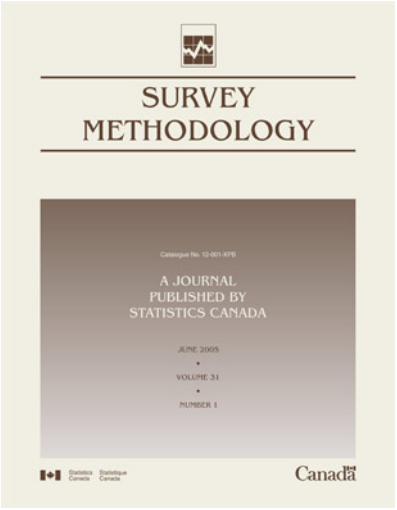
Catalogue no. 12-001-XIE

Survey Methodology

June 2002



Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1-800-263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

National inquiries line	1-800-263-1136
National telecommunications device for the hearing impaired	1-800-363-7629
Depository Services Program inquiries	1-800-700-1033
Fax line for Depository Services Program	1-800-889-9734
E-mail inquiries	infostats@statcan.ca
Website	www.statcan.ca

Information to access the product

This product, catalogue no. 12-001-XIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Publications.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About us > Providing services to Canadians.

Statistics Canada
Business Survey Methods Division

Survey Methodology

June 2002

Published by authority of the Minister responsible for Statistics Canada

October 2006

Catalogue no. 12-001-XIE
ISSN 1492-0921

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-XIF au catalogue).

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Survey Methodology, June 2002
Vol. 28, No. 1, pp. 51-61
Statistics Canada, Catalogue No. 12-001-XIE

Benchmarking Parameter Estimates in Logit Models of Binary Choice and Semiparametric Survival Models

Ian Cahill and Edward J. Chen [1]

Abstract

An approach to exploiting the data from multiple surveys and epochs by benchmarking the parameter estimates of logit models of binary choice and semiparametric survival models is developed. The goal is to exploit the relatively rich source of socio-economic covariates offered by Statistics Canada's Survey of Labour and Income Dynamics (SLID), and also the historical time-span of the Labour Force Survey (LFS), enhanced by following individuals through each interview in their six-month rotation. A demonstration of how the method can be applied is given, using the maternity leave module of the LifePaths dynamic microsimulation project at Statistics Canada. The choice of maternity leave over job separation is specified as a binary logit model, while the duration of leave is specified as a semiparametric proportional hazards survival model with covariates together with a baseline hazard permitted to change each month. Both models are initially estimated by maximum likelihood from pooled SLID data on maternity leaves beginning in the period 1993–1996, then benchmarked to annual estimates from the LFS 1976–1992. In the case of the logit model, the linear predictor is adjusted by a log-odds estimate from the LFS. For the survival model, a Kaplan-Meier estimator of the hazard function from the LFS is used to adjust the predicted hazard in the semiparametric model.

Key Words: Microsimulation; Benchmarking; Semiparametric survival models; Binary logit.

1. Introduction

Researchers often base econometric models on a survey conducted over a short period of time. In this case it may be desirable to incorporate information from a supplementary data source covering a longer period, even if measurements are only available for the dependent variable. For a broad class of non-linear models, we develop a simple method of benchmarking the parameter estimates obtained from a survey rich in explanatory variables to information from a survey with significant historical depth. A primary objective is that model predictions accord with information from the secondary data source. We demonstrate application of the method first to a simple logit model of binary choice, and secondly to a semiparametric survival model. Since the survival model can be viewed as a sequence of binary choices, while retaining an interpretation as an incompletely observed continuous time model, it provides a natural generalization of the first application.

The illustration we provide is a study of maternity leave. The Statistics Canada Survey of Labour and Income Dynamics (SLID) provides data on both the incidence of choosing a maternity leave over withdrawing from the labour force, and on the duration of maternity leave, as well as a rich set of explanatory variables. Because of this we use SLID to estimate base parameters, including those determining the effects of the explanatory variables on the incidence (the logit model) and hazard of returning to work (the survival model). The Canadian Labour Force Survey (LFS) conducted by Statistics Canada provides reasonable proxies for both the incidence and duration extending back to 1976. The SLID parameter estimates are therefore benchmarked to LFS estimates of incidence and the hazard of returning to work during the period 1976-1992, which is prior to the availability of SLID data.

The work was carried out while developing the maternity leave module of the LifePaths microsimulation model at Statistics Canada. The goal of the LifePaths project is to construct a dynamic microsimulation model encapsulating as much detail as possible on socio-economic processes in Canada, as well as the historical patterns of change in those processes. LifePaths has been employed in a broad range of policy analysis and research activities. Examples include Canada Student Loan policy (under contract to Human Resources Development Canada and the Government of Ontario), returns to education (Appleby, Boothby, Rouleau and Rowe 1999), time use (Wolfson and Rowe 1996; Wolfson 1997; Wolfson and Rowe 1998a), tax-transfer and pensions (Wolfson, Rowe, Gribble and Lin 1998; Wolfson and Rowe 1998b), and labour force careers (Rowe and Lin 1999). In addition, the task of assembling data for LifePaths has required new research into, for example, educational careers (Chen and Oderkirk 1997; Rowe and Chen 1998; Plager and Chen 1999) and earnings correlation (Chen and Rowe 1999).

LifePaths is intended to incorporate socio-economic information from all relevant sources available to Statistics Canada. Consequently the construction of the model has motivated research into application of methodologies for exploiting multiple data sources. Embedding an estimated model in LifePaths is a powerful tool for deriving implications of the model that can be compared to information

1. Ian Cahill, Partnership and Continuous Evaluations, HRDC, 140 Promenade du Portage, Phase IV 3rd floor, Room 3D475, Gatineau, Québec K1A 0J9, and Edward J. Chen, Household Survey Methods Division, Statistics Canada, R.H. Coats Building 16th floor, Tunney's Pasture, Ottawa, Ontario K1A 0T6.

from other sources. For example, Rowe and Lin (1999) derived job tenures by simulation from a model estimated using short-period longitudinal data, then compared the results with data from a cross-sectional survey. We report on one aspect of the continuing effort to build a tool providing the maximum information that can be extracted from Statistics Canada's data sources.

The paper is organized to illustrate the way in which technical problems are often encountered in the course of building LifePaths, and how their solution is integrated with the model development process. To do this, a fair amount of background detail on associated issues is provided. Section 2 outlines the context of the benchmarking problem, and section 3 presents the theory behind our solution, with some possible extensions for further work. Section 4 describes the models to which it will be applied, including some details concerning the estimation of their parameters in the base period, then section 5 describes the application of the benchmarking method to these models. We display and discuss our empirical results in section 6, then close with some overall conclusions in section 7.

2. Context of the Problem

We provide context in this section by presenting an overview of the LifePaths model structure, a brief description of data sources involved, and a discussion of how the benchmarking problem arose.

