投稿論文

Duration Analysis of High School Dropout Risks in the United States:

An Application of Survival Analysis to U.S. High School Student Data

Satoshi P. WATANABE※

高校生の就学意思とドロップアウトリスク要因に関する研究：米国のパネルデータを用いた生存分析の応用

渡邉　聡

本研究の目的は，無作為に抽出されたアメリカの高校生のパネルデータに生存分析（Survival Analysis）とよばれる統計手法を応用し，高校生のドロップアウト・リスクの要因とそのタイミングを明らかにすることである。計測された累積ハザードから，アメリカの高校生のドロップアウト・リスクは毎年6月に最も高くるなることが分かる。またベースライン・ハザード関数がワイブル（Weibull）分布に従うと仮定し，Cox比例ハザードモデルを推定した場合，ドロップアウト・リスクには正の期間依存（positive duration dependence）が存在する。アメリカにおける高校生のドロップアウト・リスクは，居住地域の失業率に対して負の相関を示し，高校卒業後に予測できる賃金や世帯所得に対しては正の相関を示している。しかし，これらのリスク要因がドロップアウトに与える影響は，最終学年に進級する時期から大きく減少しはじめる。

I. Introduction

Econometric analysis of duration data has been extensively studied by researchers in recent decades (Heckman and Singer 1984a, 1984b; Lancaster 1979). Han and Hausman (1990) and Sueyoshi (1992) studied the competing risks models of duration data. Simple

※筑波大学大学院ビジネス科学研究科
(Graduate School of Business Sciences, University of Tsukuba)

tests based on a score statistic for unobserved heterogeneity are discussed in Kiefer (1984) and Lancaster (1985a) when the heterogeneity is in the multiplicative form in the exponential and Weibull models, respectively. Other specification diagnostics are provided by Chesher (1984) and Kiefer (1985). Heckman and Singer (1984a), Lancaster and Nickell (1980) and Waldman (1985) discuss computational schemes with duration data in the presence of unobserved heterogeneity. The most applications are found in the studies of unemployment spells, particularly with respect to the impact of unemployment insurance, and the process of job search (Burdett et al. 1985; Lancaster 1985b; Lynch 1985, 1992; Meyer 1990; Moffitt 1985; Naredranathan and Nickell 1985). Others include Kennan (1985) who studied the duration of contract strikes in U.S. manufacturing, Dolton and Klaauw (1996) on salaries of U.K. teachers and their retention, and Kiefer (1985) on the role of education in labor turnover. A similar study on education and job turnover was done by Light and Omori (1996) using a competing risks model. Light (1996) also studies a hazard model of the decision to reenroll in school after leaving school for the first time.

This article uses the survival analysis approach with duration data on complete and incomplete enrollment lengths for U.S. high school students. It is important to correctly understand the factors that influence enrollment decisions by high school students and when and where during the process these factors affect their schooling failure because individuals without a high school diploma are and continue to be the most vulnerable in terms of economic success and nearly all aspects of labor activities once they leave school.[1] Not only do fewer years of formal educational training affect their life-time earnings, but also they influence workers' post-school labor activities followed by fewer opportunities to reinvest in further education and job-related training (Lynch 1992; Mincer 1988, 1993). Economic disadvantages experienced by these workers have cyclical and intergenerational impacts on their descendants by reducing their children's educational opportunities due to financial constraints and little exposure to academic environments (see, for example, Becker 1991). Though the proportion of U.S. adults with a high school diploma monotonically increased during the last century with slower increments in the 1990s as the proportion reached high figures, there still existed 12 percent of adults 25–29 years of age in 2000 who never completed high school education

(U.S. Department of Education 2002).

This study attempts to shed some light on the analysis of complete and incomplete duration data of high school enrollment lengths within the framework of a survival analysis approach. No prior studies exist that empirically analyze the conditional probabilities over time of high school dropout incidence in the hazard model framework, nor do we fully understand the relationship between students' background characteristics and the duration until enrollment failure. Although researchers have long studied the issues of schooling decisions made by young individuals (for example, Edwards 1975), dynamic models of educational attainment have rarely been considered.[2] Most previous studies concentrated on the examination of a single educational transition relying heavily on the limited discrete choice models, i.e., the logit or probit models, using various cross-sectional data or data set up accordingly. The duration data analysis allows for modeling time until failure by using the conditional probabilities of dropping out of school over time, given students remain enrolled up to a certain point in time. Censored observations and time dependent covariates are also handled with relative ease in this approach. This paper takes advantage of these analytical merits and examines the enrollment decision made by high school students in a more dynamic framework. It places a particular emphasis on the duration dependency of enrollment failure by high school students; that is, on examining whether the dropout risk for U.S. high school students increases or decreases as time elapses. It is recognized that the decision to remain enrolled in school goes hand-in-hand with the decision to work. The decision in turn is not independent of local labor demand from potential employers or the necessity to work due to financial difficulties faced by a student's family, as well as economic outcomes expected upon completing high school education. Therefore, a particular emphasis is also placed on estimating the impact of economic factors such as local labor demand, short-term credit constraints students and their family face, and economic incentives for students to complete high school, on the conditional probabilities of leaving high school without fulfilling the graduation requirements.

In analyzing the duration data on high school enrollment, it is an important step to carefully choose the correct duration distribution as a misspecified parametric model yields biased estimates of covariates and thus incorrect inferences on the effects of these

factors. However, this often is not a simple task as the shape of the hazard function or duration distribution may not explicitly be suggested by sound economic theory. We discuss this issue in Section III, which immediately follows the next section on the descriptions of the dataset used in this study.

II. DATA

The data used in this study is the 1979 cohort of *National Longitudinal Survey of Youth* (NLSY).[3] A subsample of the respondents who are in the 9th grade in 1979 is followed for the next three years until May of 1982 in which most respondents are expected to graduate from the high school in which they have been enrolled. Since NLSY is on-going panel data which contain rich information on the originally sampled 12,686 young individuals, it is possible to obtain the highest grade attained by these individuals beyond high school. However, for the purpose of the current study the duration data is trimmed or "censored" at 44 months, which represents May 1982. As all the students in the subsample enter the 9th grade in September 1978, most of them are assumed to complete their high school education sometime in the spring of 1982. About 40 percent of the incomplete cases or 278 individuals continued their education beyond 44 months. These individuals, however, are not necessarily enrolled in college or other forms of higher educational institutions, and 7 percent were still enrolled in high school after 44 months to complete their graduation requirements. This paper avoids the analysis on the lengths of continuous enrollment beyond high school as the transition from high school to college is not a simple extension of transiting from the 11th grade to the 12th (or equivalently from the 23rd month to the 24th), and the decision to enter college is influenced by factors that normally do not concern us when we simply consider continuous enrollment in secondary school. Moreover, the separation decision from college is likely affected by factors that are uncommon with those affecting the decision at a high school level, due to differences in the pattern of enrollment and influential environments, e.g., enrolled full-time or part-time, with work or without, and whether or not living away from home, etc. Therefore, rather than conducting the analysis with the data on the enrollment lengths at mixed levels, this paper focuses on the duration of high school enrollment up to 44 months for which the environments surrounding students are

considered to be more or less homogeneous. Moreover, our model is not that of schooling-transition, and we do not examine in this study the model of students' transiting from one grade to the next. Instead we attempt to cast some light on the econometric analysis of high school enrollment duration, regardless of grade repetition or transiting status, and identify some of the associated influential background characteristics.

