Proceedings of Statistics Canada Symposium 2002
Modelling Survey Data for Social and Economic Research

# THE DYNAMICS OF LOW INCOME IN FOUR COUNTRIES

Dennis Batten [1], Miles Corak [2], and Wen-Hao Chen [3]

## ABSTRACT

We model the dynamics of exiting low-income using discrete time hazard models while controlling for both observed and unobserved heterogeneity. The emphasis of this paper is on the methodology behind the econometric hazard models and estimation procedures we are using to carry out our research. An application to demonstrate the work, using longitudinal data from the United Kingdom, Germany, United States and Canada, is provided. The analysis relies on data from the Cross-National Equivalent files of the BHPS (for the UK), the GSOEP (for Germany), the PSID (for the US) and the SLID (for Canada). Longitudinal administrative data drawn from tax files from Canada are also used.

KEY WORDS: Longitudinal Survey Data, Unobserved Heterogeneity, Mixture of Distributions, Multiple Spells,

## 1. INTRODUCTION

The importance of viewing the labour market from a dynamic perspective is now accepted wisdom in both policy and academic circles. The availability of longitudinal data and the appropriate analysis of them have helped to create a clearer picture of North American and European labour markets as it relates directly to policy concerns. The most obvious example is the fuller understanding that has developed about the nature of low income. Bane and Ellwood (1986) using longitudinal data from the US are often cited as one example of a particularly cogent portrait of the low-income population and the dynamic processes that determine entry into and exit from poverty. The gradual availability of similar data in other countries has spawned a wide literature on this topic, with the most recent examples being Bradbury, Jenkins and Micklewright (2001), Jenkins (2000) and Stevens (1999).

The objective of our research is to examine the dynamics of low income in a comparative way by focusing on North America and two European countries. We use longitudinal data from Canada, the United States, Germany and the United Kingdom, and focus on developments during the 1990s.

The emphasis of this paper is on the methodology behind the econometric hazard models and estimation procedures we are using to carry out our research. In section 2, we begin with a brief overview of the data sources and bring attention to some issues that need consideration when conducting a comparative analysis of this nature. Next, in section 3 we provide a detailed discussion on how we model transition probabilities including how we control for unobserved heterogeneity, the estimation procedure we use with complex survey data, a small example to illustrate how it all works and how to estimate variances. Finally, we conclude the paper in section 4 by providing a brief summary of what we have accomplished thus far and outline how we intend to direct our future research.

## 2. DATA SOURCES AND ISSUES

The data come from the Cross-National Equivalent Files (CNEF). The CNEF brings together multiple waves of longitudinal data from Canada, United States, Great Britain, and Germany. Variables across the surveys have been defined in a similar manner in order to encourage cross-national research (Burkhauser *et al.* 2000). The data extracted from CNEF, for the use of this study, originate from the Canadian Survey of Labour and Income Dynamic

[1]Statistics Canada, Ottawa, Canada, K1Y 0T6 (dennis.batten@statcan.ca)
[2]Statistics Canada, Ottawa, Canada, K1Y 0T6 (miles.corak@statcan.ca)
[3]Statistics Canada, Ottawa, Canada, K1Y 0T6 (wen-hao.chen@statcan.ca)

(SLID, 1993-1998), the United States Panel Survey of Income Dynamics (PSID, 1990-1996), the British Household Panel Survey (BHPS, 1991-1999), and the German Socio-Economic Panel (GSOEP, 1992-1999). An advantage of CNEF data is that a set of estimates of annual income variables are provided that are not immediately available on the original data sets that have been derived from the original variables from all countries. It includes pre- and post-government household income, estimates of annual labour income, assets, imputed rent, private and public transfers, and taxes paid at household level. The availability of information on each household income source allows a full picture to be developed of the relative roles of market, family and state in determining income level.

A study which involves multi-national comparisons, incorporating data from different surveys all of which use different sampling methods, requires crucial and difficult decisions concerning various definitions and concepts. Which populations are we comparing? Considering, the limitation of each survey, the dynamic nature of the longitudinal populations, the non-response rates, etc. We need to define the unit of analysis and the unit's income. Do we use before tax or after tax income? For the CNEF we do not use the post-government variable due to the inconsistent definition across nations. How to define low income? There is no one definition of low income in the four countries, the USA being the only one with an official "poverty" line. How should we assign an "equivalent income" to each unit member? Can we use total unit income as a function of the unit size or as function of the number of adults and number of children by assigning different weights to the adults and children? For the most part, there are arguably many different answers, which are acceptable for each of these questions. Here is what we did.