2.1 Structure of the LifePaths Model

The LifePaths model simulates individual lifetimes as a series of events which modify the set of "state variables" describing the demographic, social, and economic circumstances of the individual. Waiting times to every possible event are associated with an individual, although they may be infinite. The waiting times may be conditioned on the values of state variables. The event type with the shortest waiting time occurs (its associated functions are called). Modification of any state variable at the occurrence of an event may lead to the generation of new waiting times for other events.

LifePaths initialises a case by randomly generating a "dominant" individual's sex, province of residence, age at immigration and year of birth. The year of birth can range from 1892 to 2051. Mortality and immigration assumptions are designed to reproduce provincial age-sex structures. When a dominant individual marries, enters a common-law union, or has a child, a non-dominant individual of suitable characteristics is created and is linked to the dominant individual, forming part of the case. Once created, non-dominant individuals undergo the same possible events as dominant individuals. However, since their purpose is to complete the profile of the dominant actor, they are usually filtered from all tabular reports.

LifePaths presently includes models of fertility, mortality, marriage (including common-law unions), educational careers, labour force careers, maternity leave, hours of work, earnings, taxes, and transfers. The model of the labour force careers describes transitions between the states "paid employee," "self-employed," and "not employed." It also includes a model of retirement and student work. The model of secondary and post-secondary educational careers at the provincial level is mature and highly developed.

2.2 The Data Sources

The estimation of base parameters for the model of maternity leave was carried out using data from SLID covering maternity leaves beginning in the period 1993-1996. Using data from 1997 allowed us to follow most maternity leaves to completion rather than using extensively censored data. This is a household survey designed to permit both longitudinal and cross-sectional analysis of people's financial and work situations. Starting in 1993, SLID follows the same respondents for six years, with new rotation groups introduced every three years. Each rotation groups includes about 15,000 households with 30,000 adults. From this survey we obtain the month of child birth, monthly data on labour force status, and a rich set of explanatory variables including job tenure, an indicator of self-employment, birth order of the child, presence of an employed spouse, province of residence, education level, and age. We can also determine if a mother who left a job within 4 months of birth has returned to the same job within 16 months. This is used as a practical definition of maternity leave and becomes our unit of analysis, with a slight expansion to include the 1% of cases where a mother returned to a different job from a labour market state of absence in the previous month. Using this unit of analysis we get a sample size of 835 births. As we show in section 6, this sample size is adequate to reveal some key explanatory factors. More precisely, several factors are found to be significant at the 95% confidence level. This sample contains about 730 unique mothers, representing over 87% of the sample of births. This means that there will be some correlation between observations as a result of those mothers who have two or more maternity leaves within the observation period, but we did not feel that it is of sufficient magnitude to warrant any special statistical tools.

The LFS is a monthly household survey focussing on labour force status, and also reporting a number of demographic characteristics. The survey is normally used exclusively for cross-sectional analysis. For the LifePaths project, however, a file covering the period from 1976 to 1995 was constructed that follows individuals as they rotate through the six monthly rotation groups of the survey, providing a six-month window on each individual's labour market activity. Since the number and ages of children are recorded each month, it is possible to observe the

appearance of a new child. Since all surveys throughout the period are used, the sample size is very large, and about 26,000 births are observed.

In the LFS window we note the labour force status of a new mother when the child is first reported. This is the key to estimating the probability of choosing a maternity leave, rather than leaving the labour force. We begin by considering $P(\mathrm{E})$, the proportion of such mothers who are employed. If the mother is "employed, at work," we suppose that they took a brief absence from their job – less than a month. If they are "employed, absent from work," it may be that they have chosen to take a maternity leave absence from their job and then return to it. However this may not always be the case. A new mother who we observe as employed and absent (EA) may later make a transition out of employment (to NE). To correct for this, considering mothers with a child of age less than a year observed in a window, we calculate the proportion $P(\mathrm{EA} \to \mathrm{NE})$ of transitions out of the "employed, absent from work" state that are to a not-employed state. We also estimate the proportion $P(\mathrm{NE} \to \mathrm{OJ})$ of mothers who return to an old job (OJ) after having left employment. The estimate is obtained by using observations on mothers with a young child who make transitions from a not-employed state to a job with a start date earlier than the previous month. Our estimate of the probability of choosing a maternity leave is now $P(\mathrm{E}) - P(\mathrm{EA} \to \mathrm{NE}) + P(\mathrm{NE} \to \mathrm{OJ})$.

It is also possible to observe mothers with a child of age less than a year making a transition from the status "employed, absent from work for personal or family responsibilities" to the status "employed, at work." We use this transition as a proxy for the return to work after a maternity leave. Since the duration of absence is reported in the previous month, this is the key to benchmarking the survival model.

The preceding discussion illustrates the weakness of the LFS data for a study of maternity leave, relative to SLID data. In addition to having fewer explanatory variables available than in SLID, we must accept proxies for the dependent variables. Nevertheless, we require the historical depth of the LFS. This relationship between the data sets is the context of the benchmarking problem described in the next section.

Both the SLID and the LFS have complex sample designs involving detailed stratification, and complex methods for calculating observation weights. We always make use of observation weights, both in estimation and in the calculation of frequencies. The methods used are fairly simple, and are discussed in sections 4 and 5.

2.3 The Benchmarking Problem

The context of our benchmarking problem is a model of women choosing between leaving the labour force or taking a maternity leave, and if they choose a leave, deciding how long that leave should be. The first decision is represented by a binary logit model, and the second by a semiparametric survival model, both including a vector of explanatory variables and associated parameters. In LifePaths, the decisions are made as part of the maternity leave choices event, which always occurs in the middle of a pregnancy. SLID is quite adequate for estimation of the base parameters of both these models. However, since a major goal of the LifePaths project is to incorporate historical patterns of change in socio-economic processes, it was necessary to benchmark the SLID parameter estimates to annual estimates of dependent variable means obtained from the LFS.

In this problem, we assume stable observed characteristics of the population. There are two reasons for this. First, LifePaths is a work in progress, and the benchmarking exercise we report on was carried out at a stage when other parts of the model that predict these characteristics were being extensively revised. In section 3.3, we touch on the consequences of evolving population characteristics. Second, we suppose that the primary reason for systematic change in observed outcomes between time periods is change in some factors not included in the measured characteristics of individuals. In the case of our application we observed a trend towards choice of maternity leave over leaving the labour force which seems to be due to social change rather than changes in the composition the population. We also observed a change in the distribution of maternity leave durations that appears to be due to changes in the Unemployment Insurance (UI) program implemented in Bill C–21 in 1990. At that time Parental Benefits were introduced, which extended the period during which many mothers could receive benefits from 15 to 25 weeks. Many mothers return to work at a time close to when they have exhausted UI benefits.