The summary statistics of the variables used in the analysis are reported in Table 1 for 1,336 high school students who were in the 9th grade in 1979. Approximately one half (52 percent) of the sample consists of male students and 34 percent nonwhite. Students' average age increases linearly from 14.7 years in 1979 to 17.7 in 1982. Average AFQT score is 33 and the score is missing for 1 percent of the sample. Average highest grade completed by fathers (11.1 years) is slightly higher than that of mother's (10.9 years) with missing values for 15 percent and 7 percent of students, respectively. In 1979, 64 percent

Table 1 Descriptive statistics

Variable	1979 Mean (S.D.)	1980 Mean (S.D.)	1981 Mean (S.D.)	1982 Mean (S.D.)
1 if Male	0.52 (0.50)	———	———	———
1 if Nonwhite	0.34 (0.47)	———	———	———
Age [a, b]	14.69 (0.72)	15.65 (0.71)	16.64 (0.70)	17.65 (0.71)
AFQT [b]	32.90 (25.68)	———	———	———
1 if AFQT missing	0.01 (0.10)	———	———	———
Father's highest grade [b]	11.09 (3.69)	———	———	———
1 if father's grade missing	0.15 (0.36)	———	———	———
Mother's highest grade [b]	10.87 (2.93)	———	———	———
1 if mother's grade missing	0.07 (0.25)	———	———	———
Type of curriculum:				
1 if vocational [a]	0.08 (0.27)	0.10 (0.30)	0.12 (0.32)	0.12 (0.32)
1 if commercial [a]	0.01 (0.11)	0.02 (0.13)	0.02 (0.14)	0.02 (0.12)
1 if college preparatory [a]	0.23 (0.42)	0.28 (0.45)	0.27 (0.44)	0.29 (0.45)
1 if general program [a]	0.64 (0.48)	0.58 (0.49)	0.50 (0.50)	0.43 (0.50)
1 if curriculum unknown [a]	0.05 (0.21)	0.02 (0.15)	0.10 (0.29)	0.15 (0.36)
Unemployment rate (%):				
1 if < 6.0% [a]	0.51 (0.50)	0.32 (0.47)	0.21 (0.41)	0.07 (0.25)
1 if 6.0-8.9% [a]	0.39 (0.49)	0.51 (0.50)	0.45 (0.50)	0.31 (0.46)
1 if 9.0-11.9% [a]	0.05 (0.23)	0.11 (0.31)	0.24 (0.43)	0.38 (0.49)
1 if 12.0-14.9% [a]	0.02 (0.16)	0.05 (0.23)	0.06 (0.24)	0.12 (0.32)
1 if ≥ 15.0% [a]	0.00 (0.00)	0.00 (0.00)	0.03 (0.16)	0.12 (0.32)
1 if type unknown [a]	0.01 (0.12)	0.00 (0.06)	0.00 (0.07)	0.01 (0.10)
Net total family income [a, b]	15,881 (12,142)	18,539 (13,574)	19,935 (14,622)	21,989 (16,521)
1 if family income missing [a, b]	0.18 (0.38)	0.19 (0.39)	0.23 (0.42)	0.26 (0.44)
Sample size	1,336	1,336	1,336	1,336

[a] Denotes covariates that vary across one-year intervals, but are assumed constant within intervals. All other covariates are constant within and across intervals.

[b] Denotes continuous covariates, which enter hazards as deviations from sample mean.

of the high school students are enrolled in a general academic curriculum, followed by 23 percent in a college preparatory program and with only 8 percent and 1 percent enrolled in vocational and commercial curriculums, respectively. These proportions shifted somewhat from a general program towards other curriculum types during the following years. As the surveyed intervals fall in the recessionary period in the early 1980s, the unemployment rates for the labor market of students' local residence rose from 1979 to 1982. Finally, the average net total income for students' family increased from 15,881 dollars in 1979 to 21,989 dollars in 1982.

Of 1,336 high school students with the characteristics described in Table 1, we have 348 completed or unsuccessfully ended enrollment lengths. Among the 280 students for whom the reasons why they left school are available, approximately 10 percent reported that they are getting married or became pregnant (Table 2). About 19 percent reported they stopped going to school simply because they did not like it. Twelve percent had home responsibilities or financial difficulties, or chose to work with no detailed explanations. Fourteen percent of students quit high school due to poor grades or because they were expelled or suspended from school. A similar proportion (14 percent) chose to drop out as they received a GED, and 8 percent did not return to high school after they moved away from school. Finally, 23 percent reported other reasons with no further explanations.

Using the sample of these 1,336 high school students who were in the 9th grade in 1979, various specifications of the relative risk model described in Section IV are estimated

Table 2 Reasons respondents left high school

	Percent (Number of cases)
Getting married/pregnancy	10.4 (29)
Did not like school	18.9 (53)
Home responsibilities/financial difficulties/chose to work	11.8 (33)
Poor grades/expelled or suspended	14.3 (40)
Received a GED	13.9 (39)
Moved away from school	7.8 (22)
Other	22.9 (64)
Sample with reasons available	100.0 (280)

Note: The reasons are unavailable or unreported for 68 students.

to examine high school students' dropout behavior with particular interest in the duration dependency and the effects of demographic and family characteristics on the failure risk. Included variables in the model are the dummies for male and nonwhite. A continuous variable for student age is also included. Although the sample consists of mostly homogeneous students in terms of their age, the variable in the model is not time-homogeneous. Young individuals often learn gradually through experiences inside and outside school the outcomes of dropping out and the importance of staying in school. This aging effect on the hazards may be captured by this covariate. In addition, all the continuous variables in the model are entered as deviations from sample mean so that λ has an interpretation as the hazard for the mean individual in the sample.

Since the standardized test scores are not available in NLSY, a continuous but time constant variable for *Armed Forces Qualification Test* (AFQT) score is used as a proxy for individual ability. However, the score is not available for 1 percent of high school students in the sample. For these students, AFQT score is set to zero and the dummy for "AFQT missing = 1" is included. The predicted effect of students' ability on the school exit rate is indeterminate as the cost of being in school in terms of lost wages may be higher for those with high abilities which are likely correlated with a better distribution of wage offers. Students with higher abilities at the same time have lower cost in terms of effort levels they have to put in, and individuals with higher innate capacity are likely to succeed in school with less effort. The effect of students' ability on the hazard may be uncovered by this covariate. Parents are considered to offer significant support and encouragement for their child's school performance and play critical role models. The highest grade completed by both parents is included to control for the parental influences students may receive though adolescence. For students who did not provide this information, the highest grade completed is set to equal zero and the dummy variables are included for the missing cases. Students' conditional probability of leaving school is also assumed to depend on the skills they acquire in school. In order to capture the varying effects of curriculum type on finding jobs, the dummy variables are entered for vocational, commercial, college preparatory, with the general program as the omitted group. The dummy is also included for students whose curriculum type is unavailable.

This study places a particular emphasis on the impacts of three economic factors to

which high school students are considered to be receptive; (1) local labor demand in the market of student's residence, (2) financial resources available to students and students' family, and (3) economic incentive of completing high school. In order to capture the separate effects of labor demand, the ordinal dummies for unemployment rates (lower than 6.0 percent; 6.0–8.9 percent; 9.0–11.9 percent; 12.0–14.9 percent) are included with "higher than 15.0 percent" as the omitted category. The short-term credit constraints faced by students and their family are controlled by a continuous variable for the net total family income. The net total family income measures the total income in the individual's household from such sources as labor earnings (net of the respondent's), gifts, alimony, unemployment insurance, and public assistance programs. Financial difficulties proxied by this variable may be expected to have a negative impact on the enrollment lengths of students. Finally, the effect of students' incentive for completing high school is measured by the wage differentials under two alternative regimes, predicted at the time immediately following the potential duration of maximum enrollment lengths (44 months). The potential wage gains expected if one completes high school is considered to affect students' enrollment behavior on an assumption that high school students are myopic in the sense that they do not take into account the discounted lifetime earnings but only consider the immediate return to their high school education. Thus, the wage rates are computed for each student if he completes a full length of potential enrollment ($\ln \omega'$) and if he quits school before the 44th month and start working ($\ln \omega$), both estimated at an arbitrary time t shortly following 44 months. As a practical matter of computation, the wages under the two alternatives are predicted with the parameter estimates from two separate log-wage regressions using the subsamples of individuals who completed the full length and reported wages in 1982 and those who quit school and reported wages in the same year.[4] The predicted wage gap ($\ln \hat{\omega}' - \ln \hat{\omega}$) is then included in the hazard model as deviation from its mean.