Each individual living in the same household, whether or not they are related to each other by blood or marriage, are assigned an income value that is equivalent for each household member. Equivalent Income is calculated as a function of total household income adjusted to constant 1997 dollars and then divided by the square root of the size of the household, where household incomes are measured as market income prior to government transfers. Market income is defined in the same way as total income without taking into account public transfers, social security pension and taxes. This can be regarded as an estimate of potential disposable income for each household member under the assumption of equal sharing. Furthermore, annual income rather than income at the time of interview is used. An individual is defined as being in low income, in a particular year, if the household income falls below one half of the national median in that year.

The samples are meant to be representative of all individuals in the population including children and non-working people. However, in order to fit into our transition models, only individuals with spells that had a fresh start of low-income (non-left censor) during the period of observation of the study are included. Once an individual is in low income, we follow the low income spell until it ends or it is right-censored at the end of the survey period. Multiple spells of low income by the same individual are included in the study and assumed to be independent of each other, therefore our most basic unit of study is a low income spell and we are interested in observing, after an individual begins a low income spell, the end of that particular spell of low income. Those individuals who had discontinuity at any point during the period of study are dropped (for data missing for one or more years).

We also included, from Canada, the Longitudinal Administrative Data (LAD), enabling us to compare two data sources within Canada. LAD is created from various administrative data files including tax filings and the child tax benefit file. LAD presently spans 19 years (1982-2000) and contains information on individuals and their families with additional years of information added for the selected individual as they become available. Individuals contained on the LAD file are selected using a Bernoulli sampling scheme. This sampling scheme makes it easy to select individuals, and account for population changes such as births, deaths, immigrants, emigrants, etc (Small Area and Administrative Data Division, 2002).

## 3. MODELING TRANSITION PROBABILITIES

At time $t = 0$ we observe the start of low income spell $i$. For any other time $t$, $t > 0$ let $y_{it}$, $i = 1,...,n$ denote the outcome for spell $i$ where $y_{it} = 1$ if the $i^{th}$ spell exits low income at time $t$ and $y_{it} = 0$ otherwise. Assume that the conditional probability of success at time $t$ is modeled as

$$log\,it(\, p_{it}\,) = \mathbf{x}_{it}{}^{T}\boldsymbol{\beta}\,, \tag{3.1}$$

where $logit(\,a\,) = log(\,a\,/(\,1-a\,))$, $\mathbf{x}_{it}$ is a $p \times 1$ vector of independent variables, and possibly interactions of these independent variables, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters. We are primarily interested in making inferences about $\boldsymbol{\beta}$. Let us assume first that the $i^{th}$ spell is observed $T_i$ times until a success or a right censoring occur. The likelihood function for individual $i$ under the assumption of independence from year to year is

$$L_i = \prod_{t}^{T_i} f_{it}\,, \tag{3.2}$$

where $f_{it} = (1 - p_{it})^{1-\mathbf{y}_{it}}(p_{it})^{\mathbf{y}_{it}}$ in case of Bernoulli distribution probability distribution for a given year. $\boldsymbol{\beta}$ can be estimates by using standard statistical software packages which generally use the Newton-Raphson or Fisher scoring method. The estimator $\hat{\boldsymbol{\beta}}$ is the solution of an estimating equation of the form

$$U(\boldsymbol{\beta}) = \sum_i u_i(\boldsymbol{\beta})\,, \tag{3.3}$$

where $u_i(\boldsymbol{\beta}) = \sum_t (y_{it} - p_{it})\mathbf{x}_{it}$ in the case of a logit model for $p_{it}$ as in (3.1).

