

Estimating Labour Market Transitions and Continuations using Repeated Cross Sectional Data

Pierre Brochu
University of Ottawa*

May 2009

Abstract

This paper proposes a population cohort approach for estimating labour market continuations (or transitions) using repeated cross sectional data. I show that the continuation probability can be written as a ratio of two unconditional means that do not condition on past labour market status. As such, I can construct a consistent set of standard errors that account for the full variability of cross sectional data. Using Current population Survey data, I show that existing methods tend to systematically underestimate the true standard errors which can lead the researcher to incorrectly conclude that job stability had decreased.

JEL Classification: C41, J64

Key Words: repeated cross section data, duration analysis, employment, job stability

*Department of Economics, University of Ottawa, 55 Laurier Avenue East, Desmarais Building, Room 10105, Ottawa (Ontario) K1N 6N5, Canada. Phone: (613) 562-5800 ext. 1424. Fax: (613) 562-5999. Email: pbrochu@uottawa.ca

1 Introduction

There is a long tradition of exploring labour market transitions in economics. Although the unemployment-employment transition has been the most frequently explored, other transitions or continuations have also been examined, such as the transition out of the labour force (e.g. Jones and Riddell (1999)) and the continuation of a job (job stability, e.g. Brochu (2008); Heisz (2005); Neumark, Polsky, and Hansen (1999)).

While using panel data to estimate these labour market transitions is generally the preferred approach, there are circumstances where that approach is problematic. For example, limited historical coverage (Canadian panels) makes it impossible to differentiate between cyclical and secular changes in job stability.¹ With the absence of this differentiation, one cannot address the real question of interest in the job stability literature: how and why has job stability changed?² In such instances, repeated cross sectional data sets offer a valid alternative.

In this paper, I propose a population cohort approach for estimating the continuation (or transition) probability when using repeated cross-section data. The proposed non-parametric approach is empirically tractable, and its identifying assumptions are relatively mild and easy to interpret.³ The approach takes advantage of the fact that repeated cross sectional data sets, like the Current Population Survey (U.S.) and the Labour Force Survey (Canada), are representative of the country's population.

The existing empirical labour literature that examines continuation probabilities using non-parametric methods have applied (or approximated) panel tools when faced with cross sectional data (e.g. Neumark, Polsky, and Hansen (1999)). Approximating panel methods makes the researcher focus exclusively on a specific group of individuals; for the job stability literature which examines the employment-continuation, its employed workers. Yet, their cross sectional samples of employed workers are drawn from a population that changes over time - some workers will lose their jobs while others will find employment. As such, researchers are faced with the difficult task of estimating *one* transition probability using samples drawn from *two* different populations.

The key insight of this paper is that by using a population cohort approach, one can write the labour market continuation (or transition) as a ratio of two means; means that do not condition on past labour market status. Based on this result, I propose a cross sectional estimator for the continuation probability, and another for its standard errors.

Using the proposed population cohort framework, I also re-examine methods used in the empirical labour literature (e.g. job stability literature). Through this lens, I clearly identify the full

¹See Brochu (2006) for a detailed discussion of panel data limitations when exploring for changes in job stability using North American data.

²This problem is not confined to North America data. Researchers (e.g. Güell and Hu (2006)) that have studied the Spanish labour market faced similar difficulties. Panel data only became available after significant labour reforms, i.e. the introduction of fixed term contracts.

³See Moffitt (1993) and Verbeek and Vella (2004) for a discussion of other approaches to estimating dynamic models using repeated cross sections.

set of underlying assumptions of the traditional non-parametric estimator and provide a consistent estimator for its standard errors—both of which fill gaps in the literature.

I show with Current Population Survey data that the existing approaches tend to underestimate the true standard errors because they do not capture the full variability of cross sectional data. I also show that this matters at the inference stage; one can (incorrectly) conclude that job stability has changed—when in fact it has not.