Finally, it should be noted that the time varying covariates used in the analysis are collected on a yearly basis in NLSY and thus assumed constant for each year but vary across one-year intervals. That is, time dependent variables such as students' age, curriculum type they are enrolled, unemployment rate for the labor market in the current residence and the net total family income vary only across one-year intervals but remain

constant within each year. Therefore, the month-specific risks such as "summer effect" may not be fully controlled unless the failure events are strongly associated with characteristics that remain constant throughout that year.

III. DURATION OF HIGH SCHOOL ENROLLMENT AND DROPOUT RISKS

One natural question that may arise among several on the issues of high school enrollment is whether the risk of unsuccessful leave without completing the academic requirements is duration dependent. In other words, does the risk of high school dropout increase or decrease with time? If we assume that everyone "tries out" first and discovers the chance of academic success as more scholastic information becomes available through their school life, we may expect increasing probabilities of dropout risks as time elapses, in which case we would expect positive duration dependence. If students find out as they near the graduation date that their chance of receiving a diploma may be slim, then the discouraged students might even quit school right before that date. Then, we would observe that the probabilities of leaving school clutter prior to graduation. Alternatively, under an assumption that less able students (e.g., students with lower standardized test scores) tend to give up and quit school at an early stage of their high school career, then these students are more prone to leave school sooner than the others, in which case we would expect negative duration dependence as the students who remain in school longer have lower hazards than those who leave early. There also may be the case that the exogenous shocks such as financial difficulties due to parent's being seriously ill or becoming unemployed may arrive at a constant rate for each student in the fixed time intervals, leading us to believe that everyone faces the constant risk of failure throughout their entire school career. Or do they?

Before we move on to the models and data issues in detail, it is worthwhile to discuss the empirically estimated hazard function, integrated hazard and the Kaplan-Meier survivor. The hazard function or the conditional probability of a failure at time t

$$\lambda(t)=\lim_{h\to 0}\frac{P(t\leq T<t+h \mid T\geq t)}{h}$$

is expressed as

$$\lambda(t)=f(t)/S(t)$$

where t represents time, and the survivor $S(t)=1-F(t)$ is defined as the function of the probability distribution of duration

$$F(t)=\Pr(T<t)$$

which specifies the probability that the random variable T is less than some actual value t. The corresponding density function is then $f(t)=dF(t)/dt$. Thus, $\lambda(t)$ is roughly the rate at which spells will be completed at duration t, given that they last until t. The integrated hazard is defined

$$\Lambda(t)=\int_0^t \lambda(u)du$$

with the relation to the survivor function

$$S(t)=\exp[-\Lambda(t)].$$

The empirical Kaplan-Meier survivor is expressed as a function of the hazards

$$\hat{S}(t)=\prod_{i=1}^{j}(1-\hat{\lambda}_i)$$

where $\hat{\lambda}(t_i)=h_i/n_i$; h_i is the number of completed spells and n_i the number of risk set.

Table 3 shows the completed lengths of enrollment in months for 348 high school students who were in the 9th grade in 1979, with 988 lengths censored at 44 months. Additional information is also reported in the table on the risk set (n_i), the number of failure events (h_i), hazards and integrated hazards estimated based on the Kaplan-Meier survivor.[5]

The hazard function is plotted in Figure 1. Although we do not observe a discernable trend, we recognize more noticeable spikes in the mid- to late periods, most notably at 33 months followed by a sharp drop. The hazards then rise again at 37 and 41 months. The exponential distribution is appropriate in estimating a parametric model of duration data if the hazard is constant over time, while the Weibull may be a sensible assumption if the duration distribution reveals monotonic time dependency. However, the

Table 3 Empirical estimates of the Kaplan-Meier survivor, hazard, and integrated hazard on high school enrollment data

Ordered duration	Duration in months	Risk set	Number of failures	Hazard	Integrated hazard	K-M Survivor
1	3	1,336	1	.0008	.0010	.999
2	4	1,335	3	.0023	.0030	.997
3	5	1,332	1	.0008	.0040	.996
4	6	1,331	1	.0008	.0040	.996
5	7	1,330	6	.0045	.0090	.991
6	8	1,324	9	.0068	.0161	.984
7	9	1,315	15	.0114	.0274	.973
8	10	1,300	1	.0008	.0284	.972
9	11	1,299	4	.0031	.0315	.969
10	12	1,295	6	.0046	.0356	.965
11	13	1,289	10	.0078	.0440	.957
12	14	1,279	4	.0031	.0471	.954
13	15	1,275	7	.0055	.0524	.949
14	16	1,268	8	.0063	.0587	.943
15	17	1,260	15	.0119	.0704	.932
16	18	1,245	7	.0056	.0758	.927
17	19	1,238	14	.0113	.0877	.916
18	20	1,224	10	.0082	.0954	.909
19	21	1,214	11	.0091	.1054	.900
20	22	1,203	1	.0008	.1054	.900
21	23	1,202	5	.0042	.1098	.896
22	24	1,197	9	.0075	.1177	.889
23	25	1,188	12	.0101	.1278	.880
24	26	1,176	14	.0119	.1393	.870
25	27	1,162	10	.0086	.1485	.862
26	28	1,152	14	.0122	.1602	.852
27	29	1,138	10	.0088	.1696	.844
28	30	1,128	13	.0115	.1803	.835
29	31	1,115	11	.0099	.1912	.826
30	32	1,104	16	.0145	.2058	.814
31	33	1,088	28	.0257	.2319	.793
32	35	1,060	7	.0066	.2383	.788
33	36	1,053	7	.0067	.2446	.783
34	37	1,046	12	.0115	.2562	.774
35	38	1,034	9	.0087	.2653	.767
36	39	1,025	2	.0020	.2666	.766
37	40	1,023	8	.0078	.2744	.760
38	41	1,015	14	.0138	.2890	.749
39	42	1,001	11	.0110	.2998	.741
40	43	990	2	.0020	.3011	.740

Note: 988 observations are censored at enrollment length of 44 months.

shape of the hazard in Figure 1 is not strictly monotonic, and it is not quite obvious from the figure whether the hazard is increasing or decreasing as time passes.

As a supplement to the graphical exploration of Figure 1, Figure 2 illustrates the