3. Benchmarking Methodology

In this section we present the method in an abstract form in order to clarify the assumptions, develop notation, and to reveal the similarity between the application to binary choice and to survival analysis.

3.1 Application to Binary Choice

The basic model for the benchmarking methodology relates to binary choice. Since we are not primarily interested in changes in the population, we simplify the analysis by assuming that the explanatory variables or individual characteristics in period τ are represented by a series of independent identically distributed random X^{τ}. We recognise that this is quite a strong assumption. Nevertheless, for the reasons discussed in section 2.3, we use it our empirical work. Section 3.3 shows that it is a fairly simple matter to extend the theory to incorporate trends in the independent variables.Consider a linear predictor given by

$$\eta^{\tau}(x) = \beta' x + \gamma^{\tau} \qquad (3.1)$$

where β is a vector of coefficients constant over time, x is a possible outcome of X^τ, and γ^τ represents a parameter specific to period τ. Notice that x contains no "constant term." Let Y^τ be a random variable, jointly distributed with X^τ, that takes the values 1 if an event occurs and 0 if it does not. Suppose that the probability of the event, conditional on characteristics x, is given by

$$E(Y^\tau \mid X^\tau = x) = \pi^\tau(x) = F(\eta^\tau(x)) \tag{3.2}$$

where we require F to be a continuous distribution function. The values of the function will then be bounded by zero and one, and it will have an inverse g, so that

$$\eta^\tau(x) = g(\pi^\tau(x)). \tag{3.3}$$

In the context of generalised linear models, g is called a link function. We begin by finding maximum likelihood estimates of the base parameters $\hat{\beta}$ and $\hat{\gamma}^{\tau_0}$ using data for the time period τ_0 (in our case this is the period when SLID data are available). Of course these data must include variables corresponding to outcomes of both X^τ and Y^τ. It remains to estimate γ^τ for each period τ. Equations (3.1) and (3.3) imply that

$$\begin{aligned} E\{\eta^\tau(X^\tau) - \eta^{\tau_0}(X^{\tau_0})\} &= \gamma^\tau - \gamma^{\tau_0} \\ &= E\{g(\pi^\tau(X^\tau))\} - E\{g(\pi^{\tau_0}(X^{\tau_0}))\}. \end{aligned} \tag{3.4}$$

Since we have observations only on the outcomes of Y^τ from the LFS for every period, we estimate the terms γ^τ by

$$\hat{\gamma}^\tau = \hat{\gamma}^{\tau_0} + g(\hat{\pi}^\tau) - g(\hat{\pi}^{\tau_0}) \tag{3.5}$$

where $\hat{\pi}^\tau$ is an estimate of $E(Y^\tau)$. Using the LFS, this estimate is the weighted frequency of the event in the time period τ (taking each weight from the month where a child is first observed). To justify this procedure we use equation (3.4) and assume an approximation

$$\begin{aligned} E\{g(\pi^\tau(X^\tau))\} - E\{g(\pi^{\tau_0}(X^{\tau_0}))\} \cong\ & g(E\{\pi^\tau(X^\tau)\}) \\ & - g(E\{\pi^{\tau_0}(X^{\tau_0})\}). \end{aligned} \tag{3.6}$$

Inaccuracy will arise due to Jensen's inequality in regions where g is convex or concave. Nevertheless, if g can be locally approximated by a linear function in the regions where $\pi^\tau(X^\tau)$ and $\pi^{\tau_0}(X^{\tau_0})$ are concentrated, then (3.6) may be quite accurate. The fact that g has an inflection point at 0.5 may aid the approximation when probabilities are dispersed around this value.

Fortunately we are able to test the adequacy of the estimator by simulating the estimated model in LifePaths and comparing the predicted frequencies of the event with corresponding weighted frequencies observed in the data. The results indicate that it is quite adequate for our application.

3.2 Application to Survival Analysis

We will show in section 5.2 that the approach outlined above can also be extended for use with a semiparametric survival model by adding an index t representing the duration in the current state, so that (3.5) becomes

$$\hat{\gamma}^\tau(t) = \hat{\gamma}^{\tau_0}(t) + g(\hat{\pi}^\tau(t)) - g(\hat{\pi}^{\tau_0}(t)) \tag{3.7}$$

where $\hat{\pi}^\tau(t)$ represents the empirical hazard function.

3.3 Trends in the Independent Variables

The benchmarking method may be improved by taking the changes in observed characteristics into account. As we noted in section 2.3, this would be considered when other parts of LifePaths are in a more mature form. To do this we relax the assumption that the random vectors X^τ are identically distributed. Equation (3.4) then becomes

$$\begin{aligned} E\{\eta^\tau(X^\tau) - \eta^{\tau_0}(X^{\tau_0})\} &= \gamma^\tau - \gamma^{\tau_0} \\ &\quad + \beta'\{E(X^\tau) - E(X^{\tau_0})\} \\ &= E\{g(\pi^\tau(X^\tau))\} \\ &\quad - E\{g(\pi^{\tau_0}(X^{\tau_0}))\} \end{aligned} \tag{3.8}$$

Based on this, we might estimate γ^τ by

$$\hat{\gamma}^\tau = \hat{\gamma}^{\tau_0} + g(\hat{\pi}^\tau) - g(\hat{\pi}^{\tau_0}) - \hat{\beta}'(\bar{x}^\tau - \bar{x}^{\tau_0}) \tag{3.9}$$

where $\bar{x}^\tau$ is the vector of mean values of the characteristics in period τ. Of course it may not be possible to obtain all of the mean values from the same data source. The method would extend to the survival model case in the same manner as (3.7) to give

$$\begin{aligned} \hat{\gamma}^\tau(t) = \hat{\gamma}^{\tau_0}(t) &+ g(\hat{\pi}^\tau(t)) - g(\hat{\pi}^{\tau_0}(t)) \\ &- \hat{\beta}'(\bar{x}^\tau(t) - (\bar{x}^{\tau_0}(t)). \end{aligned} \tag{3.10}$$

4. Models and the Estimation of Base Parameters

As explained in section 3.1, the base parameters $\hat{\beta}$ and $\hat{\gamma}^{\tau_0}$ are estimated by maximum likelihood using data from the period τ_0. We use data from SLID on all maternity leaves beginning in the period 1993-1996 (our base period τ_0). We do not attempt to estimate annual changes in the constant term γ throughout this period.

