

# BENCHMARKING PARAMETER ESTIMATES IN LOGIT MODELS OF BINARY CHOICE AND SEMIPARAMETRIC SURVIVAL MODELS

Ian Cahill and Edward Chen<sup>1</sup>

## 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 with 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: Benchmarking; Binary logit; Microsimulation; Semiparametric survival models.

## RÉSUMÉ

Une approche pour exploiter les données de nombreuses enquêtes et époques par l'analyse comparative des estimations de paramètres de modèles logit de choix binaire et de modèles de survie semiparamétrique est élaborée. L'objectif est d'exploiter la source relativement riche de covariables socio-économiques que contient l'Enquête sur la dynamique du travail et du revenu (EDTR), ainsi que la durée historique de l'Enquête sur la population active (EPA), ce qui est rehaussé par le suivi des individus par le biais d'une entrevue prévue dans le calendrier de renouvellement de six mois. L'application de la méthode est démontrée à l'aide du module de congé de maternité du projet de microsimulation dynamique du logiciel LifePaths de Statistique Canada. Le choix du congé de maternité plutôt que la cessation d'emploi est précisé comme modèle logit binaire, alors que la durée du congé est précisée comme modèle de survie de risque proportionnel semiparamétrique, avec covariables combiné à un scénario de risques qui change chaque mois. Au départ, les deux modèles sont estimés selon la méthode du maximum de vraisemblance à partir de données regroupées de l'EDTR de 1993 à 1997, puis comparées aux estimations annuelles des EPA de 1976 à 1992. Pour ce qui est du modèle logit, la variable explicative linéaire est ajustée par une estimation d'entrées dépareillées de l'EPA. Pour le modèle de survie, l'estimateur Kaplan-Meier de la fonction de risques de l'EAP est utilisé pour ajuster le risque prédit du modèle semiparamétrique.

MOTS CLÉS : Analyse comparative des estimations de paramètres; logit binaire; microsimulation; modèles de survie semiparamétriques.

## 1. INTRODUCTION

We develop a simple method of benchmarking the parameter estimates obtained from a survey rich in explanatory variables to estimates from a survey with significant historical depth. 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.

<sup>1</sup> Ian Cahill and Edward Chen, Socio-Economic Modeling Group, Statistics Canada, 24th floor, R.H. Coats Building, Tunney's Pasture, Ottawa, Canada K1A 0T6.

The work was carried out while developing the maternity leave module of the LifePaths microsimulation model at Statistics Canada. 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 1998), and tax-transfer and pensions (Wolfson, Rowe, Gribble and Lin 1998). In addition, the task of assembling data for LifePaths has required new research into, for example, educational careers (Chen and Oderkirk 1998; Rowe and Chen 1998; Plager and Chen 1999) and earnings correlation (Chen and Rowe 1999).

The paper is organized to illustrate the way in which technical problems are often encountered in the course of building LifePaths. Section 2 outlines the context of the benchmarking problem, and section 3 presents the theory behind our solution. 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 finish by showing results and drawing conclusions in section 6.

## **2. CONTEXT OF THE PROBLEM**

### **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.

### **2.2 The Data Sources**

The estimation of base parameters was carried out using data from the Statistics Canada Survey of Labour and Income Dynamics (SLID) covering maternity leaves beginning in the period 1993-1996. 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 panels introduced every three years. 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. The sample size of about 850 births is adequate for estimation.

The Canadian Labour Force Survey (LFS) conducted by Statistics Canada is a monthly household survey focussing on labour force status, and also reporting a number of demographic characteristics. For the LifePaths project a file covering the period from 1976 to 1995 was constructed that follows individuals as they rotate through the six monthly panels 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.

### **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. The LFS 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.

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 was carried out when other parts of the model that predict these characteristics were being extensively revised. Second, we believe that change in observed outcomes between time periods is mostly due to change in factors not included in the measured characteristics of individuals. For example, we 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

#### 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  $\tau$  are represented by a series of independent identically distributed random vectors  $X^\tau$ . Consider a linear predictor given by

$$\eta^\tau(x) = \beta'x + \gamma^\tau \quad (3.1)$$

where  $\beta$  is a vector of coefficients constant over time,  $x$  is a possible outcome of  $X^\tau$ , and  $\gamma^\tau$  represents a parameter specific to period  $\tau$ . Notice that  $x$  contains no "constant term." Let  $Y^\tau$  be a random variable, jointly distributed with  $X^\tau$ , 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

$$E(Y^\tau | X^\tau = x) = \pi^\tau(x) = F(\eta^\tau(x)) \quad (3.2)$$

where  $F$  is a distribution function with an inverse  $g$ , so that

$$\eta^\tau(x) = g(\pi^\tau(x)). \quad (3.3)$$

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  $\tau_0$  (in our case this is the period when SLID data is available). Of course these data must include variables corresponding to outcomes of both  $X^\tau$  and  $Y^\tau$ . It remains to estimate  $\gamma^\tau$  for each period  $\tau$ . Equations (3.1) and (3.3) imply that