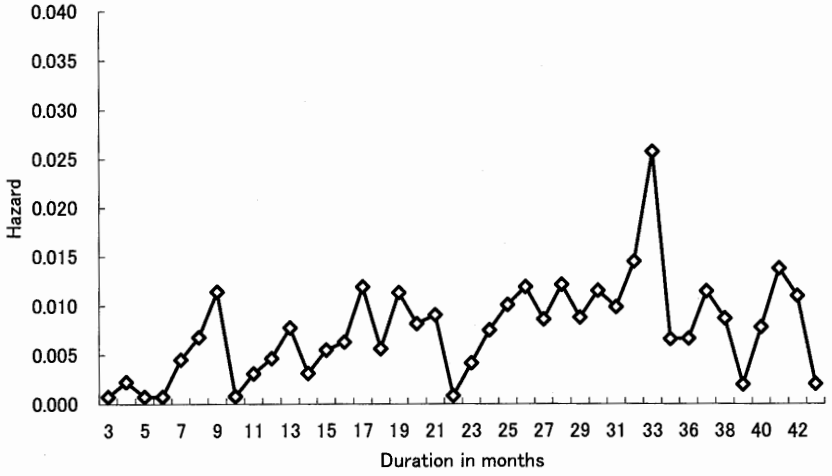

Figure 1 Empirical hazards based on the Kaplan-Meier survivor

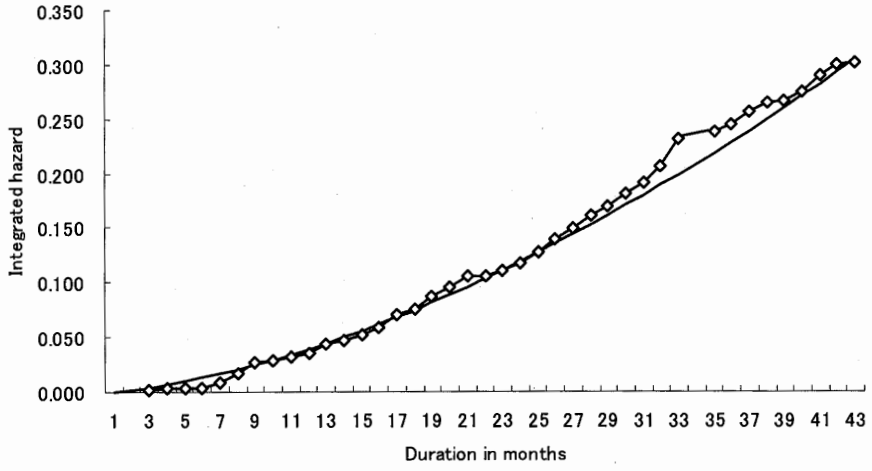

Figure 2 Integrated hazards of high school dropout

integrated hazards empirically estimated based on the Kaplan-Meier survivor or product limit estimator as reported in Table 3. The integrated hazards reveal slight convexity, and the plot of the fitted estimates on months underscores the nonlinearly increasing trend of the curve[6]. Noting that the constant hazard produces a straight line of the integrated hazards, for which the exponential distribution may be used for the baseline hazard in the Cox regression model, the figure suggests that the exponential assumption may not be an adequate one. In light of the graphical exploration of Figurer 2, the Weibull baseline hazard which allows for a more flexible model with a scale parameter seems a sensible candidate in estimating the Cox regression model of high school dropout.

In viewing Figure 1, we also consider a case with a nonmonotonic hazard distribution, which may be represented by the log-logistic distribution with parameters $\gamma>0$ and $\alpha>0$ with the hazard $\lambda(t)=\gamma\alpha t^{\alpha-1}/(1+t^{\alpha}\gamma)$. In the log-logistic model with $\alpha>0$, the hazard first increases with duration, then decreases, and when $0<\alpha\leq 1$ the hazard function decreases with duration.

Figure 2 also shows a few plateaus in the integrated hazards deviating from the fitted trend curve at around the 9th, 21st, and 33rd months counting from September of 1978, followed almost cyclically by an immediate drop in the hazards at 10, 22 and 35 months as observed in Figure 1. As these months represent June of 1979, 1980 and 1981 respectively, these figures suggest that high school students face a higher risk of leaving school especially during summer and fail to return for the new school year. The integrated hazards then seem to subside in the very last periods before the prospective graduation date, though the hazard function in Figure 1 still reveals two noticeable spikes at months 37 and 41.

Although our preliminary analysis is graphical and informal, the empirical estimates of the hazard function and the associated integrated hazards certainly provide us with sensible insights in choosing an appropriate hazard distribution. The graphical analysis is particularly useful in the absence of strong support of economic theory about the shapes of the hazard function, an issue which most applied studies have been subject to in the econometric analysis of duration data.[7]

IV. RELATIVE RISK MODEL WITH WEIBULL BASELINE HAZARD

One reasonably expects that the characteristics causing a lower hazard will be more concentrated among the individuals in the population who are still remaining after individuals with a higher hazard have left. For example, students with longer enrollment lengths may disproportionately be found among children of affluent families with highly educated parents, who are unlikely to suffer from financial constraints, etc. Intuitively, more mobility prone students are the first to leave the population leaving the less prone students behind, thus leading us to conclude with the misapprehension of stronger negative duration dependence than actually exists.[8] This sorting effect may mask a much larger increase in the true hazards of the average population. These problems, however, are solved by using the relative risk model which controls for the effects of regressors in a similar manner as the ordinary regression models.

In view of our graphical exploration using Figure 2 which reveals convexity with possibly increasing hazards, we estimate the relative risk model with the Weibull baseline hazard

$$\lambda(t)=\alpha t^{\alpha-1}\exp\{X(t)\beta\}$$

where α estimates a parameter which would be greater than unity if the hazard is increasing as time elapses (positive duration dependent), less than unity if decreasing (negative duration dependent), and unity for the constant hazard which reduces to the exponential model.[9] $X(t)$ is a vector of observed covariates, and β is a corresponding vector of parameters to be estimated. It is noted that $X(t)$ includes both time dependent and independent regressors. The importance of time dependent covariates and censored observations makes a duration data analysis especially useful as it is difficult to estimate a time-varying hazards model with explanatory variables that are fixed constant over time.

The relative risk model with a specific distributional assumption on the baseline hazard, however, is estimated at the expense of possible misspecification. It is well understood (Heckman and Singer 1984a; Kiefer 1988; Lancaster, 1985) that parametrically misspecified hazard model is inconsistently estimated with respect to the effects of regressors. Therefore, the estimates of the partial likelihood approach suggested by Cox (1972, 1975) are also obtained in comparison with the estimation results from the Weibull

model. With Cox's approach where t and x are factored out into separate but multiplicative terms in the form $\lambda(t,x,\alpha,\beta)=\lambda_0(t,\alpha)\phi(x,\beta)$, the partial likelihood function is expressed

$$\prod_i \frac{\lambda(t_j,x_j,\beta)}{\sum_{i=j}\lambda(t_i,x_i,\beta)}=\prod_i \frac{\phi(x_j,\beta)}{\sum_{i=j}\phi(x_i,\beta)} \qquad \text{for all } j$$

where $\lambda_0(t,\alpha)$ is the baseline hazard and $\phi(x,\beta)=\exp\{X(t)\beta\}$ in our model. In words, the likelihood is formed as the product of the individual contributions to the partial likelihood, and the log-likelihood function is defined

$$L(\beta)=\sum_{i=1}^{n}\left\{\ln\phi(x_j,\beta)-\ln\left[\sum_{i=j}^{n}\phi(x_j,\beta)\right]\right\}.$$

Thus, the estimation of the β parameters requires no *a priori* distributional assumptions on the shape of the baseline hazard λ_0. The intercept term, however, is not estimated in the partial likelihood approach as it is absorbed into λ_0. Although the coefficients of the regressors are consistently estimated, the disadvantage of the partial likelihood estimation is that it makes no suggestions on duration dependency of failure risk and the baseline hazards need to be estimated indirectly using the consistent estimates of β (Cox and Oakes 1985) or directly for each discrete interval by more generalized maximum likelihood estimation.[10]

V. DISCUSSION

The results from parametric estimation with the Weibull baseline hazard are reported in Table 4 for various specifications. The first specification is estimated with characteristic variables of students' individual and demographic background only. The result indicates that the risk of dropping out of high school is significantly greater for male students than females. This may be due to the higher opportunity cost of staying in school for male students because their potential market wages are higher compared to those of their female counterparts. It is somewhat striking that nonwhite students face a lower hazard relative to white students. Again, the reason may be that the potential wages for white students are higher than those for nonwhite students. Students' age also negatively influences the probability of their unsuccessful leave from school, which implies that the

Table 4 Parametric estimates of the relative risk model for high school dropouts with the weibull baseline hazard

