0.211	0	2
Brothers					
simple	5697	9.075%	0.280	0	1
age-weighted	5697	5.847%	0.212	0	2

4.5.2 Estimation Results

This section presents the results of the estimation of the size of the social utility gained by conforming or digressing from the average behavior of the group (siblings in the family). Table 4.4 presents five versions of the estimated model for school dropout. The first column contains the estimation results for a model that assumes no social interaction effects among the siblings. In correspondence with the equation 4.3.4, the estimated coefficients for the individual and family characteristics that are used in the model are in fact the difference of the effects in the case of the dropout versus enrollment: $(\beta_1 - \beta_{-1})$.

The effect of gender is significant, confirming the stylized fact that girls have lower national dropout rates. The odds of dropping out of school increase with age. The racial minorities have a higher risk of dropout, as do the teenagers from low income families. Living in urban areas, or in an intact family doesn't significantly change the odds of dropping out. Overall the effects of the covariates are consistent with the findings of empirical models of young adults dropout behavior.

Models (2) and (3) include the social interaction term of table 4.3 into the regression. Model (2) assumes that the schooling outcome of all the agents (siblings) in interaction with the teenager have equal effect on his dropout choice. Model (3) assumes that the interaction with the dropout siblings who are closer in age to the teenager is more effective compared to those siblings who are much older or younger than the teen. Both models estimate a positive and significant value for the gamma coefficient in the equation (4.3.4).

Table 4.4. Estimation of school dropout, with different specifications of social interaction

	no social interaction	with social interaction	with weighted social interaction	with lagged social interaction	with lagged weighted social interaction
	(1)	(2)	(3)	(4)	(5)
social		1.803			
		6.23			
social-w			2.81		
			7.43		
Lag(social)				1.617	
				5.86	
Lag(social-w)					1.874
					5.38
sex	-0.872	-0.871	-0.925	-0.877	-0.916
	-2.81	-2.83	-2.98	-2.86	-2.98
age	0.079	0.074	0.081	0.058	0.067
	3.17	2.96	3.19	2.29	2.64
black	1.796	1.578	1.561	1.601	1.645
	4.85	4.33	4.2	4.4	4.48
hispanic	1.342	1.183	1.269	1.2	1.287
	2.91	2.59	2.76	2.64	2.83
raceother	1.668	1.487	1.483	1.501	1.545
	3.46	3.12	3.06	3.15	3.23
low income HH	0.232	0.241	0.238	0.277	0.271
	1.25	1.3	1.27	1.5	1.46
high income HH	-0.787	-0.821	-0.806	-0.796	-0.783
	-3.43	-3.56	-3.48	-3.45	-3.41
missing income info	-0.026	-0.041	-0.071	0.001	-0.01
	-0.08	-0.12	-0.2	0	-0.03
urban	-0.234	-0.215	-0.236	-0.213	-0.236
	-0.98	-0.9	-0.98	-0.89	-0.99
dropout dad	0.017	0.049	0.095	0.051	0.074
	0.05	0.14	0.27	0.15	0.22
dropout mom	-0.825	-0.711	-0.758	-0.723	-0.776
	-2.11	-1.84	-1.94	-1.87	-2.01
intact family	0.098	0.11	0.132	0.105	0.119
	0.3	0.34	0.4	0.33	0.37
mom-kid age diff	-0.022	-0.015	-0.015	-0.016	-0.018
	-0.76	-0.53	-0.53	-0.58	-0.63
log-likelihood	-1328.773	-1312.804	-1302.031	-1314.702	-1316.189
N	8749	8749	8749	8749	8749

The models reported are random effects logistic regressions.
t-values reported beneath each estimate.

The models (4) and (5) use the lagged value of the social interaction term in an attempt to alleviate the reflection problem in the simultaneous choices of the siblings. The estimation result of these two model shows that although the size of the conformity gain in social utility is reduced moderately, the social interaction effect is still significantly increasing the odds of dropping out of school.

The next step into evaluating the model is to distinguish between the interactions with sisters versus the brothers. Table 4.5 shows the estimation of school dropout linked to the value of the groups aggregate behavior. To estimate these models the sample size is dropped to represent the teenagers who have both teenage brothers and sisters in the family. The impact of dropout bothers seems to be larger than that of dropout older sisters, although small sample size inflates the estimation errors and reduces the significance of the estimates.

Table 4.5. Estimation of school dropout, the effect of gender composition of the sibship social effect

	Simple			Weighted		
Concurrent						
Siblings social effect	1.803			2.81		
	6.23			7.43		
Brothers social efect		1.049			1.638	
		3.36			3.41	
Sisters social effect			1.952			2.918
			4.83			5.74
Lagged						
Siblings social effect	1.617			1.874		
	5.86			5.38		
Brothers social efect		1.326			2.074	
		4.25			4.17	
Sisters social effect			3.674			3.981
			3.36			2.49
N	8749	5033	2379	8749	5033	2379

The models reported are random effects logistic regressions.

t-values reported beneath each estimate.

The same set of covariates as in table 5.4 are used in the regressions but not reported in the table.

Another interesting sub-sample is the group of teenagers who have more than one sibling. In these type of families the effect of siblings is not determined in a one to one fashion, as the teenager is exposed to the behavior of at least 2 other siblings who might show consistent, or conflicting outcome. The top part of table 4.6 contains the result of estimation in the subsample of families with 3 or more teenage kids. The behavior complementarity seems to be stronger in the families where the teenager is subject to more than one source of sibling behavior.