4.1 The Binary Logit Model

We adopt the logit model to represent a mother's choice between taking a maternity leave and withdrawing from the labour force. From now on we adopt a more conventional econometrics notation and use a subscript i to index a random variable or outcome associated with an individual i. We suppose that a random variable Y_i^τ takes values 0 or 1, with $Y_i^\tau = 1$ indicating that new mother i with vector of characteristics x_i in period τ chooses to take a maternity leave, conditional on her having been employed, and that

$$\pi_i^\tau = P(Y_i^\tau = 1) = F(\eta_i^\tau) = \frac{\exp(\eta_i^\tau)}{1+\exp(\eta_i^\tau)} \qquad (4.1)$$

where $\eta_i^\tau = \beta' x_i + \gamma^\tau$ is the linear predictor of equation (3.1) and F is the logistic distribution function. We estimate the base parameters $\hat{\beta}$ and $\hat{\gamma}^{\tau_0}$ using N observations from SLID by maximising the log-likelihood $\ln L(\beta, \gamma^{\tau_0})$ where

$$\begin{aligned} L(\beta,\gamma^\tau) &= P(Y_1 = y_1, Y_2 = y_2, \ ..., \ Y_N = y_N) \\ &= \prod_{y_i=0} [1 - F(\eta_i^\tau)] \prod_{y_i=1} F(\eta_i^\tau) \\ &= \prod_i [F(\eta_i^\tau)]^{y_i} [1 - F(\eta_i^\tau)]^{1-y_i} \qquad (4.2) \end{aligned}$$

and

$$\ln L(\beta,\gamma^\tau) = \sum_i \left\{ \begin{array}{l} y_i \ln F(\eta_i^\tau) \\ \quad + (1-y_i)\ln[1 - F(\eta_i^\tau)] \end{array} \right\}. \qquad (4.3)$$

Longitudinal SLID weights in the year of the child's birth are scaled to sum to the sample size, and are then used to weight the terms of the log-likelihood and its derivatives. The weighted score equations are

$$\begin{aligned} \frac{\partial L(\beta,\gamma^\tau)}{\partial \beta} &= \sum_i w_i x_i y_i - \sum_i w_i x_i F(\eta_i^\tau) = 0 \\ \frac{\partial L(\beta,\gamma^\tau)}{\partial \gamma^\tau} &= \sum_i w_i y_i - \sum_i w_i F(\eta_i^\tau) = 0. \qquad (4.4) \end{aligned}$$

The solution, which maximises the log-likelihood, was found by Newton-Raphson iteration. The logit model has been used often by statisticians and econometricians, and there is an extensive literature. For example, see Chambless and Boyle (1985), Roberts, Rao, and Kumar (1987), and Morel (1989).

4.2 The Semiparametric Survival Model: Basic Form

For mothers who have chosen to take a maternity leave from their job, we use a survival model to describe the duration of their leave. The probability density function (pdf) of the distribution has a complex shape, as can be seen from the graphs in section 6.4. There is spike at durations of less than a month and a mode which appears to represent the maximum Unemployment Insurance special benefits entitlement available to mothers after 1990 (15 weeks of Maternity Benefits, plus 10 weeks of Parental Benefits, plus a two-week waiting period). We began the study by estimating various fully parametric models, including a log-logistic survival model combined with a logit model to predict durations of less than a month, but were unable to obtain an adequate fit. To solve this problem, we follow Prentice and Gloeckler (1978), Han and Hausman (1986) and Meyer (1990), by nonparametrically estimating the effect of time on the hazard of returning to work. The hazard of returning to work is specified in a proportional hazards form:

$$\lambda_i^\tau(t) = \lambda_0^\tau(t) \exp\{\beta' x_i(t)\} \qquad (4.5)$$

where $\lambda_0^\tau(t)$ is the unknown baseline hazard at leave duration t and time period τ, $x_i(t)$ is a vector of explanatory variables for mother i, and β is a vector of coefficients. The data tell us which of the intervals [0,1), [1,2), [2,3), … contains the spell duration (in our case the units are months), and the model can be interpreted as an incompletely observed continuous time hazard model with no restriction on the form of the baseline hazard. If T_i^τ is the duration of leave for mother i during period τ, then for $t = 1, 2, 3, \ldots$, the probability that the spell lasts until time t, given that it has lasted until $t-1$, can be written as

$$\begin{aligned} P(T_i^\tau > t \mid T_i^\tau \geq t-1) &= \exp\left[-\int_{t-1}^{t} \lambda_i^\tau(u)\,du \right] \\ &= \exp\left[-\exp\{\beta' x_i(t)\} \int_{t-1}^{t} \lambda_0^\tau(u)\,du \right] \qquad (4.6) \end{aligned}$$

if we assume that $x_i(t)$ is constant on the interval between $t-1$ and t. In order to apply the theory of section 3, we can rewrite equation (4.6) as

$$\begin{aligned} 1 - \pi_i^\tau(t) &= P(T_i^\tau \geq t \mid T_i^\tau \geq t-1) \\ &= \exp[-\exp\{\beta' x_i(t) + \gamma^\tau(t)\}] \\ &= \exp[-\exp\{\eta_i^\tau(t)\}] \qquad (4.7) \end{aligned}$$

where

$$\gamma^\tau(t) = \ln\left[\int_{t-1}^{t} \lambda_0^\tau(u)\,du \right]. \qquad (4.8)$$

One may censor any ongoing observations at some large duration T. Again we can estimate the base parameters $\hat{\beta}$ and $\hat{\gamma}^{\tau_0}$ using N observations from SLID by maximising the log-likelihood $\ln L(\gamma^{\tau_0}, \beta)$. Since we will always be referring to data from the base period for the remainder of section 4, we drop superscripts τ_0.

The likelihood function is given by

$$L(\gamma,\beta) = \prod_{i=1}^{N} \left[\begin{array}{c} [1 - \exp\{-\exp(\eta_i(k_i))\}]^{\delta_i} \\ \prod_{t=1}^{k_i} \exp\{-\exp(\eta_i(t))\} \end{array} \right] \qquad (4.9)$$

where $\gamma = [\gamma(1), \gamma(2), \ ..., \ \gamma(T)]'$, C_i is a censoring time, $\delta_i = 1$ if $T_i \leq C_i$ and 0 otherwise, $k_i = \min(\text{int}(T_i), C_i)$. The log-likelihood is therefore

$$\ln L(\gamma, \beta) = \sum_{i=1}^{N} \left[\begin{array}{r} \delta_i \ln[1 - \exp\{-\exp(\eta_i(k_i))\}] \\ - \sum_{t=1}^{k_i} \exp(\eta_i(t)) \end{array} \right]. \quad (4.10)$$

Weights from the months that a child is first observed are scaled to sum to the sample size, and then used to weight the terms of the log-likelihood function and its derivatives. The weighted log-likelihood function is maximised by the quasi-Newton algorithm of Broyden, Fletcher, Goldfarb, and Shanno (BFGS), using an implementation based on Dennis and Schnabel (1983).

4.3 The Semiparametric Survival Model: with Work-to-Birth Gap Decision

The situation in our application is complicated somewhat by our desire to model the duration from leaving the job until the birth (the work-to-birth gap), as well as the hazard of returning to work from a maternity leave. The model of work-to-birth gap is estimated separately, based on SLID data. Examination of the mean gap duration for each year in the LFS data indicates that this duration has been fairly stable over time, so the model is not benchmarked. Nevertheless, a modification of the semiparametric survival model is necessary to incorporate the separate model of work-to-birth gap. This can be accomplished by assuming that the work-to-birth gap decision, possibly involving health considerations, acts to constrain the desired total duration. This means that the above model would apply to the desired total duration, which is unobservable, and might be labelled T^*.