### 3.1 Unobserved Heterogeneity

One major goal of our study is to identify independent variables that predict the probability of success which, in the context of our problem, is observing the end of a low income spell. The vector of independent variables may include the duration of the spell along with socio-economic and demographic characteristics. However, the probability of success may be influenced by a latent class variable. An unobservable discrete variable, say $Z$, indicates the latent class of the $i^{th}$ spell. The variable is assumed to take $G$ distinct values, each of which corresponds to a distinct probability of success. Therefore, the probability of success of the $i^{th}$ spell at time $t$ depends on observed as well as unobserved characteristics. For example, in the case of $G = 2$ groups the unobserved component is introduced into the model in the same manner as the observed variables, that is,

$$h(\,\mu_{igt}\,) = \lambda_g \mathbf{x}_{it} + \lambda_g \tag{3.4}$$

where $\mu_{igt} = E(\,y_{it} \mid z_i = g, \mathbf{x}_{it}, \boldsymbol{\beta}\,)$ is the conditional mean with $\boldsymbol{\beta}_g^T = (\,\boldsymbol{\lambda}_g^T, \boldsymbol{\alpha}^T\,)$ the effect of the observed and unobserved variables for the g$^{th}$ group and $h(.)$ is the link function. One objective is to identify the latent class effect, which is statistically equivalent to determining whether $\boldsymbol{\alpha}_g$ differs across latent groups.

## 3.2 Estimation

A Mixture of distributions framework, which covers the latent class problem, is used when the data can be viewed as arising from two or more populations mixed in varying proportions. Each observed $y_i$ is drawn from a super-population $P$ which is a mixture of a finite number of $G$ populations $P_1,...,P_G$ in some proportions $\pi_1,...,\pi_G$, with

$\sum_g \pi_g = 1$ and $\pi_g \geq 0$, $g = 1,...,G$.

Given the data, and a known form of the distributions, we wish to estimate the model parameters and the mixing distribution. A mixture of distributions can be handled by using the Expectation Maximization (EM) algorithm (Dempster, Laird, and Rubin 1977). This procedure can be implemented using standard statistical software like SAS and writing an iterative code using a SAS defined function at the maximization step. For example, if with a logit link function in SAS, then the *proc logistic* function would be used for the maximization step. This can be achieved by essentially creating $G$ copies of the data and then using the conditional probabilities as weights under the census case. The calculation of the weights and the estimation of the parameters are repeated until convergence is achieved. Define the latent group membership indicator variables for spell $i$ as $z_{ig} = 1$ if $i \in P_g$ and $z_{ig} = 0$ if $i \notin P_g$. The EM algorithm is applied to the mixture of distributions by treating the variable $z_{ig}$ as missing data. The likelihood for the complete data for spell $i$ is given by

$$L_i = \prod_{g=1}^{G} (\pi_g f_{ig})^{z_{ig}} \tag{3.5}$$

where $f_{ig} = f_{ig}(y_i^1,...,y_i^T \mid \beta, g)$. Using some initial value for $\boldsymbol{\varphi}' = (\boldsymbol{\beta}', \boldsymbol{\pi}')$, say $\boldsymbol{\varphi}^{(m)}$, the E step requires the calculation of the pseudo complete log-likelihood based on the incomplete data

$$Q(\boldsymbol{\varphi}, \boldsymbol{\varphi}^{(m)}) = E(l_i \mid y_i, \boldsymbol{\varphi}^{(m)})$$

where $l_i = \log L_i$. That is, each indicator variable $\boldsymbol{z_{ig}}$ is replaced by its conditional expectation, $\tau_{ig}$, where.

$$\tau_{ig}^{(m)} = \pi_g^{(m)} f_{ig}(\beta^{(m)}) / \sum_j \pi_j^{(m)} f_{ij}(\boldsymbol{\beta}_j^{(m)}) .$$

The intent of the M step is to choose the value of $\boldsymbol{\varphi}$, say $\boldsymbol{\varphi}^{(m+1)}$, that maximizes $Q(\boldsymbol{\varphi}, \boldsymbol{\varphi}^{(m)})$ subject to $\sum_g \pi_g = 1$. To accomplish the maximization step for each EM iteration, it suffices to solve

$$\partial l_i / \partial \boldsymbol{\varphi} = \partial \log f(\mathbf{y}_i | \mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\pi}) / \partial \boldsymbol{\varphi} , \tag{3.6}$$

giving rise to the following two estimating equations

$$\partial l_i / \partial \boldsymbol{\beta} = \sum_g \tau_{ig}^{(m)}(s) f(\mathbf{y}_i | \ z = g_i, \boldsymbol{\beta}) \Big/ \partial \boldsymbol{\beta} \tag{3.7}$$

$$\partial l_i / \partial \boldsymbol{\pi} = \sum_g \tau_{ig}^{(m)}(s) f(\mathbf{y}_i | \ z = g_i, \boldsymbol{\pi}) \Big/ \partial \boldsymbol{\pi} . \tag{3.8}$$