Finally, a researcher may prefer the use of repeated cross sections even when panel data is available. Repeated cross sections like the Current Population Survey (CPS) and the Canadian Labour Force Survey (LFS) have much larger sample sizes that allow for a more detailed analysis (e.g. Baker (1992)); one can focus on more narrowly defined groups without having to worry as much about small samples. Depending on the question (and group) of interest, panel attrition may become a significant problem. One can, for example, expect younger and more mobile workers to have higher attrition rates. For these groups, a self-selection bias may arise. The proposed estimator, however, does not face this difficulty—it is designed for cases where the same individual cannot be followed over time.

The remaining sections of this paper are divided as follows: Section 2 provides a discussion of standard cross sectional approaches; Section 3 proposes an alternative framework that is suitable to cross sectional analysis; Section 4 re-interprets the existing methods using the proposed framework; Section 5 provides some empirical examples; and finally, Section 6 provides some final remarks.

2 Existing Approach

Assume the researcher is interested in the retention rate of an “at-risk group”, say, individuals with time-invariant characteristics c who have been employed for s periods at time t .⁴ Following the existing cross sectional literature (e.g. Neumark, Polsky, and Hansen (1999); Heisz (2005)), one can present the retention rate simply as the fraction of “at-risk” individuals in the *population* that remains with the same employer in the next period

$$R_t^{s,c} = \frac{N_{t+1}^{s+1,c}}{N_t^{s,c}} \quad (1)$$

where $N_t^{s,c}$ is the number of people in the population that have time-invariant characteristics c who have been unemployed for s periods at time t .⁵

Although one cannot follow individuals over time with repeated cross sectional data, researchers (e.g. Baker (1992); Neumark, Polsky, and Hansen (1999)) take advantage of the fact that base weights of representative cross sections, like the Labour Force Survey (Canada) and the Current Population Survey (U.S.), sum up to their respective populations. The cross sectional estimator

⁴The framework can easily be generalized to explore any transition or continuation probability.

⁵I also follow the notation of the literature by treating N as a count measure.

takes the form

$$\hat{Q}_t^{s,c} = \frac{\tilde{n}_{t+1}^{s+1,c}}{\tilde{n}_t^{s,c}} \quad (2)$$

where $\tilde{n}_t^{s,c}$ is the sum of the base weights of all individuals with characteristics c who have been employed s periods with the same employer as of period t . By using weights as counts, the denominator (numerator) of Equation (2) directly estimates the denominator (numerator) of Equation (1).

Directly estimating population counts is very intuitive, and it is reasonable to think that the accuracy of the estimator will improve with larger samples; yet, this cannot be proven in any statistical sense. In addition, one cannot lay bare all underlying identifying assumptions without such a proof. A clear understanding of the consistency requirements is critical considering that an important selling point of cross sectional data sets like the CPS and LFS have been their large sample sizes.

Most importantly, the lack of precision carries over to the inference stage. Given the functional form of the estimator, there is no standard way to construct standard errors. In Section 5, I use CPS data to show that the existing approaches tend to systematically underestimate the true standard errors because they do not capture the full variability of cross sectional data.

3 Proposed Approach

In this section I propose an alternative representation of the retention rate (one that differs from Equation (1)). Based on proposed representation, I suggest an estimator for the retention rate (and for its standard errors) that is well-suited to cross sectional data.

I start by assuming a population cohort.⁶ Having a population cohort simply means that there is more than one period of information for each individual in the *population*. It does not impose any restrictions on the type of sample. One could have a panel dataset where each *sampled* individual is followed over time, or repeated cross sections where each sample (cross section) is drawn from the same population, but at different moments in time.⁷

Let X_{it} be a vector of time invariant characteristics of individual i in period t . Further, let TEN_{it} represent the length of tenure, i.e. the number of periods the worker has been employed with the same employer as of period t . The retention rate for a randomly chosen individual with

⁶Other researchers (e.g. Deaton (1985), Moffitt (1993)) who have estimated dynamic models using repeated cross section have also relied on this assumption.