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

Since we have observations only on the outcomes of  $Y^\tau$  from the LFS for every period, we estimate the terms  $\gamma^\tau$  by

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

where  $\hat{\pi}^\tau$  is the observed frequency of the event in the time period  $\tau$ . To justify this procedure we use equation (3.4) and assume an approximation

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

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 observed frequencies.

#### 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

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

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

### 4. 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  $\tau_0$ . We use data from SLID on all maternity leaves beginning in the period 1993-1996 (our base period  $\tau_0$ ).

#### 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  $Y_i^\tau$  takes values 0 or 1, with  $Y_i^\tau = 1$  indicating that new mother  $i$  with vector of characteristics  $x_i$  in period  $\tau$  chooses to take a maternity leave, conditional on her having been employed, and that

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

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 (pooling all years) by maximising the log-likelihood  $\ln L(\tau_0)$  where

$$\ln L(\tau) = \sum_i \{y_i \ln F(\eta_i^\tau) + (1 - y_i) \ln [1 - F(\eta_i^\tau)]\}. \quad (4.2)$$

Longitudinal SLID weights in 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.

## 4.2 The Semiparametric Survival Model

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. We follow 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:

$$\lambda_i^\tau(t) = \lambda_0^\tau(t) \exp\{\beta'x_i(t)\} \quad (4.3)$$

where  $\lambda_0^\tau(t)$  is the unknown baseline hazard at leave duration  $t$  and time period  $\tau$ ,  $x_i(t)$  is a vector of explanatory variables for mother  $i$ , and  $\beta$  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  $T_i^\tau$  is the duration of leave for mother  $i$  during period  $\tau$ , then for  $t = 1, 2, 3, \dots$ , the probability that the spell lasts until time  $t$ , given that it has lasted until  $t - 1$ , can be written as

$$P(T_i^\tau > t | T_i^\tau \geq t - 1) = \exp\left[-\int_{t-1}^t \lambda_i^\tau(u) du\right] = \exp\left[-\exp\{\beta'x_i(t)\} \int_{t-1}^t \lambda_0^\tau(u) du\right] \quad (4.4)$$

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 (4.4) as

$$1 - \pi_i^\tau(t) = P(T_i^\tau \geq t | T_i^\tau \geq t - 1) = \exp[-\exp\{\beta'x_i(t) + \gamma^\tau(t)\}] = \exp[-\exp\{\eta_i^\tau(t)\}] \quad (4.5)$$

where

$$\gamma^\tau(t) = \ln\left[\int_{t-1}^t \lambda_0^\tau(u) du\right] \quad (4.6)$$

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  $\tau_0$ .

The log-likelihood function is given by

$$\ln L(\gamma, \beta) = \sum_{i=1}^N \left[ \delta_i \ln[1 - \exp\{-\exp(\eta_i(k_i))\}] - \sum_{t=1}^{k_i} \exp(\eta_i(t)) \right] \quad (4.7)$$

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

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

## 5. BENCHMARKING THE MODELS

### 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 get

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

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

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

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  $\tau$ , and  $\hat{\pi}^{\tau_0}$  is the frequency from SLID data.

## 5.2 Extension to the Survival Model

From equation (4.5) 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 get

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

For benchmarking we can use equation (3.7) with the observed frequencies in period  $\tau$  represented by the empirical hazard or Kaplan-Meier estimator given by

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

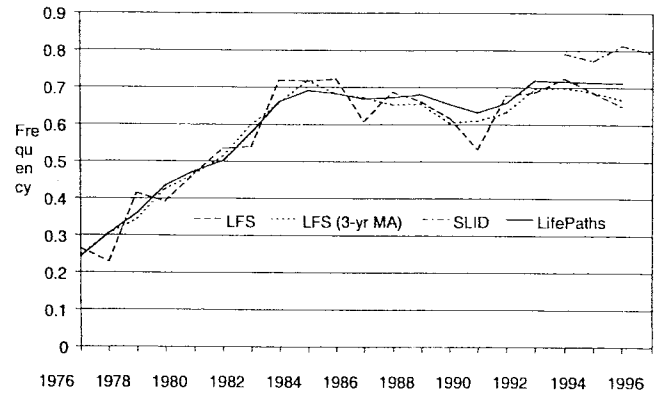
where, for spells beginning in period  $\tau$ ,  $d^{\tau}(t)$  is the number of individuals who fail in the interval  $(t-1, t]$  and  $r^{\tau}(t)$  is the number of individuals in view at duration  $t$ , including those censored at time  $t$  (censoring can only occur at the end of intervals). Numbers of individuals were calculated from sample counts by applying the LFS weight from the month that a new mother returns to work. Equation (3.7) together with equation (5.5) yields