Table 4.6. Estimation of school dropout; Sub-sample of teens who have at least 2 teenage siblings

	Sub-sample of families with 3 or more teenage kids			
	(2')	(3')	(4')	(5')
social	2.333			
	4.2			
socialw		3.573		
		5.05		
L.social			2.573	
			4.91	
L.socialw				3.017
				4.5
ll	-400.33	-394.715	-396.58	-398.038
N	1728	1728	1728	1728

	The whole sample			
	(2)	(3)	(4)	(5)
social	1.803			
	6.23			
socialw		2.81		
		7.43		
L.social			1.617	
			5.86	
L.socialw				1.874
				5.38
ll	-1312.804	-1302.031	-1314.702	-1316.189
N	8749	8749	8749	8749

Finally the models in table 4.7 compares the results of the previous models with the ones that include the Family Fixed effect variables are added to the regressions. The bottom panel of table this table is a copy from table 4.4 for comparison. The top panel of table 4.5.7 shows that after controlling for the family fixed effects, the estimates of the social utility term are still positive

Table 4.7. Estimation of school dropout, using family fixed effects

	Models with Family Fixed Effects			
	(2")	(3")	(4")	(5")
social	0.451			
	1.944			
social-w		0.679		
		2.57		
Lag(social)			0.58	
			2.301	
Lag(social-w)				0.522
				1.865
N	9862	9862	8749	8749
log-likelihood	-1122.533	-1121.558	-986.6	-988.572

	Models without Family fixed effect			
	(2)	(3)	(4)	(5)
social	1.803			
	6.23			
social-w		2.81		
		7.43		
Lag(social)			1.617	
			5.86	
Lag(social-w)				1.874
				5.38
N	8749	8749	8749	8749
log-likelihood	-1312.804	-1302.031	-1314.702	-1316.189

The models reported are paneldata logistic regressions.
t-values reported beneath each estimate.

and significant, although reduced in magnitude. This suggests the presence of siblings conformity in the framework of within family peer pressure, even after controlling for the contextual effects.

4.6 Conclusion

The estimation of school dropout through perceived sibling behaviors shows the signs of significant peer effects. An improved set of social interaction variable were introduced to the conventional discrete choice model of teens' decision about school dropout, considering the teens in the same family as members of a small social group. Using these weighted group behavior, the models enables us to address the non uniformity of the peer effect that is coming from different agents who are in contact with the teenager within the family.

To ensure the exogeneity of the siblings group effect, I applied a combination of two strategies to identify of the peer group behavior from the Manski's contextual effects. The models were estimated using the lagged aggregate outcome of the siblings, which reduces the problem of simultaneity of the choices of the siblings that could be due to the unobserved factors affecting both. The estimated results using this method, which purges the contextual effects out of the estimation, indicates that after keeping the individual and family characteristics of the teen constant, there is still positive and significant increase in the odds of school dropout event when the teens are subject to dropout siblings in the household.

Using an additional method to identify the siblings effects, the robustness of the results was confirmed through the use of household fixed effects estimations. This family fixed effects model predicts the effect of the weighted average of dropout siblings in a family is an increased odds of dropout vs graduation/enrollment by a factor of 1.68, equivalent to 0.24 percentage points increase in the probability of teen's dropout.

Appendix A (Appendix to Chapter 2)

Table A.1. Effect of birth order in various family sizes on dropout outcome of later-borns, Logistic regresssion results with all covariate