⁷Said differently, I assume that the cross sections are drawn from the same pool of individuals, with each cross section consisting of its own random draw.

characteristics c who has been employed for s periods at time t , $R_t^{s,c}$, is formally expressed as⁸

$$R_t^{s,c} = Prob(TEN_{it+1} = s + 1 | TEN_{it} = s, X_{it} = c) \quad (3)$$

Given its conditional structure, Equation (3) is a good starting point for panel data. One can condition on an individual working in the first period of a panel, and therefore estimate the sample analog of Equation (3). With repeated cross sections, however, this is not possible since one cannot follow individuals over time.

I propose an alternative representation that is more conducive to cross sectional analysis. One can rewrite Equation (3) as⁹

$$R_t^{s,c} = \frac{E(1[TEN_{it+1} = s + 1, X_{it+1} = c])}{E(1[TEN_{it} = s, X_{it} = c])} \quad (4)$$

where $1[\cdot]$ is an indicator function that equals 1 if the conditions inside the bracket hold, and zero otherwise. The proof can be found in Appendix A.1. One can estimate this retention rate using two repeated cross sections—by replacing the population means with their respective sample analogs, i.e.

$$\hat{R}_t^{s,c} = \frac{\sum_{i=1}^{n_{t+1}} 1[TEN_{it+1} = s + 1, X_{it+1} = c] / n_{t+1}}{\sum_{i=1}^{n_t} 1[TEN_{it} = s, X_{it} = c] / n_t} \quad (5)$$

where n_t is the sample size in year t .

Equation (4) is a key insight of this paper. It is conducive to cross sectional analysis because the numerator does not condition on period t events. This holds true because an individual who has been with the same employer for $s + 1$ periods as of time $t + 1$, had to have been with the same employer in the previous period (and have one less period of tenure).¹⁰

Conditioning on only time-invariant characteristics is a sufficient but not a necessary condition to be able to present the retention rate as a ratio of two means. One needs to be able to follow the at-risk group over time.¹¹ More precisely, one needs to be able to infer - from a period $t + 1$ cross section - whether an individual who remains with the same employer as of period $t + 1$, would have been part of the at risk group in period t . As such, one can estimate a broad range of retention rates. One can not only condition on gender, race, education, but also on age, industry and occupation. I elaborate on the latter three categories below.

Assuming age invariance for a 1-year retention rate, for example, would be untenable.¹² For-

⁸This is simply another way of writing Equation (1).

⁹For ease of presentation, I assume that the $[t, t + 1]$ time span is sufficiently short as to treat age as time-invariant. I later show how to address ageing when t and $t + 1$ are sufficiently far apart.

¹⁰A similar argument would hold true if one were looking at other labour market continuation/transition probabilities.

¹¹For $R_t^{s,c}$, the at-risk group consists of individuals with characteristics c who have been with the same employer for s periods as of time t .

¹²The Canadian job stability literature has typically focussed on 1-year rates (e.g. Heisz (2005); Brochu (2008)).

tunately, the above method can easily deal with ageing. If one can identify the age of a worker in year $t + 1$, we know that he was 1 year younger in year t . As such, one can consistently identify the at-risk group over time. Assume, for example, one is interested in the retention rate of 20-29 year old men that have been with the same employer 10 years as of 1988. The identifying assumption only requires that we to be able to identify - in the 1989 cross section - male workers that are 21 to 30 years old with 11 years of job tenure.

Industry status is also not time-invariant (from the researcher’s point of view). Industry affiliation is job related, and as such, one cannot typically identify (in the data) the industry to which the unemployed worker previously belonged. But as previously mentioned, one only needs to identify whether workers who continued with the same employer in period $t + 1$ would have been in the at-risk group in the previous period. To do so, one must assume that job tenure (i.e. the employer-employer relationship) ends if the individual switches industry. It is a similar story for occupation. A change in occupation classification must signal the end of the employer-employee relationship. This is a relatively mild assumption as long as the occupations are not too narrowly defined.