Variable	Specification 1 (standard error) [a]	Specification 2 (standard error) [a]	Specification 3 (standard error) [a]	Specification 4 (standard error) [a]	Specification 5 (standard error) [a]
Constant	-12.386 (0.148) ***	-12.561 (0.152) ***	-13.664 (0.366) ***	-13.897 (0.367) ***	-13.949 (0.365) ***
Male	0.306 (0.112) ***	0.300 (0.112) ***	0.337 (0.112) ***	0.359 (0.112) ***	0.014 (0.136)
Nonwhite	-0.206 (0.125) *	-0.250 (0.127) **	-0.226 (0.125) *	-0.273 (0.127) **	-0.153 (0.129)
Age [b, c]	-1.716 (0.071) ***	-1.676 (0.070) ***	-1.546 (0.071) ***	-1.484 (0.072) ***	-1.440 (0.072) ***
AFQT score [c]	-0.028 (0.004) ***	-0.026 (0.004) ***	-0.026 (0.004) ***	-0.024 (0.004) ***	-0.026 (0.004) ***
1 if AFQT score missing [c]	0.421 (0.404)	0.411 (0.404)	0.438 (0.406)	0.395 (0.406)	0.791 (0.413) *
Father's highest grade completed [c]	-0.025 (0.021)	-0.007 (0.021)	-0.031 (0.020)	-0.008 (0.021)	-0.037 (0.021) *
1 if father's grade missing [c]	0.139 (0.141)	0.011 (0.145)	0.103 (0.139)	-0.048 (0.144)	-0.360 (0.160) **
Mother's highest grade completed [c]	-0.059 (0.023) ***	-0.062 (0.022) ***	-0.069 (0.022) ***	-0.070 (0.022) ***	-0.083 (0.022) ***
1 if mother's grade missing [c]	0.104 (0.201)	0.130 (0.201)	0.123 (0.200)	0.167 (0.200)	0.428 (0.210) **
Type of curriculum:					
Vocational [b]	-0.245 (0.178)	-0.255 (0.179)	-0.273 (0.179)	-0.291 (0.180)	-0.109 (0.182)
Commercial [b]	0.484 (0.289) *	0.359 (0.292)	0.375 (0.289)	0.199 (0.293)	0.294 (0.295)
College preparatory [b]	-0.629 (0.192) ***	-0.604 (0.192) ***	-0.748 (0.194) ***	-0.736 (0.194) ***	-0.740 (0.194) ***
General program [b]					
Unknown [b]	1.105 (0.236) ***	1.097 (0.237) ***	0.864 (0.237) ***	0.852 (0.238) ***	0.827 (0.238) ***
Unemployment rate:					
< 6.0% [b]	———	———	1.702 (0.365) ***	1.787 (0.366) ***	1.967 (0.370) ***
6.0-8.9% [b]	———	———	1.407 (0.352) ***	1.386 (0.352) ***	1.479 (0.354) ***
9.0-11.9% [b]	———	———	0.586 (0.363)	0.527 (0.363)	0.568 (0.363)
12.0-14.9% [b]	———	———	0.595 (0.404)	0.606 (0.405)	0.532 (0.405)
≥ 15.0% [b]	———	———			
1 if unemployment rate missing [b]	———	———	2.556 (1.063) **	2.179 (1.066) **	2.136 (1.069) **
Family income/1,000 [b, c]	———	-0.026 (0.006) ***	———	-0.029 (0.006) ***	-0.029 (0.006) ***
1 if family income missing [b]	———	0.118 (0.145)	———	0.078 (0.133)	0.084 (0.145)
Predicted wage differential [c]					-2.427 (0.535) ***
Alpha	2.817 (1.047) ***	2.853 (1.046) ***	2.898 (1.046) ***	2.955 (1.046) ***	2.992 (1.046) ***
Log likelihood	-1,599.5	-1,588.8	-1,566.3	-1,552.5	-1,541.1
Sample size	1,336	1,336	1,336	1,336	1,336

[a] Asymptotic normal standard errors.

[b] Denotes covariates that vary across one-year intervals, but are assumed constant within intervals. All other covariates are constant within and across intervals.

[c] Denotes continuous covariates, which enter hazards as deviations from sample me.

* Significant at .10 level; ** significant at .05 level; *** significant at .01 level.

failure risk decreases as students get older. This may be because young individuals gradually learn the importance of high school education as they acquire more experience through activities inside and outside school. The failure risk also decreases with students' ability measured by AFQT scores, suggesting that students with higher ability are likely to remain in school longer. The result therefore indicates that the effect of lower per-unit effort cost may never be dominated by the effect of higher opportunity cost of staying in school, i.e., in terms of lost wages, for high ability students. Although father's highest grade completed may not have a significant impact on his child's enrollment lengths, the mother's educational experiences appear to strongly influence the child's enrollment duration in high school.

The type of skills students acquire in school may also affect the probabilities of their quitting as the students with skills directly connected to the workplace may be eager to join the labor force or may be able to find jobs with ease with stronger demand from their potential employers. The estimate from the first specification somewhat supports this view. Although a vocational curriculum has a negative but insignificant effect on the failure risk, students in a commercial program may face a higher probability of dropping out compared with those in a general program. Students in a college preparatory curriculum, in contrast, have significantly lower probabilities of leaving their school. However, the lower risk for students in a college preparatory program may be due to their high motives or attitudes towards schooling rather than the contents of their academic curriculum which may not be directly connected to the workplace. The alpha parameter is estimated with a significant and positive effect, which indicates that the probabilities of students' dropping out of high school increases as time elapses even after controlling for their family and demographic characteristics.

The second specification in Table 4 is estimated with the same set of covariates as the first specification along with the net total income a student's family earned in the past calendar year. Under an assumption that a student of a family with lower household income faces a stronger need to become financially independent or to support his family, the risk of unsuccessful leave may be inversely related with the family income. The result in the second column supports this view. The coefficient estimates on students' demographic characteristics and duration dependency are very similar with those from

the first specification, and the effect of net total family income is significantly negatively influencing the probabilities of students' dropping out.

Specification 3 is estimated with the same group of demographic variables in the first specification, along with dummy variables for unemployment rates for the labor market of students' residence. The effect of the local unemployment rate is estimated as a proxy for the labor demand in the local labor market. The result shown in the third specification indicates that students residing in areas with lower unemployment rates face a significantly higher dropout risk, implying that the enrollment duration may be countercyclical. However, the local economic condition is considered to impose a critical influence not only on labor demand for high school students but also for the income of their parents. For example, students and families living in a sluggish local economy may face a higher unemployment rate as well as lower average income. As we see in Specifications 2 and 3, the former captures a counter effect of the latter on students' dropout risk. In order to separate these effects of labor demand for high school students and the family's short-term credit constraints, the fourth specification is now estimated with both the dummy variables for local unemployment rates and the net total family income. The resulting estimates are again very similar with those obtained above. The probability of dropout risk is positively correlated with local economic conditions, and students of affluent families face significantly lower risk of leaving school without receiving a diploma.

Finally, the last specification is estimated with the wage differentials predicted under two alternative regimes; wages if students complete a potential full length of high school enrollment (44 months) and if they drop out and start working at some time prior to the potential maximum length. The estimate of the coefficient indicates that the wage differentials predicted immediately following the 44th month has a significantly negative impact on the risk of enrollment failure. In other words, students with a larger potential return to high school education face a lower risk of dropping out. The most affected variables in the coefficient estimates by the inclusion of the predicted wage differentials are the dummy variables for male and nonwhite students. Note in the previous specifications 1–4 that white male students have a significantly higher dropout risk. Inclusion of the potential wage gap, however, absorbs the significance of these individual

characteristics. The result implies that the dropout risk is greater for white male students because their potential wage gains from completing the full length of high school enrollment are so discouragingly small that extended length of enrollment is not justified for these myopic students. Equivalently, the result suggests that the potential wage gains are higher for nonwhite female students relative to their white male counterparts, at least at the time of job entry immediately following 44 months.

The expected length of completed enrollment for the Weibull model is estimated by[11]

$$E(T)=\int_0^{\infty}(1-F(t))dt=\Gamma\left(\frac{1}{\alpha}+1\right)\exp\left(-\frac{X\beta}{\alpha}\right). \qquad (1)$$

Using eq. (1) and our estimates of β in Specification 5 of Table 4, a "typical" student is expected to remain enrolled beyond 44 months (about 61 months).[12] A one month increase in student's age is expected to increase the expected length of enrollment by about 4 percent or 2.5 months. An increase in the AFQT score by 10 percentile is associated with a 9 percent or 5.5 month increase in students' enrollment length. An additional year of education by parents also has a positive impact on their child's enrollment by about 1.0 and 3.0 percent, respectively, for father and mother. Students in a college preparatory program have a longer expected length of enrollment than those in a general curriculum by 28 percent or 17 months. Moving from an area with 6.0–8.9 percent of unemployment rates to the area of the lowest unemployment (less than 6.0 percent) decreases the expected length of enrollment by 15 percent or by 9 months, suggesting that the duration of high school enrollment is strongly countercyclical. An unexpected increase in the net total family income by 10,000 dollars increases the enrollment length by 10 percent or 6 months. Finally, a one-tenth increase in the potential log-wage gap increases the expected length of enrollment by about 8 percent or 5 months.