Both estimating equations (3.7) and (3.8) are of the form of (3.3). If the data are from a survey with a complex design, a design consistent estimator of (3.3) is given by

$$\hat{U}(\boldsymbol{\beta}) = \sum_i w_i(s) u_i(\boldsymbol{\beta}) , \tag{3.9}$$

where $w_i(s)$ is the survey weight for the $i^{th}$ spell in sample $s$.

## 3.3 Characterization of Low Income Dynamics

For the purpose of this paper, we will illustrate the method by modeling the probability of a low income spell ending using only a single explanatory variable, namely, the duration of the low income spell before ending. Moreover, a common time period (1993-1998) will be chosen over all data sets to ensure simplicity and comparability of the countries. Assume the following model for exiting "low" income:

$$logit(p_{it(g)}) = \alpha_g + \lambda_{g2} d_{2i}(t) + \lambda_{g3} d_{3i}(t) + \lambda_{g4} d_{4i}(t)$$

under a mixture of $G = 2$ distributions. Where $d_{ji}(t) = 1$ if $t = j$ years and $0$ otherwise and $\lambda_{gt}$ is the associated regression effect for the $i^{th}$ unit to exit low income after $t$ years under the $g^{th}$ group. The parameter estimates are given in Table (3.1) for the five data sets studied. Table (3.2) displays the probability a low income spell will end after particular length of time (1,…,4+).

The characteristics of the two groups depicted in Tables (3.1) and (3.2) appear to be similar for each data set. Low income spells belonging to group one appear to have an extremely low probability of exiting at the beginning of the spell, but the probability of exiting increases the longer the spell lasts. This group appears to characterize approximately 29% of the low income spells in Great Britain, 36% and 37% of the low income spells from the LAD and SLID data sets respectively in Canada, and 37% of the low income spells in Germany. The second group, which characterizes the rest of the low income spells in each country, tells an opposite story. Low income spells belonging to this group appear to have an easier time exiting low income in the early stages of the spell. However, the longer a low income spell lasts, characterized by group two, the less likely that low income spell will end.

Table 3.1: Parameter estimates of probability of exiting low income 1993-1998

| ***Model Parameters*** | ***Canada*** | | | | ***Germany*** | | ***Great Britain*** | | **** United States*** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | LAD | | SLID | | | | | | | |
| G | ***1*** | ***2*** | ***1*** | ***2*** | ***1*** | ***2*** | ***1*** | ***2*** | ***1*** | ***2*** |
| ***∀*** | ***-2.81*** | ***-0.02*** | ***-3.15*** | ***0.24*** | ***-3.18*** | ***-0.02*** | ***-3.78*** | ***-0.09*** | ***0.46*** | ***-0.86*** |
| $8_2$ | ***0.81*** | ***-0.78*** | ***0.30*** | ***-0.06*** | ***0.75*** | ***-0.51*** | ***0.31*** | ***-0.32*** | ***-1.98*** | ***-0.45*** |
| $8_3$ | ***1.56*** | ***-2.46*** | ***1.75*** | ***1.42*** | ***1.27*** | ***-2.48*** | ***1.40*** | ***-1.70*** | | |
| $8_{4+}$ | ***1.77*** | ***-4.95*** | ***1.85*** | ***-4.76*** | ***1.92*** | ***-5.52*** | ***3.03*** | ***-4.16*** | | |
| ***B*** | ***0.36*** | ***0.64*** | ***0.37*** | ***0.63*** | ***0.37*** | ***0.63*** | ***0.29*** | ***0.71*** | ***0.37*** | ***0.63*** |

** Note: an error reading the CNEF data file for the 1997 United States data inhibited analysis for that country past the $2^{nd}$* year