It should be emphasized that the population cohort assumption implies that the samples (and their underlying population) must include both workers and non-workers. Restricting the sample to only include employed workers (as is the case in the job stability literature) would violate the population cohort assumption. The pool of employed workers in the population does not remain constant over time—some individuals will lose their jobs, while others will find employment. The population cohort assumption also requires that one impose a different age restriction for the period $t + 1$ sample if the $[t, t + 1]$ interval is sufficiently large. The same ageing argument that affected how one estimated the retention rates also applies to the sample. Assume one is interested in 1-year retention rates of prime age (25-54) workers (and of its sub-groups).¹³ The period t sample will consist of all individuals 25 to 54 years of age. The period $t + 1$ sample, however, should consist of 26 to 55 year olds. This will ensure that the population cohort assumption is met—that we are drawing from the same underlying population.

A population cohort approach is possible with both the LFS and the CPS, because they are both representative of their respective population. The LFS and CPS are also carried out at frequent intervals - on a monthly basis. As a result, slippage (changes in population) due to immigration, emigration or deaths will be minimal if one carries out monthly retention rates. In Section 4, I show how the proposed estimator can be adjusted when the cross sections are further apart as is the case in the American literature.

Given the simple functional form of the proposed retention rate estimator, i.e. $\hat{R}_t^{s,c}$, one can easily generate consistent standard errors. In Appendix A.3 I show how to do so by first deriving

For data reasons, the American literature has focussed on 4-year (and 8-year) rates. The use of cross sections that are 4-year (or 8-year) apart introduced a further complication: it becomes unreasonable to assume a population cohort. I will address this issue in the next section.

¹³One could, for example, also be interested in the retention rate of male workers 20 to 29 years of age.

the asymptotic properties of $\hat{R}_t^{s,c}$. The standard error estimator is consistent because it accounts for the full variability of cross sectional data.

Finally, applying Equation (5) to survey data where the probability of being selected is not the same across observations is straightforward. One replaces the sample means with weighted ones.¹⁴ This reflects the traditional use of weights where the weights are only used to reflect the varying probability of selection.

4 Links Between Methods

In this section I re-examine the existing retention rate estimator within a population cohort framework. I show that this estimator can be expressed as a function of sample means. As such, I can identify the full set of underlying assumptions of this estimator, and provide a consistent estimator for its standard errors.

The American job stability literature (e.g. Swinnerton and Wial (1995); Neumark, Polsky, and Hansen (1999)) estimated 4-year retention rates. It would be untenable to assume that the two cross sections - drawn 4-years apart - come from the same underlying population; the American population will have changed due to deaths, emigration and immigration.¹⁵ Fortunately, the population-cohort framework can be extended to deal with such compositional changes.

Assuming that compositional changes break the tenure spell, one can write the retention rate as a function of two population means

$$R_t^{s,c} = \frac{adj_t E(1[TEN_{it+1} = s + 1, X_{it+1} = c])}{E(1[TEN_{it} = s, X_{it} = c])} \quad (6)$$

where adj_t is the population growth (or adjustment) factor. If, for example, the population size increased by 20%, the population growth factor would be 1.2. The arguments are similar to those made in Section 3 when I relaxed the time-invariant characteristics assumption. One needs to be able to identify in the the American *population* who continued with the same employer in year $t + 4$, and would have been in the at-risk group in year t . An intuitive proof is left to Appendix A.3.

A death easily meets the identifying assumption; changes due to immigration and emigration, however, require further attention. One requires that the migrant change employer upon arrival in his new country. This empirical strategy would be appropriate if job transfers (where workers stays with same employer) are not the driving force behind migration patterns.

The existing approach, i.e. Equation (2), is in fact an estimator of $R_t^{s,c}$ as presented above. This become apparent if one rewrites Equation (2) as

¹⁴Where the weights are normalized to sum up to 1 in each sample period.