	Two-kid families			Three-kid families			Four-kid families			Five-plus-kid families		
	All	Males	Females	All	Males	Females	All	Males	Females	All	Males	Females
secondborn	0.225***	0.408***	-0.036	0.117*	0.063	0.191*	-0.004	0.165	-0.100	0.212**	0.202	0.311**
	(0.051)	(0.066)	(0.081)	(0.056)	(0.079)	(0.082)	(0.068)	(0.092)	(0.108)	(0.076)	(0.104)	(0.115)
thirdborn				0.341***	0.449***	0.230*	-0.122	-0.196*	0.029	0.183*	0.049	0.379**
				(0.061)	(0.083)	(0.094)	(0.075)	(0.099)	(0.118)	(0.081)	(0.113)	(0.118)
fourthborn							0.062	0.024	0.111	0.327***	0.129	0.509***
							(0.086)	(0.115)	(0.135)	(0.081)	(0.112)	(0.119)
fifthborn										0.347***	0.295**	0.441***
										(0.073)	(0.102)	(0.108)
sex	-0.490***			-0.224***			-0.392***			-0.424***		
	(0.051)			(0.048)			(0.055)			(0.049)		
age12				-2.674***	-2.133**		-2.660***		-1.893*			
				(0.733)	(0.757)		(0.741)		(0.775)			
age13	-2.091***	-2.424***	-1.597***	-1.666***	-1.967***	-1.410***	-2.208***	-2.163***	-2.156***	-2.173***	-1.803***	-2.820***
	(0.308)	(0.444)	(0.432)	(0.273)	(0.401)	(0.378)	(0.352)	(0.461)	(0.551)	(0.302)	(0.373)	(0.540)
age14	-1.861***	-2.193***	-1.385***	-1.665***	-1.823***	-1.537***	-2.585***	-3.225***	-1.951***	-1.976***	-2.045***	-1.977***
	(0.220)	(0.313)	(0.315)	(0.209)	(0.295)	(0.301)	(0.316)	(0.529)	(0.405)	(0.218)	(0.315)	(0.302)
age15	-1.332***	-1.422***	-1.140***	-1.331***	-1.330***	-1.405***	-1.589***	-1.659***	-1.459***	-1.472***	-1.458***	-1.554***
	(0.157)	(0.205)	(0.247)	(0.157)	(0.212)	(0.236)	(0.186)	(0.246)	(0.287)	(0.156)	(0.218)	(0.227)
age16	-0.663***	-0.800***	-0.404*	-0.572***	-0.719***	-0.439**	-0.818***	-0.946***	-0.598**	-0.796***	-0.874***	-0.762***
	(0.119)	(0.157)	(0.186)	(0.115)	(0.161)	(0.166)	(0.134)	(0.179)	(0.206)	(0.120)	(0.169)	(0.171)
age17	-0.208*	-0.274*	-0.057	-0.221*	-0.324*	-0.120	-0.276*	-0.352*	-0.136	-0.380***	-0.374**	-0.409**
	(0.100)	(0.130)	(0.157)	(0.096)	(0.134)	(0.142)	(0.108)	(0.144)	(0.170)	(0.101)	(0.140)	(0.147)
age18	0.056	0.030	0.129	-0.026	-0.004	-0.052	-0.001	-0.076	0.146	-0.028	-0.003	-0.079
	(0.087)	(0.113)	(0.140)	(0.085)	(0.115)	(0.127)	(0.094)	(0.124)	(0.149)	(0.087)	(0.121)	(0.128)
black	0.276***	0.435***	0.062	0.115	0.578***	-0.505***	0.182**	0.543***	-0.329**	0.398***	0.671***	0.102
	(0.061)	(0.079)	(0.098)	(0.059)	(0.079)	(0.095)	(0.067)	(0.087)	(0.109)	(0.063)	(0.087)	(0.091)
hispanic_nomiss	0.603***	0.732***	0.381**	0.123	0.276*	-0.081	0.493***	0.723***	0.155	0.397***	0.610***	0.243*
	(0.075)	(0.097)	(0.121)	(0.075)	(0.111)	(0.105)	(0.077)	(0.103)	(0.118)	(0.074)	(0.104)	(0.109)
raceother	0.147	0.044	0.365**	0.303***	0.378**	0.202	-0.177*	-0.094	-0.344**	-0.020	0.073	-0.162
	(0.084)	(0.111)	(0.130)	(0.078)	(0.116)	(0.109)	(0.082)	(0.108)	(0.130)	(0.076)	(0.105)	(0.114)
mom_age_nomiss	-0.007	-0.007	-0.008	-0.006	0.010	-0.024**	-0.022***	-0.004	-0.048***	-0.003	0.001	-0.006
	(0.005)	(0.006)	(0.008)	(0.005)	(0.007)	(0.007)	(0.005)	(0.007)	(0.009)	(0.005)	(0.006)	(0.007)
mi_mom_age	-0.235	-0.499*	0.061	0.028	0.294	-0.313	-0.964***	-0.944***	-1.096***	0.007	0.306	-0.262
	(0.163)	(0.220)	(0.252)	(0.151)	(0.208)	(0.227)	(0.177)	(0.242)	(0.278)	(0.151)	(0.208)	(0.230)
pov125	0.812***	0.678***	0.989***	0.849***	0.632***	1.070***	0.842***	0.615***	1.117***	0.689***	0.520***	0.906***
	(0.057)	(0.076)	(0.087)	(0.053)	(0.075)	(0.078)	(0.061)	(0.081)	(0.096)	(0.054)	(0.075)	(0.081)
pov400	-1.056***	-0.880***	-1.474***	-0.976***	-0.907***	-1.116***	-0.893***	-0.779***	-1.283***	-1.056***	-0.972***	-1.262***
	(0.080)	(0.094)	(0.158)	(0.081)	(0.103)	(0.132)	(0.105)	(0.123)	(0.211)	(0.119)	(0.145)	(0.215)
mi_pov	0.309**	0.302*	0.348	0.086	0.096	-0.063	0.606***	0.615***	0.423*	0.499***	0.284	0.715***
	(0.116)	(0.152)	(0.181)	(0.114)	(0.147)	(0.187)	(0.116)	(0.143)	(0.208)	(0.111)	(0.154)	(0.163)
urban	-0.042	-0.090	0.061	-0.035	-0.164	0.154	-0.296***	-0.304***	-0.282*	0.041	-0.121	0.204*
	(0.065)	(0.084)	(0.102)	(0.062)	(0.084)	(0.092)	(0.068)	(0.088)	(0.110)	(0.066)	(0.088)	(0.103)
mi_urban	-0.008	0.011	-0.050	-0.211	-0.286	-0.187	-0.218	-0.263	-0.153	-0.089	-0.029	-0.200
	(0.151)	(0.191)	(0.250)	(0.146)	(0.194)	(0.227)	(0.151)	(0.200)	(0.236)	(0.135)	(0.171)	(0.226)
hgc_res_dad_nomiss	0.005	-0.004	0.018	0.018	-0.011	0.051**	-0.025*	0.005	-0.055**	0.060***	0.092***	0.024
	(0.011)	(0.015)	(0.019)	(0.011)	(0.016)	(0.017)	(0.012)	(0.016)	(0.018)	(0.012)	(0.017)	(0.017)
mi_hgs_res_dad	-0.080	-0.276	0.187	0.393**	-0.001	0.817***	-0.444**	0.095	-1.024***	0.998***	1.299***	0.637**
	(0.153)	(0.195)	(0.252)	(0.149)	(0.208)	(0.223)	(0.157)	(0.218)	(0.239)	(0.155)	(0.219)	(0.220)
hgc_res_mom_nomiss	0.018	0.023	0.018	-0.002	0.003	-0.004	-0.004	0.028	-0.036*	-0.032**	-0.048**	-0.013
	(0.011)	(0.015)	(0.018)	(0.010)	(0.015)	(0.015)	(0.011)	(0.015)	(0.017)	(0.010)	(0.015)	(0.015)
mi_hgs_res_mom	0.149	0.201	0.217	-0.057	-0.011	0.048	-0.206	0.385	-0.904***	-0.916***	-1.141***	-0.632**
	(0.162)	(0.217)	(0.249)	(0.150)	(0.211)	(0.220)	(0.159)	(0.217)	(0.254)	(0.155)	(0.221)	(0.222)
live_w_both_nomiss	-0.125*	-0.043	-0.218*	0.000	-0.136	0.176*	0.098	0.001	0.124	-0.048	-0.021	-0.067
	(0.060)	(0.081)	(0.092)	(0.058)	(0.078)	(0.086)	(0.064)	(0.083)	(0.107)	(0.060)	(0.084)	(0.086)
mi_live_w	-0.085	0.062	-0.243	0.123	0.231*	-0.082	0.178	0.053	0.351*	0.142	-0.341*	0.489***
	(0.088)	(0.117)	(0.137)	(0.084)	(0.111)	(0.133)	(0.097)	(0.131)	(0.153)	(0.088)	(0.139)	(0.117)
hhsize_nomiss	0.206***	0.161***	0.262***	0.140***	0.092***	0.208***	0.073***	0.067***	0.082***	0.050***	0.041**	0.062***
	(0.017)	(0.023)	(0.025)	(0.015)	(0.021)	(0.022)	(0.015)	(0.019)	(0.024)	(0.011)	(0.016)	(0.016)
mi_hhsize	0.889***	0.912**	0.723	0.820***	0.614*	1.227**	0.874***	0.826**	0.885	0.806***	0.893**	0.433
	(0.261)	(0.313)	(0.494)	(0.235)	(0.302)	(0.377)	(0.260)	(0.309)	(0.501)	(0.243)	(0.301)	(0.439)
Constant	-3.275***	-3.205***	-4.011***	-3.022***	-2.880***	-3.572***	-1.166***	-2.471***	-0.023	-2.498***	-2.610***	-2.927***
	(0.221)	(0.285)	(0.351)	(0.211)	(0.285)	(0.317)	(0.231)	(0.313)	(0.364)	(0.222)	(0.310)	(0.323)
Observations	23,485	11,893	11,592	22,293	11,387	10,760	13,402	7,103	6,179	12,686	6,056	6,630