Table 3.2: Probability to exit low income after t years 1993-1998

| ***Duration*** | ***Canada*** | | | | ***Germany*** | | ***Great Britain*** | | ***United States*** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | LAD | | SLID | | | | | | | |
| <u>G</u> | ***1*** | ***2*** | ***1*** | ***2*** | ***1*** | ***2*** | ***1*** | ***2*** | ***1*** | ***2*** |
| ***1*** | ***0.057*** | ***0.495*** | ***0.041*** | ***0.560*** | ***0.040*** | ***0.495*** | ***0.022*** | ***0.476*** | ***0.613*** | ***0.297*** |
| ***2*** | ***0.119*** | ***0.310*** | ***0.055*** | ***0.545*** | ***0.081*** | ***0.371*** | ***0.030*** | ***0.399*** | ***0.179*** | ***0.212*** |
| ***3*** | ***0.223*** | ***0.077*** | ***0.198*** | ***0.840*** | ***0.129*** | ***0.076*** | ***0.085*** | ***0.143*** | | |
| ***4+*** | ***0.261*** | ***0.007*** | ***0.214*** | ***0.012*** | ***0.221*** | ***0.004*** | ***0.321*** | ***0.014*** | | |
| ***B*** | ***0.36*** | ***0.64*** | ***0.37*** | ***0.63*** | ***0.37*** | ***0.63*** | ***0.29*** | ***0.71*** | ***0.37*** | ***0.63*** |

The observed data sets appear to display similar patterns when only the duration variable is used to explain the probability for a low income spell ending. However, when additional explanatory variables (sex, age, family structure, etc) are considered the fitted models tell different stories across the four countries.

## 3.4 Variance Estimation

To make appropriate and adequate comparisons between and within countries as well as to investigate the significance of our estimates we require good variance estimation. In this section we will outline examples of different comparisons of interest and propose a variance estimator.

To test the significance of our underlying parameters we need to conduct a hypothesis test that will reveal whether our parameters differ from zero. For example, to test if a parameter in Canada is significant we conduct the following test $H_0 : \boldsymbol{\varphi}^{(1)}_{\mathbf{CAN}} = 0$. If our interest is to investigate if significant differences exist between countries or, in the case of Canada, between populations represented by two different samples, a different hypothesis will need to be tested. For example, to test if the effects in Canada and the United States are different we test $H_0 : \boldsymbol{\varphi}^{(1)}_{\mathbf{CAN}} = \boldsymbol{\varphi}^{(1)}_{\mathbf{USA}}$, and we test $H_0 : \boldsymbol{\varphi}^{(1)}_{\mathbf{LAD}} = \boldsymbol{\varphi}^{(1)}_{\mathbf{SLID}}$ to investigate any differences between the two available data sets in Canada. Where $\boldsymbol{\varphi}^{(1)} \subseteq \boldsymbol{\varphi}$. To conduct such statistical procedures we need an estimate of the variance of our estimated coefficient vector. The variance is

$$Var(\hat{\boldsymbol{\varphi}}) = E_M Var_S(\hat{\boldsymbol{\varphi}}) + Var_M E_S(\hat{\boldsymbol{\varphi}})$$

where $E_M$ and $Var_M$ are the model based expectation and variance respectively, $E_S$ and $Var_S$ are the expectation and variance respectively under the sample design. An estimate of the variance is given by

$$est(Var(\hat{\boldsymbol{\varphi}})) = v_S(\hat{\boldsymbol{\varphi}}) + est(Var_M E_S(\hat{\boldsymbol{\varphi}})) .$$

When $\boldsymbol{n/N}$ is small $\boldsymbol{est(Var_M E_S(\hat{\varphi}))}$ is small. Therefore, a good approximation for $Var(\hat{\boldsymbol{\varphi}})$ is

$$est(Var(\hat{\boldsymbol{\varphi}})) \approx v_S(\hat{\boldsymbol{\varphi}})$$

For calculating $v_S(\hat{\boldsymbol{\varphi}})$ we assume that the survey design can be approximated by with replacement selection of primary sampling unites (PSU) from each stratum and then use a replication method of variance estimation such as the Jackknife or Bootstrap method.