¹⁵Births are not problematic because we focus on the working age population. Assume one was interested in 4-year retention rates for prime aged (25-54) workers. The year t sample would consist of individuals aged 25 to 54, while the year $t + 4$ sample would only look at at those 29 to 58 years of age.

$$\hat{Q}_t^{s,c} = \hat{adj}_t \left(\frac{\sum_{i=1}^{n_{t+1}} nw_{it+1} 1[TEN_{it+1} = s + 1, X_{it+1} = c]/n_{t+1}}{\sum_{i=1}^{n_t} nw_{it} 1[TEN_{it} = s, X_{it} = c]/n_t} \right) \quad (7)$$

where nw_{it} is the normalized base weight of individual i in year t ,¹⁶ and $\hat{adj}_t = \frac{\sum_{i=1}^{n_{t+1}} bw_{it+1}}{\sum_{i=1}^{n_t} bw_{it}}$ (with bw_{it} representing the base weight).¹⁷

Given that the sum of the base weights add up to the target population in both the CPS and Canadian LFS, \hat{adj}_t is an estimate of the population growth. The second term of Equation (7) is simply the weighted version of $\hat{R}_t^{s,c}$ (which was discussed at the end of Section 3).

By rewriting the existing estimator as a function of the proposed one, I can make two contributions. First, I can clearly identify the underlying assumptions of this estimator - namely that changes in population break the tenure spell. Second, I can easily construct consistent standard errors. They will be similar to those of $\hat{R}_t^{s,c}$, but with an adjustment made for the population change.¹⁸ In the next section, I use an empirical example to show that existing approaches to estimating standard errors are downward biased, i.e. that they do not fully account for the variability introduced by not being able to follow individuals over time. I also show that this can lead the researcher to incorrectly conclude that job stability has changed.

5 Empirical Examples

In this section, I empirically compare existing and proposed approaches using two large scale data sets: the Canadian LFS and the U.S. CPS. This section is divided into two parts. In the first part, I focus on the retention rate estimators, and rely on repeated cross sections from the LFS files to illustrate the differences. The second part focusses on the standard errors. Using the CPS, I show that the bias of existing approaches to estimating standard errors can lead the researcher to falsely reject the null hypothesis of no change (or difference) in the retention rate.

5.1 Example 1: LFS data

The LFS is a large monthly household survey of approximately 54,000 households per month, with a focus on gathering information about labour market activities of Canadians. The LFS is a rich source of tenure data. As part of their regular questionnaire, respondents are asked when they started working for their present employer. Table 1 compares retention rate estimators using the master LFS files.¹⁹ I present estimates of the 1-year retention rate for select groups in the year

¹⁶Where the base weights are normalized to sum up to 1 in each sample period.

¹⁷In Appendix A.4., I show that Equation (2) and (7) are numerically equivalent.

¹⁸If one treats \hat{adj}_t as a constant, one can easily show that $se(\hat{Q}_t^{s,c}) = \hat{adj}_t \cdot se(\hat{R}_t^{s,c})$

¹⁹These files were accessed on site at the British Columbia Interuniversity Research Data Centre (BCIRDC). The BCIRDC is run and sponsored by the University of British Columbia, University of Victoria and Simon Fraser University, in collaboration with Statistics Canada.

2000.²⁰

Two important conclusions can be drawn from Table 1. One, the two methods generate very similar 1-year retention rate estimates. As Equation (7) indicates, the two estimates will only differ by a scaling factor, i.e. the population growth factor. Table 1 shows the 2000 growth factor only slightly exceeded 1. As a robustness check, I estimated the population growth factor for each year over the 1977-2003 period. The growth factor averaged 1.0033 over this period, and was close to 1 in all years—despite the fact that the Canadian population has increased over time due to immigration. This is because one is only looking at whether an individual remains with the same employer in the next year, and not, say, 10 years from now. The second conclusion is that the standard errors are relatively small. The loss in efficiency due to the fact that one cannot follow individuals over time in a cross sectional approach is compensated by the large samples of the LFS. As a result, the the probability estimates are very precise. Both sets of standard errors were estimated using proposed methods.