The results are derived from pooled logit regression models with cluster standard errors.

Standard errors in prantesis; significant levels are .10, .05, .01 indicated by *, **, ***

Included but not reported in the regression are dummy variables for years of survey.

Table A.2. School Dropout, considering the schooling decision of the sibling; subsample of two-kid families*

	All		Males		Females	
	(1)	(2)	(1)	(2)	(1)	(2)
Dropout oldesr sibling	0.788*	1.336***	0.615	0.886	1.042	2.043**
	(0.332)	(0.349)	(0.400)	(0.453)	(0.595)	(0.625)
Age difference between siblings	-0.232	-0.201	-0.197	0.086	-0.295	-0.823*
	(0.126)	(0.184)	(0.155)	(0.240)	(0.203)	(0.392)
Same-sex siblings pairs	-0.163	-0.351	0.008	-0.331	-0.362	-0.265
	(0.240)	(0.350)	(0.269)	(0.452)	(0.375)	(0.665)
Sex	-0.704**	-1.173***				
	(0.258)	(0.356)				
Age first dropped out	0.258***	0.214***	0.258***	0.217***	0.296***	0.217***
	(0.014)	(0.019)	(0.018)	(0.024)	(0.030)	(0.036)
Black	0.318	0.857*	0.946**	0.439	1.359**	1.776*
	(0.324)	(0.389)	(0.342)	(0.486)	(0.460)	(0.732)
Hispanic	-0.183	1.076*	0.656	1.517**	-0.492	0.576
	(0.345)	(0.424)	(0.370)	(0.571)	(0.629)	(0.828)
Other race	0.606	0.221	0.199	0.566	-1.300**	-0.105
	(0.362)	(0.523)	(0.436)	(0.687)	(0.461)	(1.033)
Low income household	0.192	1.075	0.123	0.706	1.642***	1.887
	(0.225)	(0.560)	(0.501)	(0.772)	(0.385)	(1.010)
High income household	-0.965***	0.606*	0.098	0.527	0.427	0.755
	(0.268)	(0.261)	(0.288)	(0.346)	(0.367)	(0.447)
Urban residence	0.078	-0.128	-0.722	-0.427	0.303	0.377
	(0.239)	(0.500)	(0.522)	(0.637)	(0.488)	(0.907)
Father's education	-0.044	0.539	0.574	0.731	0.047	-0.207
	(0.046)	(0.629)	(0.630)	(0.835)	(0.514)	(1.149)
Mother's education	0.057	-0.166	-0.666	-0.892	-0.471	0.796
	(0.052)	(1.017)	(0.680)	(1.243)	(1.194)	(2.099)
Mother's age at the birth of respondant	-0.078**	0.537	1.030	0.859	0.139	-0.345
	(0.028)	(0.620)	(0.541)	(0.858)	(0.603)	(1.145)
Observations	4,494	4,645	2,297	2,367	2,009	2,278
		519		258		248

(1) Pooled logit regression, clustered robust standard errors

(2) Random effects logit regression

Standard errors in prantesis; significant levels are .10, .05, .01 indicated by *, **, ***

Included but not reported in the regression are dummy variables for years of survey, and the dummy variables for the missing values on parents education, mothers age and family intactness.