In cases where the desired duration was shorter than the work-to-birth gap, the mother might return to work as soon as possible after the birth. This means that in cases where we observe a significant work-to-birth gap (greater than a month), and the mother returns soon after birth (within a month), all that is known about desired duration is that

$$T^* \leq T$$

where T is the total duration of leave. This is equivalent to a situation labelled "left censoring" by Cox and Oaks (1984, page 178), where observation does not start immediately and some individuals have already failed before it does.

From such an observation we get a contribution to the likelihood function and its logarithm given by

$$L_i = 1 - \prod_{t=1}^{k_i} P(T^* \geq t \mid T^* \geq t-1) = 1 - \prod_{t=1}^{k_i} \exp[-\exp(\eta_i(t))] \quad (4.11)$$

and

$$\ln(L_i) = \ln\{1 - \exp[-\sum_{t=1}^{k_i} \exp(\eta_i(t))]\}. \quad (4.12)$$

Unfortunately the log-likelihood expression does not simplify like the corresponding expression for "right-censored" observations. In spite of this, Monte Carlo experiments indicate that estimation is not a problem even in heavily censored data sets.

Longitudinal SLID weights in year of the child's birth are used in same manner as for the basic form of the survival model.

5. Benchmarking the Models

To begin the benchmarking procedure we must invert the distribution function F given in equation (3.2) to find the link function g. We then apply equation (3.5) in the case of the logit model, and equation (3.7) in the case of the survival model.

5.1 Application to the Binary Logit Model

To benchmark the logit model we first invert the logistic distribution function in equation (4.1) to obtain

$$\eta_i^\tau = g(\pi_i^\tau) = \ln\left(\frac{\pi_i^\tau}{1-\pi_i^\tau}\right) \quad (5.1)$$

where g is the well-known logit function. We can then apply equation (3.5) and (5.1) to obtain

$$\hat{\gamma}^\tau = \hat{\gamma}^{\tau_0} + g(\hat{\pi}^\tau) - g(\hat{\pi}^{\tau_0}) = \hat{\gamma}^{\tau_0} + \ln\left(\frac{\hat{\pi}^\tau/(1-\hat{\pi}^\tau)}{\hat{\pi}^{\tau_0}/(1-\hat{\pi}^{\tau_0})}\right) \quad (5.2)$$

where for $\tau < \tau_0$, each $\hat{\pi}^\tau$ is the frequency of choosing maternity leave calculated from LFS data for maternity leaves beginning in year τ, and $\hat{\pi}^{\tau_0}$ is the frequency from SLID data.

5.2 Extension to the Survival Model

From equation (4.7) we get

$$\pi_i^\tau(t) = 1 - \exp[-\exp\{\eta_i^\tau(t)\}] = F\{\eta_i^\tau(t)\} \quad (5.3)$$

where

$$\eta_i^\tau(t) = \beta' x_i(t) + \gamma^\tau(t). \quad (5.4)$$

In this case F is an extreme value distribution that is easily inverted to obtain

$$\eta_i^\tau(t) = \ln[-\ln(1-\pi_i^\tau(t))] = g(\pi_i^\tau(t)). \quad (5.5)$$

For benchmarking we can use equation (3.7) with the observed frequencies in period τ represented by the empirical hazard or occurrence/exposure ratio given by

$$\hat{\pi}^{\tau}(t) = d^{\tau}(t) / r^{\tau}(t) \qquad (5.6)$$

where, for spells beginning in period τ, $d^{\tau}(t)$ is the number of mothers who fail in the interval $(t-1, t]$ and $r^{\tau}(t)$ is the number of mothers in view at duration t, including those censored at time t (censoring can only occur at the end of intervals). Numbers of mothers were calculated from sample counts by applying the LFS weight from the month that a new mother returns to work. The empirical hazard and the corresponding estimator for the survivor function implied by the product law of probabilities were studied by Kaplan and Meier (1958). The use of the empirical hazard in equation (3.7) together with equation (5.5) yields

$$\hat{\gamma}^{\tau}(t) = \hat{\gamma}^{\tau_0}(t) + \ln\left(\frac{\ln[1-\hat{\pi}^{\tau}(t)]}{\ln[1-\hat{\pi}^{\tau_0}(t)]}\right). \qquad (5.7)$$

6. Empirical Results

The results of estimation in the base period, and the results of simulation with benchmarked parameter estimates are presented for both models. The simulation results are compared with annual survey sample frequencies of choosing a maternity leave in the case of the logit model, and with annual survey frequency distributions of maternity leave duration in the case of the survival model.

6.1 Estimation Results for the Binary Logit Model

The estimation results obtained from estimating the logit model from SLID data are presented in Table 1. Omitted dummy variable categories, which form the reference categories for the variables used in the model, were province of residence Ontario and highest education level "some post secondary." Individual and family income variables were tested, but were found not to be significant, and so were not included in the regression.

There may be some bias in the estimates, particularly those of the standard errors, due to the fact that the complex SLID sample design was accounted for only through the weights applied to the log-likelihood.

The significant positive effect of job tenure seems reasonable for a number of reasons. A lengthy tenure might indicate that the woman has acquired firm-specific human capital and has achieved some seniority. It would also be an indicator of strong attachment to the labour force generally. On the firm side, the longer the woman's job tenure, the longer the leave that the firm is likely to grant with a guarantee that she can return to her job. Also, provincial government guarantees of job security also depend on job tenure. Finally, a lengthy job tenure means that the woman will likely meet the Unemployment Insurance eligibility requirements (20 weeks of insured employment). A dummy variable indicating that UI entrance requirements were met was tested and found to be just significant at the 5% level. However, because we are not able at this stage to model changes in the UI program through the influence of covariates, because of uncertainty in interpretation, and because of high correlation with job tenure, it was not included. In the LFS, self-employed workers are reported as having a transition out of employment only when they terminate their business. Since taking a leave simply means not terminating the business, a significant positive effect for the indicator of self-employment is to be expected. Having been self-employed before the birth increases the odds of taking a maternity leave by 333%, the strongest effect that we see for an indicator variable.