All these results are in agreement with our intuitions and economic reasoning. However, these results should be accepted with skepticism of potential biases as incorrect assumptions on the baseline hazard and a misspecified functional form of the model lead to inconsistent estimates of the parameters of explanatory variables. If a Weibull model is misspecified as exponential, for example, the coefficients are likely to be underestimated when the scale parameter α is greater than unity. When α is less than unity, the coefficients

Table 5 Expected duration of high school enrollment

Student's background characteristics	Expected length of enrollment in months
Typical	60.5
Non-typical:	
Age + 1 month	63.0
AFQT score, 10.0 percentile higher than average	66.0
Father's highest grade + 1 year	61.2
Mother' highest grade + 1 year	62.2
Enroll in college preparatory curriculum	77.4
Unemployment rate < 6.0%	51.4
Net total family income, $10,000 higher than average	66.6
Predicted log-wage gap, 0.1 higher than average	65.6

Note: The typical student is a white male enrolled in a general curriculum, residing in an area with 6.0-8.9% unemployment rates with age, AFQT score, both parents' highest grade, net total family income, and predicted wage gap all equal to the sample mean which is zero as all the continuous variables are included as deviations from sample mean.

are likely to be overestimated.[13] Even if the model is correctly specified with respect to the functional form or the correct distributional assumption on the baseline hazard, neglected heterogeneity may lead to biased estimates of duration dependency and regressors, and thus incorrect inferences on the effects of these parameters. Therefore, we now consider estimating the relative risk model incorporating these issues of misspecification and unobserved heterogeneity.

VI. ISSUES OF MISSPECIFICATION AND NEGLECTED HETEROGENEITY

As we have already discussed, the relative risk or Cox regression model with a specific distributional assumption is estimated at the expense of possible biases due to misspecification. The nonparametric result from the partial likelihood approach is presented in Column 1 of Table 6 along with the estimates obtained based on the Weibull baseline hazard (Column 2). The reported coefficients of the partial likelihood estimates are very similar with those of the Weibull model, which suggests that the Weibull assumption is not seriously biasing the estimated effects of the explanatory variables. In comparison with these results, the coefficient estimates of the exponential model are also

Table 6 Parametric and nonparametric estimates of the relative risk model of high school dropout

Variable	(1) Partial likelihood estimates (standard error) [a]	(2) Weibull model (standard error) [a]	(3) Exponential model (standard error) [a]	(4) Log-logistic model (standard error) [a]
Constant	———	-13.949 (0.367) ***	-6.310 (0.346) ***	-17.031 (0.405) ***
Male	0.019 (0.136)	0.014 (0.136)	-0.035 (0.133)	0.010 (0.179)
Nonwhite	-0.166 (0.129)	-0.153 (0.129)	-0.123 (0.126)	-0.189 (0.172)
Age [b, c]	-1.481 (0.080) ***	-1.440 (0.072) ***	-0.857 (0.062) ***	-1.855 (0.095) ***
AFQT score [c]	-0.026 (0.004) ***	-0.026 (0.004) ***	-0.022 (0.004) ***	-0.037 (0.005) ***
1 if AFQT score missing	0.722 (0.415) *	0.791 (0.413) *	0.525 (0.407)	0.768 (0.597)
Father's highest grade completed [c]	-0.036 (0.021) *	-0.037 (0.021) *	-0.032 (0.021)	-0.057 (0.029) **
1 if father's grade missing	-0.328 (0.160) **	-0.360 (0.160) **	-0.265 (0.158) *	-0.402 (0.223) *
Mother's highest grade completed [c]	-0.078 (0.022) ***	-0.083 (0.022) ***	-0.063 (0.022) ***	-0.104 (0.031) ***
1 if mother's grade missing	0.410 (0.210) *	0.428 (0.210) **	0.283 (0.211)	0.311 (0.286)
Type of curriculum:				
Vocational [b]	-0.121 (0.182)	-0.109 (0.182)	-0.169 (0.180)	-0.234 (0.242)
Commercial [b]	0.275 (0.295)	0.294 (0.295)	0.290 (0.291)	0.091 (0.421)
College preparatory [b]	-0.734 (0.240) ***	-0.740 (0.194) ***	-0.699 (0.190) ***	-0.906 (0.232) ***
General program [b]				
Unknown [b]	0.730 (0.240) ***	0.827 (0.238) ***	0.410 (0.234) *	0.926 (0.352) ***
Unemployment rate:				
< 6.0% [b]	1.960 (0.365) ***	1.967 (0.370) ***	1.549 (0.363) ***	2.465 (0.422) ***
6.0-8.9% [b]	1.463 (0.349) ***	1.479 (0.354) ***	1.320 (0.348) ***	1.864 (0.394) ***
9.0-11.9% [b]	0.564 (0.362)	0.568 (0.363)	0.533 (0.362)	0.599 (0.405)
12.0-14.9% [b]	0.497 (0.405)	0.532 (0.405)	0.569 (0.403)	0.574 (0.466)
≥ 15.0% [b]				
1 if unemployment rate missing [b]	2.191 (1.065) **	2.136 (1.069) **	1.840 (1.064) *	3.043 (1.449) **
Family income/1,000 [b, c]	-0.029 (0.006) ***	-0.029 (0.006) ***	-0.019 (0.006) ***	-0.037 (0.007) ***
1 if family income missing [b]	0.092 (0.145)	0.084 (0.145)	0.084 (0.145)	0.076 (0.191)
Predicted wage differential	-2.368 (0.530) ***	-2.427 (0.535) ***	-1.783 (0.507) ***	-3.322 (0.708) ***
Alpha	———	2.992 (1.046)	Fixed at 1.00	3.788 (1.046)
Log likelihood	———	-1,541.1	-1,734.2	-1,545.6
Sample size	1,336	1,336	1,336	1,336

[a] Asymptotic normal standard errors.

[b] Denotes covariates that vary across one-year intervals, but are assumed constant within intervals. All other covariates are constant within and across intervals.

[c] Denotes continuous covariates, which enter hazards as deviations from sample means.

* Significant at .10 level; ** significant at .05 level; *** significant at .01 level

presented in Column 3. As we expected, the exponential model generally produces underestimated results when the Weibull is the correct distribution of the baseline hazard. Although the model is better fit than the exponential, the log-logistic model in Column 4, in contrast, overestimated the effects of covariates, especially for significant estimates. These results underscore the importance of fitting the parametric model with the correct baseline hazard distribution. Overall, the Weibull distribution seems a more adequate assumption for the current sample than the exponential or the log-logistic in specifying the hazard model.

It is a statistical fact that neglected unobservables bias estimated hazards towards negative duration dependence (Heckman and Singer 1984). Not only does the neglected heterogeneity impose an incorrect inference on the duration dependency, but also it leads to an incorrect conclusion of the inferences on the effects of covariates.[14] In order to take into account the impact of the unobserved heterogeneity, we follow Lancaster and Nickell (1980) and assume multiplicative heterogeneity

$$\gamma=\nu\exp(X\beta)$$

where ν is assumed Gamma distributed with mean 1 (by normalization) and variance σ^2. Then, our survivor function becomes

$$1-F(t;\alpha,\beta)=[1+\sigma^2 t^{\alpha}\exp(X\beta)]^{-\sigma^{-2}} \tag{2}$$

with the density

$$f(t;\alpha,\beta)=[1+\sigma^2 t^{\alpha}\exp(X\beta)]^{-\sigma^{-2}-1}\alpha t^{\alpha-1}\exp(X\beta) \tag{3}$$

Substituting (2) and (3) into the likelihood function

$$L(t;\alpha,\beta)=\prod_{i=1}^{n} f(t;\alpha,\beta)^{\delta_i}[1-F(t;\alpha,\beta)]^{1-\delta_i}$$

we obtain the log-likelihood

$$\log L(t;\alpha,\beta)=\sum_{i}^{k_1}\delta_i\left\{\left(-\sigma^{-2}-1\right)\log\left[1+\sigma^2 t^{\alpha}\exp(X\beta)\right]+\log\alpha+(\alpha-1)\log t+X\beta\right\}$$
$$+\sum_{i}^{k_2}(1-\delta_i)\left\{\left(-\sigma^{-2}\right)\log\left[1+\sigma^2 t^{\alpha}\exp(X\beta)\right]\right\}. \tag{4}$$

The log-likelihood function (4) is then maximized with respect to the parameters α, β and σ^2. The resulting maximum likelihood estimate of σ^2 was extremely small and approaching zero. The other parameter estimates did not alter significantly nor did the value of the log-likelihood. There seems, therefore, to be little evidence of neglected heterogeneity left in the model with this sample, and we may conclude that the unobserved heterogeneity is not seriously biasing our parameter estimates of α and β, and all the results in Tables 4 and 5 with the Weibull baseline hazard remain valid.