Table A.3. Fixed effects estimation of the effect of dropout older sibling, on the dropping out

	(1)	(2)	(3)	(4)
Dropout oldesr sibling	1.106***	1.432***	0.665***	0.156
	(0.110)	(0.118)	(0.136)	(0.115)
Female	-0.207*	-0.380**		-0.057
	(0.097)	(0.140)		(0.134)
Have an older sibling	0.056	0.140*		0.090
	(0.040)	(0.056)		(0.069)
Age first dropped out	0.212***	0.173***	0.027***	0.100***
	(0.006)	(0.007)	(0.008)	(0.006)
age12	-1.581**	-2.697***	-4.099***	-3.192***
	(0.585)	(0.683)	(0.786)	(0.729)
age13	-0.974***	-2.119***	-3.502***	-2.304***
	(0.285)	(0.388)	(0.468)	(0.392)
age14	-1.054***	-2.261***	-3.617***	-2.414***
	(0.257)	(0.330)	(0.395)	(0.330)
age15	-0.445	-1.495***	-2.845***	-1.681***
	(0.228)	(0.270)	(0.324)	(0.268)
age16	0.068	-0.625**	-1.871***	-0.794***
	(0.192)	(0.223)	(0.269)	(0.222)
age17	0.349*	-0.002	-0.924***	-0.156
	(0.155)	(0.182)	(0.216)	(0.180)
age18	0.430***	0.341*	-0.151	0.230
	(0.110)	(0.148)	(0.168)	(0.145)
Black	0.152	0.794***		-1.766*
	(0.117)	(0.172)		(0.797)
Hispanic	0.014	0.439*		-0.425
	(0.148)	(0.213)		(1.191)
Other race	0.230	0.407		0.367
	(0.158)	(0.224)		(1.767)
Low income household	0.460***	0.531***	0.183	0.296**
	(0.083)	(0.099)	(0.114)	(0.097)
High income household	-0.749***	-0.721***	-0.228	-0.202
	(0.116)	(0.144)	(0.171)	(0.147)
Missing information on hosehold income	-0.002	-0.014	0.038	-0.089
	(0.151)	(0.184)	(0.207)	(0.177)
Urban residence	-0.005	-0.003	0.028	0.124
	(0.103)	(0.126)	(0.158)	(0.139)
Father's education	0.027	0.016	0.038	-0.002
	(0.021)	(0.032)	(0.024)	(0.040)
Mother's education	-0.028	-0.013		-0.138***
	(0.020)	(0.030)		(0.031)
Intact family	0.035	-0.114		0.036
	(0.119)	(0.166)		(0.171)
Mother's age at the birth of respondant	-0.035***	-0.040**		-0.027
	(0.010)	(0.014)		(0.014)
Household size	0.093***	0.086***		0.043*
	(0.019)	(0.022)		(0.022)
Constant	-3.673***	-5.087***		
	(0.382)	(0.553)		
Number of observations	20,702	20,702	5,709	7,034
Number of groups		2,308	631	546

(1) Pooled logit regression, clustered robust standard errors
(2) Random effects logit regression
(3) Individual fixed effects regression
(4) Family fixed effects regression
Standard errors in prantesis; significant levels are .10, .05, .01 indicated by *, **, ***
Included but not reported in the regression are dummy variables for years of survey, and the dummy variables for the missing values on parents education, mothers age and family intactness.

Appendix B (Appendix to Chapter 3)

Table B.1. Kaplan Meijer estimation of the survivor function of school dropout

Time(Age)	At risk	Fail	Censored	Survivor Function	s.e.	95% C.I.	
Those without older dropout siblings							
12	3416	1	0	0.9997	0.0003	0.9979	1
13	3415	18	0	0.9944	0.0013	0.9913	0.9964
14	3397	36	0	0.9839	0.0022	0.9791	0.9876
15	3361	68	0	0.964	0.0032	0.9572	0.9697
16	3293	165	0	0.9157	0.0048	0.9059	0.9245
17	3128	212	0	0.8536	0.006	0.8413	0.8651
18	2916	151	0	0.8094	0.0067	0.7959	0.8222
19	2765	59	0	0.7922	0.0069	0.7782	0.8054
20	2706	22	343	0.7857	0.007	0.7716	0.7991
21	2341	8	462	0.783	0.0071	0.7688	0.7965
22	1871	8	514	0.7797	0.0071	0.7653	0.7933
23	1349	5	587	0.7768	0.0072	0.7623	0.7906
24	757	3	588	0.7737	0.0074	0.7588	0.7878
25	166	0	162	0.7737	0.0074	0.7588	0.7878
26	4	0	4	0.7737	0.0074	0.7588	0.7878
Those with older dropout siblings							
12	619	3	0	0.9952	0.0028	0.985	0.9984
13	616	20	0	0.9628	0.0076	0.9446	0.9752
14	596	22	0	0.9273	0.0104	0.9038	0.9452
15	574	64	0	0.8239	0.0153	0.7915	0.8517
16	510	83	0	0.6898	0.0186	0.6518	0.7246
17	427	74	0	0.5703	0.0199	0.5303	0.6082
18	353	60	0	0.4733	0.0201	0.4335	0.5121
19	293	15	0	0.4491	0.02	0.4096	0.4878
20	278	10	71	0.433	0.0199	0.3936	0.4716
21	197	2	81	0.4286	0.02	0.3892	0.4673
22	114	0	67	0.4286	0.02	0.3892	0.4673
23	47	0	35	0.4286	0.02	0.3892	0.4673
24	12	0	12	0.4286	0.02	0.3892	0.4673