Table 1
Binary Logit Parameter Estimation Results

Parameter	Estimate of Coefficient	Contribution to Odds Ratio*	Std Error of Coefficient	Prob-Value
Constant	−6.432	0.002	2.995	0.0318
NFLD	−0.829	0.436	0.741	0.2636
PEI	0.931	2.537	1.612	0.5633
NS	−0.456	0.634	0.541	0.3992
NB	0.207	1.230	0.675	0.7596
QUE	−0.361	0.697	0.247	0.1437
MAN	−0.490	0.613	0.503	0.3306
SASK	−0.163	0.850	0.458	0.7218
ALTA	−0.200	0.819	0.325	0.5379
BC	−0.120	0.887	0.300	0.6899
Job Tenure (mths)/10	0.094	1.099	0.026	0.0003
Self-employed?	1.203	3.330	0.418	0.0040
Age (Years)	0.479	1.614	0.199	0.0160
(Age^2)/10	−0.071	0.931	0.033	0.0296
< High School Grad	−0.702	0.496	0.357	0.0490
High School Grad	−0.148	0.862	0.276	0.5913
University Grad	−0.292	0.747	0.229	0.2027
First Child?	−0.525	0.592	0.192	0.0063

Log-likelihood = −381.553

Number of Observations = 835

Observations are given the SLID longitudinal weight from the year of birth, scaled to sum to the sample size

* This is the exponential of the coefficient. It may be interpreted as the proportional change in the odds ratio due to a unit change in the corresponding independent variable.

The effect of the first child indicator also seems reasonable. The odds for maternity leave for a first-time mother is only 59% of the odds for maternity leave for a mother of more than one child, given that all other characteristics are the same – *i.e.* first-time mothers are more inclined to job separation than the mothers who already have children. This may be partly a consequence of the fact that our sample consists of mothers who have been employed within 4 months of the birth. Mothers who have more than one child tend to space them within a few years at most. If they are employed just before a second or subsequent births, they will have already demonstrated that they returned to work after an absence that must have been less than the gap between births. This at least rules out some common patterns of withdraw from the labour force – for example staying at home until all children are in school.

The effect of age is more difficult to interpret since the effect on the log-odds ratio is non-linear. By drawing a graph of the term $-0.479 * age - 0.0071 * age^2$ one can see that, as age increases, the log-odds of taking a maternity leave first increases, but that the rate of increase declines until a level point at the maximum log-odds is reached by the age of 34. Since the number of mothers declines considerably after this age, the subsequent decline may not be meaningful. One might hazard a conjecture that, among young mothers, being relatively older indicates more attachment to the labour force and thus a stronger tendency to take a maternity leave, while among older mothers, who are past the stage of first entering the labour force, this effect is reduced. However, the results are probably not precise enough to draw any firm conclusion about this.

6.2 Simulation Results for the Benchmarked Binary Logit Model

The benchmarking exercise consists of adjusting the constant term of the model in the manner described by (5.2) for each year in the period 1975 – 1992. The constant term is not adjusted after 1992, partly because the LFS data do not indicate a strong trend after 1992. The model is then incorporated in LifePaths and a simulation is run. For each year from 1976 to 1995, Figure 1 shows both the frequency of choosing a leave in the LifePaths simulation, and the frequency estimated from the LFS. For the period 1993 – 1995, estimates from SLID are also presented.

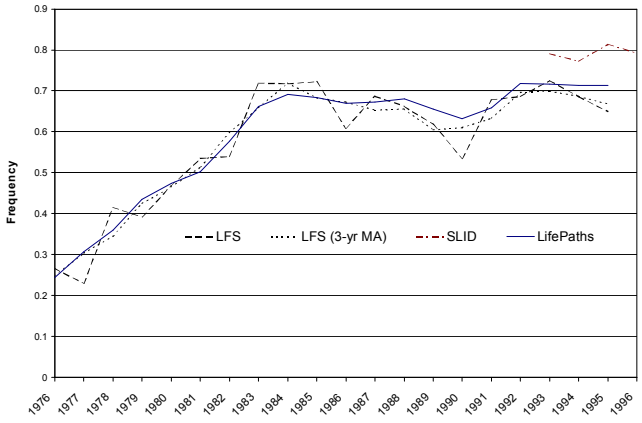


Figure 1. Frequency of Choosing a Maternity Leave 1976–1996.

The simulation captures the change over time revealed by the LFS data during the period 1976 – 1992. There is no benchmark adjustment implemented in the LifePaths simulation after 1992, so that the base parameters estimated from pooled SLID data 1993 – 1996 are effective. The simulated frequency is slightly lower than the observed SLID frequency during this period. Two possible sources of error are an insufficiently flexible specification of the binary choice model, and differences between the SLID estimates of explanatory variables and those provided by LifePaths.

6.3 Estimation Results for the Survival Model

The results obtained from estimating the semiparametric survival model from SLID data are presented in Table 2. As in the binary logit model estimation, omitted dummy variable categories were province of residence Ontario and highest education level "some post secondary." Since the dependent variable is the hazard of returning to work, a positive coefficient for a covariate indicates an influence that tends to shorten the duration of maternity leave.

The estimates of the constant terms in the duration-dependent linear predictor given by (4.7) are denoted in Table 2 by GAMMAi, $i = 1, 2, \ldots, 15$. This represents the influence of the baseline hazard incorporating the influence of duration.

Table 2
Survival Model Parameter Estimation Results

Parameter	Estimate	Std Error	Prob-Value
Job Tenure (mths) /10	−0.030	0.010	0.0024
NFLD	0.195	0.426	0.6470
PEI	0.307	0.490	0.5313
NS	0.173	0.253	0.4940
NB	0.109	0.293	0.7091
QUE	0.111	0.117	0.3411
MAN	−0.402	0.253	0.1116
SASK	−0.303	0.213	0.1539
ALTA	0.270	0.154	0.0798
BC	−0.440	0.148	0.0030
Self-Employed?	1.665	0.157	0.0000
Age	−0.253	0.041	0.0000
Age** 2 / 10	0.043	0.007	0.0000
First Child?	−0.301	0.090	0.0009
< High School Grad	0.508	0.206	0.0135
High School Grad	−0.124	0.125	0.3212
University Grad	−0.374	0.108	0.0006
Employed Spouse?	0.109	0.151	0.4703
Gamma1	2.570	0.609	0.0000
Gamma2	−1.136	0.816	0.1636
Gamma3	−0.466	0.719	0.5176
Gamma4	0.780	0.640	0.2232
Gamme5	1.425	0.627	0.0231
Gamma6	2.755	0.613	0.0000
Gamma7	3.640	0.612	0.0000
Gamma8	3.413	0.620	0.0000
Gamma9	3.465	0.630	0.0000
Gamma10	3.387	0.649	0.0000
Gamma11	4.579	0.655	0.0000
Gamma12	4.285	0.785	0.0000
Gamma13	3.645	1.110	0.0010
Gamma14	3.746	1.281	0.0034
Gamma15	6.215	2.415	0.0101

log-likelihood = −1165.06
Number of Observations 3,411
Observations are given the SLID longitudinal weight from the year of birth, scale to sum to the sample size

Again, individual and family income variables were tested and found not to be significant. Both this finding and the importance of a self-employment indicator as a predictor of early return to work accord with the findings of Marshall (1999). Marshall found that education variables were not significant in determining whether a mother would return to work within a month. We find however, that university graduation has a significant negative effect on the hazard (positive effect on duration). Job tenure has a significant negative effect on the hazard, possibly reflecting its relationship with Unemployment Insurance entitlement and job security.