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

## 6. RESULTS AND CONCLUSIONS

### 6.1 Results for the 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



**Figure 1:** Frequency of Choosing a Maternity Leave

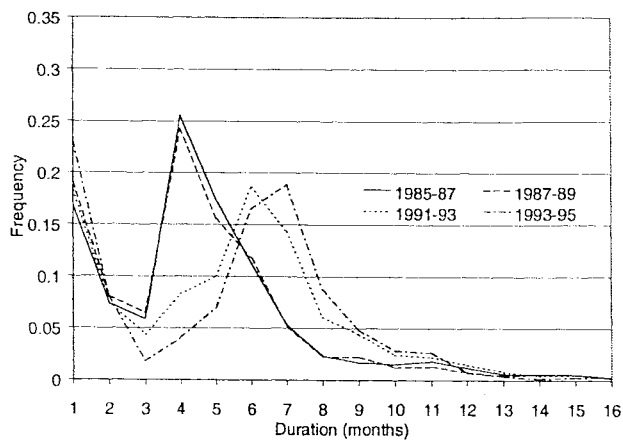
the LFS data do not indicate a strong trend after 1992. The model is then incorporated in LifePaths and a simulation is run. For each year from 1976 to 1995, Figure 1 shows both the frequency of choosing a leave in the LifePaths simulation, and the frequency estimated from the LFS. For the period 1993-1995, estimates from SLID are also presented.

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.

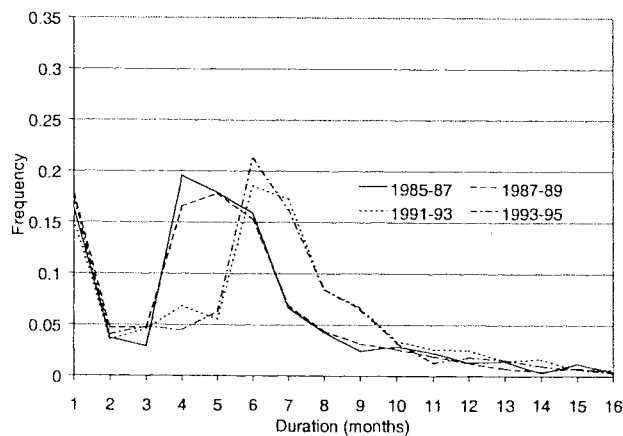
### 6.2 Results for the Survival Model

In the case of the semiparametric survival model, benchmarking consists of adjusting all of the background hazard parameters according to (5.7) 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,



**Figure 2:** LifePaths: Distribution of Leave Durations for 1985-1989 and 1991-1995



**Figure 3:** LFS Data: Distribution of Maternity Leave Durations for 1985-1989 and 1991-1995

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. To balance periods before and after 1990 using available data, in Figures 2 and 3 we use the overlapping periods 1985-1987, 1987-1989, 1991-1993, and 1993-1995.

### 6.3 Conclusions

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.

### ACKNOWLEDGEMENTS

The authors wish to express their thanks to Steve Gribble and the Socio-economic Modelling Group at Statistics Canada for useful comments, 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, and to Adrienne ten Cate for fruitful discussions.

### REFERENCES

- APPLEBY, J., D. BOOTHBY, 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.
- CHEN, E.J., and ROWE, G. (1999). Trend Correlation of Labour Market Earnings in Canada: 1982 to 1995. *Statistical Society of Canada 1999 Proceedings of the Survey Methods Section*, 173-179.
- CHEN, E.J., and ODERKIRK, J. (1997). Varied Pathways: The Undergraduate Experience in Ontario, Feature article. *Education Quarterly Review*, Statistics Canada, Vol. 4, No. 3, 47-62.
- 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.
- 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, Vol. 5, No. 4, 10-35.
- 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.
- WOLFSON, M.C., and ROWE, G. (1998). LifePaths – Toward an Integrated Microanalytic Framework for Socio-Economic Statistics. 26<sup>th</sup> 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, in M. Corak (Ed), *Government Finances and Generational Equity*, Statistics Canada, 107-127.