Finally, given consistently estimated β, a sensible nonparametric estimate of the baseline hazard $\lambda_0(t)$ is obtained as suggested by Cox and Oakes (1985)

$$\hat{\lambda}_0(t_j)=\frac{d(t_j)}{\sum_{i=j}^{n}\phi(x_i,\hat{\beta})} \tag{5}$$

where $d(t_j)$ is the number of spells ended at duration t_j and equals 1 in the absence of ties. The denominator of (5) is evaluated at the estimated values of β. The baseline hazards obtained at the average values of covariates are plotted in Figure 3, along with nonparametrically obtained hazards (copied from Figure 1). It is found in the figure that most spikes in the early and mid-intervals are trimmed an-d the fluctuations previously

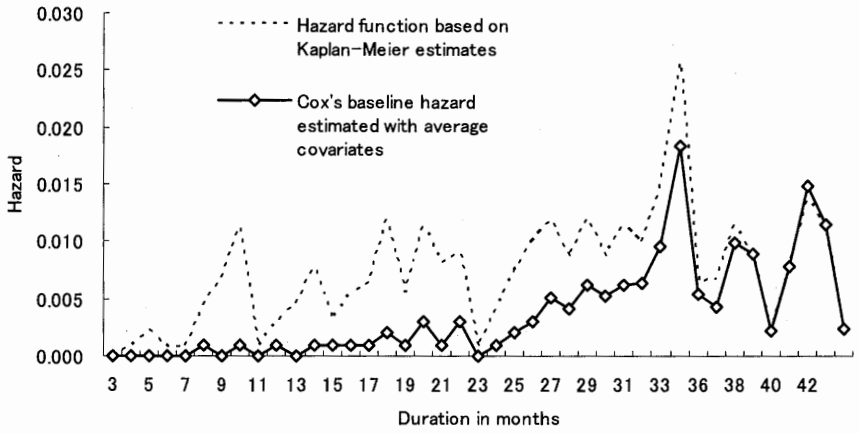

Figure 3 A comparison of emprically estimated baseline hazards and Cox baseline hazards with average covariates

observed have significantly been absorbed by the effects of the regressors entered into the relative risk model. In contrast, the regressors do not seem to have enough control on the shape of the hazards in the later months of high school enrollment duration. Figure 3 shows that the family and demographic characteristics significantly influence the enrollment behavior of children in the early periods in high school but these characteristics tend to lose their influences as students enter the senior year.

Cameron and Heckman (1998) examine the empirical regularity that the effects of family background and resources on the probability of transiting from one grade to the next diminish at higher levels of education. They show that the pattern of declining logit coefficients for higher grade transitions is critically dependent on choices of functional forms for the distribution of unobservables. Determining whether our result with respect to the diminishing effects of family and demographic characteristics is subject to the implication of their results requires further study. Provided that the effect of neglected heterogeneity left in the model is negligible and the lower hazard students remaining in the data are homogeneous, the rising dropout risks emerging after the sophomore year is considered to be caused by shocks that are associated with month-specific factors and unobserved characteristics that typical high school students may equally experience by reaching that stage of adolescence. Although the sources of these shocks are yet to be identified, it clearly indicates that the later periods in high school is a critical stage for average high school students who are contemplating whether to complete the designated full lengths of enrollment.

VII. CONCLUDING REMARKS

In this study, duration data on high school enrollment is graphically explored to identify time dependency of dropout probabilities. In light of the graphical exploration of the data, the relative risk model is estimated with the Weibull baseline hazard in comparison with alternative distributions. The resulting estimates indicate significant effects of the family and demographic characteristics of high school students with strong positive duration dependence. In order to take into consideration the issue of parametrically misspecified hazard function, the estimates are then compared with the results from the Cox's partial likelihood estimation. The comparison of the results from the Weibull and

partial likelihood models is found to be very similar with respect to the effects of covariates, suggesting that the Weibull distribution is a reasonable assumption for the sample. When the parametric model is estimated with the exponential baseline hazard, the coefficients of the regressors are generally underestimated if the alpha parameter is greater than unity or the hazard is increasing. The coefficients of the log-logistic model, in contrast, are generally overestimated, especially for the significant estimates.

The maximum likelihood estimation of the Weibull model with the gamma distributed heterogeneity produces a similar result as the heterogeneity uncorrected estimation, with a trivial effect of neglected heterogeneity. Thus, we conclude that the obtained results with the Weibull assumption remain valid. Some of the important results show that students' ability using AFQT as a proxy variable is inversely related with their dropout risk. Parents' highest grade, particularly of mothers, also negatively influences the probabilities of their child's enrollment failure. Moreover, the probability of students' dropping out of high school is inversely related with the net total family income. When the labor demand for students is measured by the unemployment rates in the local labor market of their residence, the enrollment duration is found to be significantly countercyclical. Finally, white male students face a higher dropout risk than their nonwhite female counterparts because the potential wage gains from completing the full length of enrollment duration (44 months) are smaller for the former students.

The effects of the family and demographic characteristics on the enrollment decision of high school students diminish as students enter the junior year. In other words, the risk is not as well controlled in the later half periods of high school as in the earlier with the same family and background characteristics. In particular, some of the highest risks experienced in the senior year remain uncontrolled. However, identifying the source of the shocks causing these higher risks in the later months requires further study.

REFERENCES

Becker, Gary S. *A Treatise on the Family, Enlarged Edition,* Cambridge, MA: Harvard University Press, 1991.

Burdett, Kenneth, Nicholas M. Kiefer and Sunil Sharma. "Layoffs and Duration Dependence in a Model of Turnover," Journal of Econometrics, vol. 28, 1985, pp. 51–69.

Cameron, Stephen V. and James J. Heckman. "Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males," Journal of Political Economy, vol. 106, no. 2, 1998, pp. 262–333.

____. "The Dynamics of Educational Attainment for Black, Hispanic, and White Males," Journal of Political Economy, vol. 109, no. 3, 2001, pp. 455–499.

Chesher, Andrew. "Testing for Neglected Heterogeneity," Econometrica, vol. 52, no. 4, July 1984, pp. 865–872.

Cox, David R. "Regression Models and Life-Tables (with discussion)," Journal of Royal Statistical Society, May/August 1972, B34, pp. 187–220.

Cox, David R. and David Oakes. *Analysis of Survival Data.* London and New York: Chapman & Hall, 1985.

____. "Partial Likelihood," Biometrika, May/Aug 1975, vol. 62, no. 2, pp. 269–76.

Dolton, Peter and Wilbert van der Klaauw. "Teaching Salaries and Teacher Retention" in *Assessing Educational Practices: The Contribution of Economics,* ed. by William E. Becker and William J. Baumol. Cambridge, MA: The MIT Press, 1996.

Edwards, Linda. "The Economics of Schooling Decisions: Teenage Enrollment Rates," Journal of Human Resources, vol. 10, no.2, 1975, pp. 155–173.

Griliches, Zvi. "Wages of Very Young Men," Journal of Political Economy, vol. 84, no. 4, pt. 2, 1976, pp. S69–S85.

Han, Aaron and Jerry A. Hausman. "Flexible Parametric Estimation of Duration and Competing Risk Models," Journal of Applied Econometrics, vol. 5, 1990, pp. 1–28.