6.4 Simulation Results for the Benchmarked Survival Model

In the case of the semiparametric survival model, benchmarking consists of adjusting all of the terms GAMMAi, $i = 1, 2, \ldots, 15$ of the previous section according to (5.8) for each of the years in the period 1975 – 1992. The model is then simulated as part of LifePaths.

The frequency distribution of simulated maternity leave durations is presented and compared to the corresponding observed frequency distribution from LFS data. In order to present the results, the frequencies in 3-year periods were averaged. A key feature of the frequency distribution is an abrupt change apparently due to the introduction of parental benefits with Bill C-21 at the end of 1990. Since mothers with maternity claims in progress at the time of implementation were entitled to parental benefits, the claims beginning in 1990 represent a mixture of regimes. For this reason the year 1990 is not included in any of the 3-year averages. In Figures 2 and 3 we use disjoint 3-year periods covering 1976 – 1984. To balance periods before and after 1990 using available data, in Figures 4 and 5 we use the overlapping periods 1985-1987, 1987-1989, 1991-1993, and 1993-1995.

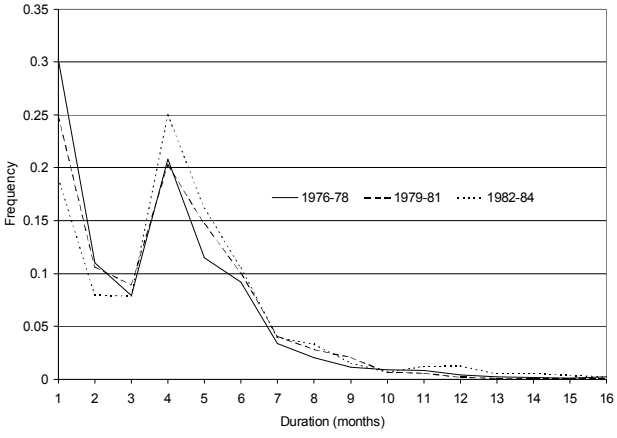


Figure 2. LifePaths: Distribution of Leave Durations for 1976–1984.

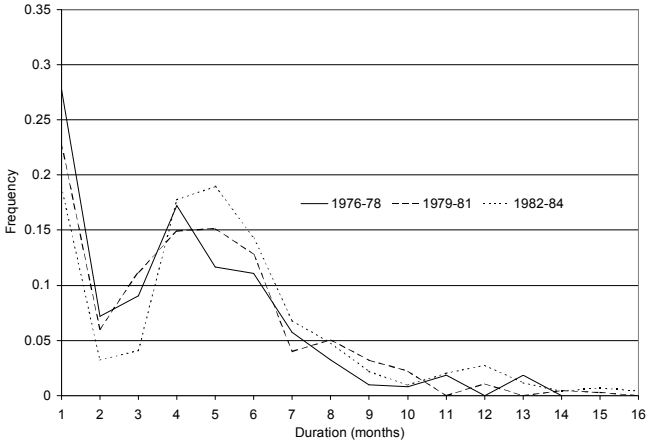


Figure 3. LFS Data: Distribution of Leave Durations for 1976–1984.

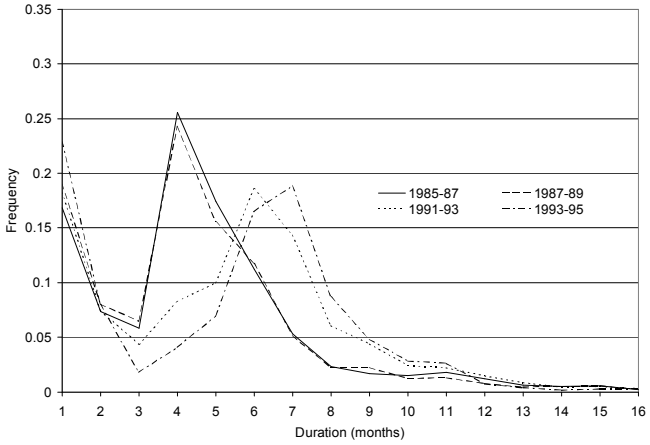


Figure 4. LifePaths: Distribution of Leave Durations for 1985–1989 and 1991–1995.

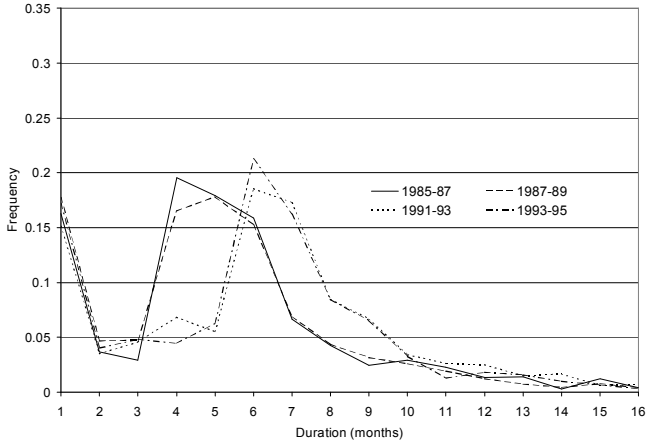


Figure 5. LFS Data: Distribution of Leave Durations for 1985–1989 and 1991–1995.

The distribution of durations derived from SLID data 1993 – 1996 is presented in Figure 6. This may be compared with the simulated data shown in Figure 4 for the period 1993 – 1995, since no benchmarking is applied after 1992.

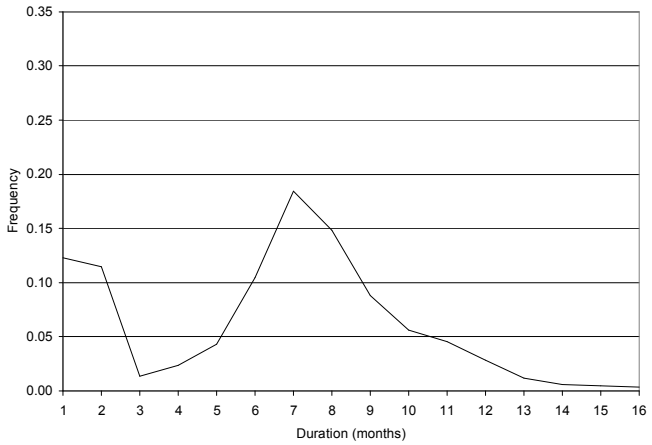


Figure 6. SLID Data: Distribution of Leave Durations for 1993 – 1996.

In Figure 7 we present the average duration of maternity leaves beginning in each year of the observed period. The average of simulated durations are compared with those from the surveys.