Table B.2. Nelson-Aalen estimation of the cumulative hazard of school dropout

Time(Age)	At risk	Fail	Censored	Cum. Hazard	s.e.	95% C.I.	
Those without older dropout siblings							
12	3416	1	0	0.0003	0.0003	0	0.0021
13	3415	18	0	0.0056	0.0013	0.0035	0.0087
14	3397	36	0	0.0162	0.0022	0.0124	0.021
15	3361	68	0	0.0364	0.0033	0.0305	0.0434
16	3293	165	0	0.0865	0.0051	0.0771	0.0971
17	3128	212	0	0.1543	0.0069	0.1413	0.1684
18	2916	151	0	0.2061	0.0081	0.1908	0.2225
19	2765	59	0	0.2274	0.0086	0.2112	0.2448
20	2706	22	343	0.2355	0.0087	0.219	0.2533
21	2341	8	462	0.2389	0.0088	0.2223	0.2568
22	1871	8	514	0.2432	0.0089	0.2263	0.2614
23	1349	5	587	0.2469	0.0091	0.2297	0.2654
24	757	3	588	0.2509	0.0094	0.2332	0.2699
25	166	0	162	0.2509	0.0094	0.2332	0.2699
26	4	0	4	0.2509	0.0094	0.2332	0.2699
Those with older dropout siblings							
12	619	3	0	0.0048	0.0028	0.0016	0.015
13	616	20	0	0.0373	0.0078	0.0248	0.0562
14	596	22	0	0.0742	0.0111	0.0554	0.0994
15	574	64	0	0.1857	0.0178	0.1539	0.2241
16	510	83	0	0.3485	0.0252	0.3024	0.4016
17	427	74	0	0.5218	0.0323	0.4622	0.589
18	353	60	0	0.6917	0.039	0.6193	0.7726
19	293	15	0	0.7429	0.0412	0.6664	0.8283
20	278	10	71	0.7789	0.0427	0.6995	0.8674
21	197	2	81	0.7891	0.0433	0.7085	0.8788
22	114	0	67	0.7891	0.0433	0.7085	0.8788
23	47	0	35	0.7891	0.0433	0.7085	0.8788
24	12	0	12	0.7891	0.0433	0.7085	0.8788

Table B.3. Cox Proportional Hazard Model: The effect of sibship and other covariate on the age of first school dropping out

	(1)	(2)	(3)	(4)
Parents highest grade of education	-0.011	-0.010	-0.010	-0.010
	(0.013)	(0.013)	(0.013)	(0.013)
Father is a dropout	0.029	0.032	0.035	0.035
	(0.078)	(0.078)	(0.078)	(0.078)
Mother is a dropout	-0.118	-0.115	-0.115	-0.115
	(0.084)	(0.084)	(0.084)	(0.084)
Mother's age at birth of the teen	-0.000	-0.000	-0.000	-0.000
	(0.005)	(0.005)	(0.005)	(0.005)
Intact family at age 14	-0.025	-0.027	-0.024	-0.024
	(0.062)	(0.062)	(0.062)	(0.062)
Low income household	0.724	0.722	0.720	0.721
	(0.069)	(0.069)	(0.069)	(0.069)
High income household	-0.937	-0.937	-0.927	-0.927
	(0.124)	(0.124)	(0.124)	(0.124)
Missing income info	0.073	0.068	0.070	0.071
	(0.071)	(0.071)	(0.071)	(0.071)
Household size	0.056	0.049	0.038	0.038
	(0.017)	(0.017)	(0.018)	(0.018)
Have an older sibling	0.161			
	(0.056)			
Have an older brother		0.181		
		(0.056)		
Have an older sister		0.114		
		(0.058)		
Older sibling within 3 years			0.190	
			(0.045)	
Older sibling older than 3 years			0.054	
			(0.024)	
Older brother within 3 years				0.186
				(0.071)
Older brother older than 3 years				0.042
				(0.041)
Older sister within 3 years				0.149
				(0.061)
Older sister older than 3 years				0.066
				(0.041)
Observations	5,817		5,817	5,817

(1)-(4) share the same covariate, but each include a different set of sibship covariates.

Cox model is stratified using gender, age and race dummies.

Standard errors in prantheses.

Table B.4. Parametric estimation of the shared failty models of dropout, Gompertz, Log normal, and Log logistic distributions

	Gompertz Hazard			Log normal Hazard		Log logistic Hazard		
	Standard	Gamma freilty	Inv. Gaussian Frailty	Standard	Gamma freilty	Standard	Gamma freilty	Inv. Gaussian Frailty
Have a dropout older sibling	1.023	0.117	0.249	-0.255	0.033	-0.266	0.029	0.026
	11.07	0.66	1.52	-10.08	1.13	-10.75	0.8	3.77
Have an older sibling	0.091	0.273	0.218	-0.011	-0.036	-0.018	-0.048	-0.005
	0.95	2.23	1.82	-0.48	-1.75	-0.77	-2.13	-0.69
Parents highest grade of education	-0.007	-0.013	-0.014	0.001	0.002	0.001	0.002	0.001
	-0.38	-0.53	-0.57	0.22	0.4	0.29	0.37	0.69
Father is a dropout	0.011	0.010	-0.011	-0.001	-0.001	-0.003	-0.006	-0.007
	0.1	0.07	-0.08	-0.04	-0.03	-0.09	-0.21	-0.75
Mother is a dropout	-0.210	-0.175	-0.185	0.030	0.009	0.039	0.020	0.007
	-1.79	-1.13	-1.2	1.05	0.32	1.32	0.68	0.75
Mother's age at birth of the teen	0.003	0.004	0.005	-0.001	-0.001	-0.001	-0.001	0.000
	0.5	0.5	0.54	-0.73	-0.59	-0.7	-0.6	0.46
Intact family at age 14	-0.108	-0.025	-0.014	0.025	-0.002	0.031	0.009	0.009
	-1.27	-0.22	-0.13	1.19	-0.09	1.44	0.4	1.36
Low income household	0.802	1.153	1.132	-0.193	-0.117	-0.205	-0.178	0.012
	7.87	7.22	7.29	-7.59	-3.62	-7.94	-4.74	1.41
High income household	-0.556	-0.806	-0.750	0.108	0.151	0.127	0.189	0.026
	-2.93	-3.42	-3.16	2.93	3.95	2.98	4.12	2.23
Missing income info	0.294	0.316	0.318	-0.058	-0.014	-0.068	-0.043	0.013
	2.79	2.08	2.13	-2.35	-0.49	-2.65	-1.28	1.41
Household size	0.049	0.061	0.077	-0.010	-0.005	-0.010	-0.007	0.000
	2.1	1.62	2.03	-1.74	-0.81	-1.74	-0.93	-0.02
Urban residence	0.025	-0.022	0.008	-0.001	0.012	-0.004	0.012	-0.006
	0.26	-0.15	0.06	-0.03	0.47	-0.16	0.42	-0.74
Female	-0.356	-0.428	-0.438	0.079	0.056	0.086	0.070	0.006
	-4.580	-4.410	-4.520	4.15	3.25	4.43	3.77	0.990
Age	-0.124	-0.215	-0.220	0.029	0.048	0.027	0.046	0.058
	-4.02	-5.22	-5.38	4.11	6.84	3.63	5.75	27.58
Black	0.368	0.581	0.602	-0.084	-0.067	-0.091	-0.100	0.016
	3.89	3.93	4.15	-3.64	-2.53	-3.85	-3.36	2.18
Hispanic	0.170	0.330	0.279	-0.040	-0.051	-0.040	-0.064	0.023
	1.380	1.790	1.550	-1.320	-1.550	-1.320	-1.690	2.400
Other race	0.169	0.164	0.220	-0.040	-0.022	-0.043	-0.030	-0.025
	1.32	0.83	1.14	-1.24	-0.63	-1.33	-0.75	-2.32
_cons	-6.146	-5.669	-5.690	3.056	2.488	3.083	2.555	1.818
	-10.79	-7.5	-7.63	22.35	16.6	22.13	15.18	44.36
Frailty variance (theta)		2.111	3.332		3.334		3.041	529.6
		4.08	4.35		6.72		5.42	42.64