Heckman, James J. and Burton Singer. "Econometric Duration Analysis," Journal of Econometrics, vol. 24, 1984, pp. 63–132.

____. "A Method for Minimizing the Impact of Distribution Assumptions in Econometric Models for Duration Data," Econometrica, vol. 52, no. 2, March 1984, pp. 271–320.

Kalbfleisch, John D. and Ross L. Prentice. *The Statistical Analysis of Failure Time Data,* Second Edition. Hoboken, NJ: John Wiley and Sons, Inc., 2002.

Kennan, John. "The Duration of Contract Strikes in U.S. Manufacturing," Journal of Econometrics, vol. 28, 1985, pp. 5–28.

Kiefer, Nicholas M. "A Simple Test for Heterogeneity in Exponential Models of Duration," Journal of Labor Economics, vol. 2, no. 4, 1984.

____. "Specification Diagnostics Based on Laguerre Alternatives for Econometric Models of Duration," Journal of Econometrics, vol. 28, 1985, pp. 135–154.

____. "Evidence on the Role of Education in Labor Turnover," Journal of Human Resources, vol. 20, no. 3, 1985, pp. 445–452.

____. "Economic Duration Data and Hazard Functions," Journal of Economic Literature, vol. 26, June 1988, pp. 646–679.

Lancaster, Tony. "Econometric Methods for the Duration of Unemployment," Econometrica, vol. 47, no. 4, July 1979, pp. 939–956.

____. "Generalised Residuals and Heterogeneous Duration Models with Applications to the Weibull Model," Journal of Econometrics, vol. 28, 1985, pp. 155–169.

Lancaster, Tony and S. Nickell. "The Analysis of Re-employment Probabilities of the Unemployed," Journal of the Royal Statistical Society A143, 1980, pp. 141–165.

Light, Audrey. "Hazard Model Estimates of the Decision to Reenroll in School," Labour Economics, vol. 2, 1996, pp. 381–406.

Light, Audrey and Yoshiaki Omori. "A Competing Risks Hazard Model of Schooling and Job Turnover," Working Papers in Economics, No. 97–02, Department of Economics, Ohio State University, December 1996.

Lynch, Lisa M. "State Dependency in Youth Unemployment: A Lost Generation?" Journal of Econometrics, vol. 28, 1985, pp. 71–84.

____. "Differential Effects of Post-School Training on Early Career Mobility," Working Paper No. 4034, National Bureau of Economic Research Working Paper Series, March 1992.

____. "Private-Sector Training and the Earnings of Young Workers." *American Economic Review,* Vol.82, No.1, March 1992: 299–312.

Mincer, Jocob. "Job Training, Wage Growth and Labor Turnover." National Bureau of Economic Research Working Paper No. 2090, Cambridge, MA, August 1988.

____. "Job Training: Costs, Return, and Wage Profiles" in *Studies in Human Capital: Collected Essays of Jacob Mincer, Vol.1.* Brookfield, VT: Edward Elgar Publishing Company. 1993: 263–281.

Meyer, Bruce D. "Unemployment Insurance and Unemployment Spells," Econometrica, vol. 58, no. 4, July 1990, pp. 757–782.

Moffitt, Robert. "Unemployment Insurance and the Distribution of Unemployment Spells," vol. 28, 1985, pp. 85–101.

Narendranathan, Wiji and Stephen Nickell. "Modelling the Process of Job Search," Journal of Econometrics, vol. 28, 1985, pp. 29–49.

Smith, Peter J. *Analysis of Failure and Survival Data,* London and New York: Chapman and Hall/CRC, 2002.

Sueyoshi, Glenn T. "Semiparametric Proportional Hazards Estimation of Competing Risks Models with Time-Varying Covariates," Journal of Econometrics, vol. 51, 1992, pp. 25–58.

U.S. Department of Education. *The Digest of Education Statistics 2001,* National Center for Education Statistics: Washington, DC, 2002.

Waldman, Donald M. "Computation in Duration Model with Heterogeneity," Journal of Econometrics, vol. 28, 1985, pp. 127–134.

Notes

(1) This assumption by no means is intended to preclude the possibility that young individuals do not reenroll once they leave school for the first time. For example, in the analysis which used NLSY data, Light (1996) shows that one third of young individuals who first leave school between 1978 and 1990 are found to return to school before 1991.

(2) Exceptions are Cameron and Heckman (1998, 2000) and preceding studies cited in theirs. Cameron and Heckman discuss a dynamic process of educational attainment for different American male cohorts (1998) and racial and ethnic cohorts (2000). They examine students' transiting from one grade to the next taking the effects of a dynamic selection bias into consideration. This paper, in contrast, is not intended to lay its basis on the grade-transition model but instead to cast light on the analysis of enrollment lengths regardless of grade repetition or transition status.

(3) The NLSY, sponsored by the Bureau of Labor Statistics of the U.S. Department of Labor, is a nationally representative sample of 12,686 young men and women who were 14–22 years old when they were first surveyed in 1979. These individuals were interviewed annually through 1994 and are currently being interviewed on a biennial basis. For more details on the sampling process, refer to NLS Handbook 1998, available from the Center for Human Resource Research at The Ohio State University.

(4) Detailed estimation results from the log-wage regressions with separate subsamples may be obtained from the author upon request.

(5) In Table 3, the integrated hazard is estimated as minus the logarithm of the Kaplan-Meier

estimator. This is equivalent to estimating $\hat{\Lambda}(t_j)=\sum_{i\leq j}\hat{\lambda}(t_i)$ when the hazards are small.

(6) A nonlinear regression of $\Lambda(t)=\gamma t^{\alpha}$, which is exactly the formula of the integrated hazards of the Weibull model, is estimated as

$$\underset{(0.00007)}{\Lambda=0.0007t^{1.6149}}, \quad \underset{(0.0287)}{R^2=0.990}$$

where Λ and t are the integrated hazards and duration in months respectively, and the numbers in parentheses are standard errors.

(7) Heckman and Singer (1984a) exceptionally demonstrate three examples of duration models constructed from economic choice theories.

(8) This implies that neglected heterogeneity potentially causes serious effects on the inferences about duration dependence, regressors or both (Lancaster, 1985; Kiefer, 1988).

(9) The corresponding distribution, survivor, density, and integrated hazard functions of the Weibull model are defined as $F(t)=1-\exp(-\gamma t^{\alpha})$, $S(t)=\exp(-\gamma t^{\alpha})$, $f(t)=\gamma\alpha t^{\alpha-1}\exp(-\gamma t^{\alpha})$, and $\Lambda(t)=\gamma t^{\alpha}$, respectively, while those of the exponential model are $F(t)=1-\exp(-\gamma t)$, $S(t)=\exp(-\gamma t)$, $f(t)=\gamma\exp(-\gamma t)$ and $\Lambda(t)=\gamma t$, with $\lambda(t)=\gamma$. Thus, it is easily seen that the exponential model is a special case of the Weibull when $\alpha=1$.

(10) See for example Meyer (1990) in which he nonparametrically estimates the baseline hazards for each discrete interval.

(11) It is easily seen in eq. (1) that the expected duration equals the inverse of $\exp(X\beta)$ when the model is specified with exponentially distributed hazard or $\alpha=1$.

(12) The "typical" student refers to a white male individual enrolled in a general curriculum and residing in an area with 6.0–8.9% unemployment rates with age, AFQT score, both parents' highest grade completed, net total family income, and predicted wage gap all estimated at the sample mean. Since all the continuous variables are included as deviations from sample means, the typical student's expected length is estimated at age, AFQT score, parents' highest grades, family income and wage gap all equal zero.

(13) See Kiefer (1988) for his regression interpretation of the hazard model, in which he provides a hint about the biases to be expected in the absence of censoring when a Weibull model is misspecified as exponential.

(14) In a hazard model with no time-varying regressors, no censoring and a Weibull baseline hazard, Lancaster (1985) derives the asymptotic bias from omission of heterogeneity. He finds that the parameter α and all the coefficients β are biased towards zero. However, elasticities with respect to the expected value of the log of duration are always correctly estimated even when the true model is a Weibull mixture.