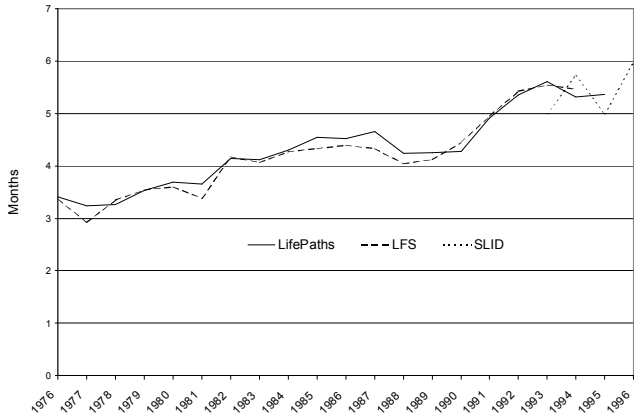


Figure 7. Average Duration of Maternity Leave 1976 – 1996.

6.5 Evaluation of Benchmarking Performance

The benchmarking method appears to be very effective in the case of the binary logit model. The trend of the LFS data is well reflected in the LifePaths simulation. In the case of the survival model, the key feature of the LFS data is the abrupt shift of the mode of the frequency distribution after 1990, apparently due to the introduction of parental benefits. This shift has been captured by the simulated data. Also the average duration of maternity leave in the simulation fits the LFS data very closely.

A noticeable divergence between the simulation and the LFS data is the height of the mode at the interval (3, 4] months in the frequency distribution of the durations from LifePaths from 1982 – 1989. This may be due to the effect of trends in the values of explanatory variables, which we have assumed to be stable. Further work is necessary to establish this. A possible extension to the model was discussed in section 3.3.

7. Conclusions

The technique that we have developed appears to be quite successful in benchmarking of the logit and survival model parameters so that the essential features of the LFS data are captured in LifePaths predictions. The key to benchmarking the logit model is the adjustment of the parameter corresponding to the "constant term" in the linear predictor that is imbedded it the logistic distribution function in order to predict the conditional expectation of the dependent variable. Section 3.1 develops the technique in a general framework that includes other models of binary choice. Particularly, it would extend to the popular probit model where a linear predictor is embedded in the standard normal distribution function. Benchmarking of the semi-parametric survival model hinges on the adjustment of all the parameters representing the baseline hazard. Our results illustrate how the entire shape of the distribution of durations predicted by the model can be made to evolve through time according to a pattern revealed by supplementary data.

Acknowledgements

The authors wish to express their thanks to Steve Gribble and members of the Socio-economic Modelling Group at Statistics Canada for useful comments throughout the development of the maternity leave module, to Geoff Rowe and Huan Nguyen for use of their computer program to follow individuals through rotations in the LFS, to Katherine Marshall for advice on the use of SLID and for sharing computer programs, to Adrienne ten Cate for fruitful discussions, and to an anonymous referee for several improvements. This work was performed when both authors worked in the Socio-Economic Modeling Group, Statistics Canada, R.H. Coats Building 24th floor, Tunney's Pasture.

References

Appleby, J., Boothby, D., Rouleau, M. and Rowe, G. (1999). Level and Distribution of Individual Returns to Post-Secondary Education: Simulation Results from the LifePaths Model. Presented at the 1999 meetings of the Canadian Economics Association.

Chambless, L.E., and Boyle, K.E. (1985). Maximum Likelihood methods for complex sample data: Logistic regression and discrete proportional hazards models. *Communications in Statistics*, *A: Theory and Methods*, 14, 177-192.

Chen, E.J., and Oderkirk, J. (1997). Varied Pathways: The Undergraduate Experience in Ontario, Feature article. *Education Quarterly Review*, Statistics Canada, 4, 3, 47-62.

Chen, E.J., and Rowe, G. (1999). Trend Correlation of Labour Market Earnings in Canada: 1982 to 1995. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, 173-179.

Cox, D.R., and Oaks, D. (1984). *Analysis of Survival Data*. New York: Chapman and Hall.

Dennis, J.E. Jr, and Schnabel, R.B. (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Englewood Cliffs, NJ: Prentice-Hall.

Han, A., and Hausman, J. A. (1986). Semiparametric Estimation of Duration and Competing Risk Models. M.I.T. Working Paper No. 450.

Kaplan, E.L., and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53, 457-81.

Marshall, K. (1999). Employment after childbirth. *Perspectives on Labour and Income*. Statistics Canada, Autumn 1999, 18-25.

Meyer, B.D. (1990). Unemployment Insurance and Unemployment Spells. *Econometrica*, 58, 757-782.

Morel, J.G. (1989). Logistic regression under complex survey designs. *Survey Methodology*, 15, 205-223.

Plager, L., and Chen, E.J. (1999). Student Debt from 1990–91 to 1995–96: An Analysis of Canada Student Loans Data. MAJOR RELEASES, *THE DAILY* and *Education Quarterly Review*, Statistics Canada, 5, 4, 10-35.

Prentice, R., and Gloeckler, L. (1978). Regression Analysis of Grouped Survival Data with Application to Breast Cancer Data, *Biometrics*, 34, 57-67.

Roberts, G.A., Rao, J.N.K. and Kumar, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.

Rowe, G., and Chen, E.J. (1998). An Increment-Decrement Model of Secondary School Progression for Canadian Provinces. *Proceedings*: *Symposium on Longitudinal Analysis for Complex Surveys*, Statistics Canada, 167-178.

Rowe, G., and Lin, X. (1999). Modelling Labour Force Careers for the LifePaths Simulation Model. *Proceedings: Symposium 99 Combining Data from Different Sources*, Statistics Canada, 57-64.

Wolfson, M.C. (1997). Sketching LifePaths: A New Framework for Socio-Economic Statistics. *Simulating Social Phenomena*, (Eds. Conte, R. Gegselmann and P. Terna), Lecture Notes in Economics and Mathematical Systems, 456, Springer.

Wolfson, M.C., and Rowe, G. (1996). Perspectives on Working Time Over the Life Cycle, Canadian Employment Research Forum Conference on Changes to Working Time, Ottawa.

Wolfson, M.C., and Rowe, G. (1998a). LifePaths – Toward an Integrated Microanalytic Framework for Socio-Economic Statistics. 26th General Conference of the International Association for Research in Income and Wealth, Cambridge, U.K.

Wolfson, M.C., and Rowe, G. (1998b). Public Pension Reforms – Analyses Based on the LifePaths Generational Accounting Framework, 26th General Conference of the International Association for Research in Income and Wealth, Cambridge, U.K.

Wolfson, M.C., Rowe, G., Gribble, S. and Lin, X. (1998). Historical Generational Accounting with Heterogeneous Populations. *Government Finances and Generational Equity* (Ed. M. Corak), Statistics Canada, 107-127.