Values of test statistic is reported below each estimate.

The likelihood estimation for the choice of inverse gaussian frailty in the Log normal model was not convergant.

References

Akerlof, George A. “Social Distance and Social Decisions.” Econometrica, 1997, 65(5), Pages 1005–27

Ammermueller A and Pischke J-S (2009), “Peer effects in European Primary Schools: Evidence from the Progress in International Reading Literacy Study.” Journal of Labor Economics 27(3), Pages 315-48.

Argys, Laura M, Daniel I. Rees, Susan L. Averett and Benjaman Witoonchart, (2006), “Birth Order and Risky Adolescent Behavior” Economic Inquiry Vol. 44, No 2, Pages 215-233.

Becker. Gary S., and H. Gregg Lewis, (1973). “On the Interaction Between the Quantity and Quality of Children,” Journal of Political Economy, N. 81, Pages 278-S88.

Becker, Gary S. and Nigel Tomes (1976) “Child Endowments and the Quantity and Quality of Children,” Journal of Political Economy.

Becker, Gary S. (1981). “A treatise on the family” Cambridge, MA: Harvard University Press

Behrman, Jere R., Robert A. Pollack, and Paul Taubman, (1982). "Parental Preferences and Provision for Progeny," Journal of Political Economy, N. 90, V.1, Pages 52-73.

Behrman, J. R. and Paul Taubman, (1986), “Birth Order, Schooling, and Earnings.” Journal of Labor Economics 4: Pages 121-145.

Birdsall, Nancy (1991) “Birth Order Effects and Time Allocation,” in Research in Population Economic , Volume 7, ed. T Paul Schultz Pages 191-217.

Bjorn, P. and Q. Vuong (1984), “Simultaneous Models for Dummy Endogenous Variables: a Game Theoretic Formulation with an Application to Household Labor Force Participation,” Working Paper, California Institute of Technlogy.

Black, Sandra E., Deverux, Paul J., and Kjell G. Salvanes (2005) “The More the Merries? The Effect if Family Size and Birth Order on Children’s Education”, Quarterly Journal of Economics, Pages 669-700.

Brock, W. A. and Durlauf, S. N. (2001a), “Discrete choice with social interactions.” Review of Economic Studies 68(2): Pages 235–260.

Brock, W. A. and Durlauf, S. N. (2001b), “Interactions-based models.” Handbook of Econometrics; Vol. 5. North Holland, Amsterdam, Pages 3297–3380.

Butcher K. F. and A. Case (1994) “The effect of sibling sex composition on women’s education and earnings” , Quarterly Journal of Economics, CIX (3), Pages 531-563

Card, D. and T. Lemieux (2001) “Dropout and Enrollment Trends in the Post-War Period: What Went Wrong in the 1970s?" In Risky Behavior Among Youths : An Economic Analysis, edited by J. Gruber. Chicago: University of Chicago Press.

Cox, D. R. (1972), "Regression Models and Life Tables" (with discussion), Journal of the Royal Statistical Society, 34, Pages 187-220.

Duncan, GJ , Boisjoly, J, and KM Harris (2001), “Sibling, peer, neighbor, and schoolmate correlations as indicators of the importance of context for adolescent development.” Demography, V 38, N 3, Pages 437-447

Durlauf S. (1999),“The Membership Theory of Inequality: Ideas and Implications.” Elites, minorities, and economic growth, Pages 161-77.

Gaviria, A. and Raphael, S. (2001), “School-based peer effects and juvenile behavior.” The Review of Economics and Statistics V. 83, Pages 257–268.

Grant B. F. , Stinson F. S. and , T. C. Harford (2001), “Age at onset of alcohol use and DSM-IV alcohol abuse and dependence: a 12-year follow up study” Journal of Substance Abuse.

Gruber J. (2000), “Risky Behavior among Youth: An Economic Analysis”, NBER working paper, W7781

Guo, G., and Rodriguez, G. (1992), "Estimating a Multivariate Proportional Hazards Model for Clustered Data Using the EM Algorithm, With an Application to Child Survival in Guatemala," Journal of the American Statistical Association, 87, Pages 969-976.