## 3.5 Future Extensions

When studying two or more types of events in the same model, the formulation in equation (3.3) can still be used. For example, when considering simultaneously the conditional probability of exiting low income $\lambda_{it}$ and the probability of re-entering low income $q_{it}$, all that is needed is to redefine the vector of independent variables and the vector of parameters. Let $\tilde{\mathbf{x}}_{it}$ be the new vector of independent variables of order $(2p \times 1)$ with $\tilde{\mathbf{x}}_{it}{}^T = (\mathbf{x}_{it}{}^T, \mathbf{0}^T)$ for exiting low income and $\tilde{\mathbf{x}}_{it}{}^T = (\mathbf{0}^T, \mathbf{x}_{it}{}^T)$ for re-entering low income, where $\mathbf{0}$ is a $(p \times 1)$ vector of 0's. Let $\boldsymbol{\beta}$ be the new vector of parameters of order $(2p \times 1)$ where $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1{}^T, \boldsymbol{\beta}_2{}^T)$, $\boldsymbol{\beta}_1$ is the vector of parameters associated with $\lambda_{it}$ and $\boldsymbol{\beta}_2$ is the vector of parameters associated to $q_{it}$. Each probability can be written in terms of $\tilde{\mathbf{x}}_{it}$ and $\boldsymbol{\beta}$; i.e., $logit(\lambda_{it}) = \mathbf{x}_{it}\boldsymbol{\beta}_1 + \mathbf{0}\boldsymbol{\beta}_2 = p_{it}$ and $logit(\mathbf{q}_{it}) = \mathbf{0}\boldsymbol{\beta}_1 + \mathbf{x}_{it}\boldsymbol{\beta}_2 = p_{it}$.

# 4. CONCLUSION

This paper demonstrated how to model transitional probabilities while accounting for unobserved heterogeneity through a mixture model. The procedure was applied to a study that investigates the income dynamics of four countries Canada, Britain, Germany and the United States. The Expectation Maximization (EM) algorithm was used to estimate the unknown parameters for the econometric hazard functions. This algorithm can be implemented using any standard statistical software package.

Although models were fitted for all data sets, we expressed the important requirement for a variance estimator before appropriate and adequate inferences can be made within and between countries. Moreover, in section 3.1 we proposed a variance estimator that will enable us to make necessary inferences. Although the proposed Jackknife estimator is widely used and accepted in the literature more work is needed before we apply it to our specific situation.

# REFERENCES

Bane, M. J. and Ellwood D. (1986). “Slipping Into and Out of Poverty: The Dynamics of Spells.” *Journal of Human Resources.* Vol. 21 No. 1, pp. 1-23.

Bradbury, B., Jenkins S. P. and Micklewright J. (2001). *The Dynamics of Child Poverty in Industrialized Countries.* Cambridge University Press.

Burkhauser, R. V., Butrica B. A., Daly M. C., and Lillard D. R. (2000). "The Cross-National Equivalent File: A product of cross-national research." in Social Insurance in a Dynamic Society, Irene Becker, Notburga Ott, and Gabriele Rolf (Eds.)

D’Ambrosio, C., Papadopoulos F. and Tsakloglou P. (2002). “Social Exclusion in EU Member States: A Comparison of two Alternative Approaches.” Paper presented to the XVI Annual Conference of the European Society for Population Economics, Bilbao.38.

Dempster, A. P., Laird N. M. and Rubin D. B. (1977). “Maximum likelihood Estimation from Incomplete Data via the EM Algorithm (with Discussion).” *Journal of the Royal Statistical Society.* Series B. Vol. 39, pp 1-38.

Jenkins, S. (2000). “Modelling Household Income Dynamics.” *Journal of population Economics.* Vol. 73, pp. 13-22.

Small Area and Administrative Data Division (2002), Longitudinal Administrative Databank, Dictionary 2000, Catalogue no. 12-585-XIE, Statistics Canada.

Stevens, A. H. (1999). “Climbing Out of Poverty, falling Back In: Measuring the Persistence of Poverty Over Multiple Spells.” *Journal of Human Resources.* Vol. 34 No. 3, pp. 557-588.

Stewart, K. (2002). “Measuring Well-Being and Exclusion in Europe’s Regions.” Paper presented to the XVI Annual Conference of the European Society for Population Economics, Bilbao.

OECD (2001). *Employment Outlook.* Paris: OECD.