Kaestner, Robert, (1997), “Are Brothers Really Better? Sibling Sex Composition and Educational Achievement Revisited” The Journal of Human Resources, Vol. 32, No. 2, Pages 250-284

Kessler, Daniel (1991), “Birth order, family size and achievement: family structure and wage determination” , Journal of Labor Economics, Vol. 9, No 9. , Pages 413-26

Kosterman, R., J.D. Hawkins, Guo, J. Catalano, R. E. Abbott, R. D. (2000) The dynamics of alcohol and marijuana initiation: patterns and predictors of first use in adolescence. Am J Public Health 90, Pages 360–366.

Hanna, E. Z., Hsiao-ye, Y., Dufour, M. C., & Whitmore, C. C. (2001). The relationship of early onset regular smoking to alcohol use, depression, illicit drug use, and other risky behaviors during early adolescence: Results from the youth supplement to the Third National Health and Nutrition Examination Survey. Journal of Substance Abuse, 13, Pages 265–282.

Haurin, R.J. and F.L. Mott. 1990. "Adolescent Sexual Activity in the Family Context: The Impact of Older Siblings." Demography 27: Pages 537-57.

Heckman, J.J. (1978), “Dummy Endogenous Variables in a Simultaneous Equation System”, Econometrica, V. 46, Pages 931-960.

Huang, C.C. and Wang, P. (2004) “Crime and Poverty" International Economic Review, 45 (3): Pages 909-938.

Hoxby C. M, (2000), “Peer effects in the classroom: Learning from gender and race variation.” NBER working paper No. W7867.

Israel, GD, Beaulieu, LJ, & Hartless, G. (2001). The influence of family and community social capital on educational achievement. Rural Sociology, 66(1). Pages 43-68

Kooreman, P. and Soetevent, A.R. (2007). “A discrete choice model with social interactions: an application to high school teen behaviour.” Journal of Applied Econometrics, V. 22, N. 3, Pages 599-624.

Laing, Derek (2009) Crime, book chapter in Principles of Modern Labor Economics, forthcoming, Summer 2009, Norton & Co., New York, NY.

Levitt, Steven D. (1998) “Juvenile Crime and Punishment”, Journal of Political Economy, Vol 106, no 6, Pages 1156-1185

Lee, L. (2007). “Identification and estimation of econometric models with group interactions, contextual factors and fixed effects.” Journal of Econometrics, V.140, N.2, Pages 333-374.

Manski, C. F. (1993), “Identification of endogeneous social effects: the reflection problem.” Review of Economic Studies 60: Pages 531–542.

Manski, C. F. (2000). “Economic analysis of social interactions.” Journal of Economic Perspectives V. 14, Pages 115–136.

Nagin, D., & Farrington, D. P. (1992a). The onset and persistence of offending. Crirninology,30, Pages 501-523.

Oettinger, Gerald S., (2000), “Sibling Similarity in High School Graduation Outcomes: Causal Interdependency or Unobserved Heterogeneity?” Southern Economic Journal 66 (3), Pages 631-648.

Ouyang, Lijing (2004), “Sibling Effect on teen risky behavior”, Duke university, working paper.

Pollak, R.A., 1976. “Interdependent preferences.” American Economic Review, V. 66 , N. 3, Pages 309–320.

Powers, Daniel A. and Hsueh, James Cherng-tay (1997) “Sibling Models of Socioeconomic Effects on the Timing of First Premarital Birth” Demography, Vol. 34, No. 4, Pages 493-511

Rodgers J., Rowe, D. and Harris, F., (1992), “Sibling differences in adolescent sexual behavior, inferring process model from family composition patterns.” Journal of marriage and family, 54, Pages 144-52.

Rose, Elaina (2007) “Siblings and Soldiers: Family Background and Military Service in the All-Volunteer Era”, mimeo, University of Washington.

Rodriguez, G. (1994), "Statistical Issues in the Analysis of Reproductive Histories Using Hazard Models," Annals of the New York Academy of Sciences, 709, Pages 266-279.

Tolan, P. H., & Thomas, P. (1995). The implications of age of onset for delinquency risk. Journal of Abnormal Child Psychology, 23, Pages 157–181.

Singer, Judith D. and John B. Willett (1994), “Designing and Analyzing Studies of Onset, Cessation, and Relapse: Using Survival Analysis in Drug Abuse Prevention Research”, In L. M. Collins, & L. A. Seitz, (Eds), Advances in data analysis for prevention intervention research (NIDA Research Monograph No. 142, Pages. 196–263). Rockville, MD: National Institute on Drug Abuse.

Sacerdote, B. (2001). “Peer effects with random assignment: results for Dartmouth roommates.” The Quarterly Journal of Economics, V. 116, Pages 681–703.

Soetevent, A. R. and Kooreman, P. (2007). “A discrete choice model with social interactions; with an application to high school teen behavior.” Journal of Applied Econometrics, V. 22 , N. 3, Pages 599 - 624

Steinberg, L. and Caufman, E. (1996), “Maturity of Judgment in Adolescence: Psychological Factors in Adolescent Decision Making”, Law and Human Behavior, N. 20, Pages 249-272.

Sulloway, Frank J. (1996), “Born to rebel: Birth order, Family dynamic, and Creative lives.”, New York, Pantheon Books.

Vaupel, J. W., K. Manton, and E. Stallard. 1979. The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography 16: Pages 439–54.

Wu, L. , Schlenger, W., and D. Galvin (2003), “The relationship between employment and substance use among students aged 12 to 17” Journal of Adolescent Health, Volume 32, Issue 1, Pages 5-15

Zimmer, RW, Toma, EF, (2000) “Peer effects in private and public schools across countries.” Journal of Policy Analysis and Management. V. 19, N. 1, Pages 75-92.