5.2 Example 2: CPS data

Within a retention rate approach, testing for differences in job stability across time or groups is straightforward—only a single restriction needs to be tested. For ease of exposition, I focus on time differences; the arguments are similar when testing across groups. The null hypothesis is $H_0 : R_j^{s,c} - R_1^{s,c} = 0$, where $R_j^{s,c} - R_1^{s,c}$ is the difference in retention rate over a $j - 1$ period. I use a t-test approach. The t-statistic, t_n , is ²¹

$$t_n = \frac{\hat{R}_j^{s,c} - \hat{R}_1^{s,c}}{\sqrt{\hat{V}_{R_j^{s,c} - R_1^{s,c}}/n}} \quad (8)$$

where $\hat{V}_{R_j^{s,c} - R_1^{s,c}}$ is the estimator of $Avar(\hat{R}_j^{s,c} - \hat{R}_1^{s,c})$.

The standard errors estimator must be able to account for the fact that one does not follow individuals over time - as is the case with the proposed approach. The literature has used two approaches. The first approach (e.g. Swinnerton and Wial (1995); Diebold, Neumark, and Polsky (1997)) apply a standard error estimator designed for panel data to the cross sectional case, i.e. it implicitly assumes that one can follow individuals over time. The second approach which was first proposed by Neumark, Polsky, and Hansen (1999), and subsequently applied to Canadian data by Heisz (2005), treats the denominator of $\hat{Q}_t^{s,c}$ in Equation (2) as a constant.²² As such, they fail to

²⁰Following Brochu (2008), all continuation probabilities condition on being 20 to 54 years of age. This imposes the following sample restrictions: the 2000 sample only included those 20 to 54 years of age, while the 2001 sample was restricted to individuals 21 to 55 years of age. See Brochu (2006) for more details.

²¹To simplify the presentation I assume that the cross sections are all of size n . The asymptotic properties of the retention rate differential are left to Appendix A.5.

²²More precisely, they present their retention rate estimator as a ratio of two estimated counts (see Equation (2)) and treat the numerator as a random variable, but the denominator as a constant when they construct their standard error estimator.

account for the full variability of the cross sectional approach.

The standard errors estimator must also be able to account for the possible correlation between $\hat{R}_j^{s,c}$ and $\hat{R}_1^{s,c}$. More precisely, $\hat{R}_2^{s,c}$ and $\hat{R}_1^{s,c}$ may be correlated since both the denominator of $\hat{R}_2^{s,c}$ and the numerator of $\hat{R}_1^{s,c}$ are functions of the same (year 2) observations.²³ By allowing both the numerator and denominator in Equation (3) to have sampling distributions, the proposed approach can easily generate the necessary covariance term. This is not the case for existing methods. The Neumark-Polsky-Hansen (NPH) method, for example, rules out the possibility of any correlation between $\hat{R}_2^{s,c}$ and $\hat{R}_1^{s,c}$.

I use CPS data to illustrate how the choice of standard errors estimator can matter at the inference stage. As with the LFS, the American CPS is a large monthly household which asks respondents about their labour market activities. But contrary to the LFS, a tenure question is not part of the regular CPS questionnaire—it is only included in select supplements. I therefore rely on 4-year retention rates—as was previously done in the American job stability literature. Finally, I rely on the $\tilde{Q}_t^{s,c}$ estimator (instead of $\tilde{R}_t^{s,c}$). Assuming no change in the underlying population would be too restrictive an assumption.

Tables 2 and 3 examine changes in 4-year retention rates from 1996 to 2000 for males and females, respectively.²⁴ Standard errors are calculated for the NPH method, the DNP method,²⁵ the proposed method, and the proposed method with no covariance term. In all cases, weights were used to make a clearer comparison of the various methods. Robustness checks for different years and sub-populations indicate that weights do not significantly affect the results.

Tables 2 and 3 indicate that accounting for the covariance term can increase or decrease the standard errors.²⁶ Even when the covariance term is negative, a consistent pattern emerges with the proposed method generating standard errors consistently larger than either DNP or NPH methods. This pattern was found to be robust for other time periods and other sub-populations. The proposed method generates standard errors that are up to 172.9% larger than the DNP estimates and up to 55.6% larger than the NPH estimates. From Tables 2 and 3 one can observe a systematic differential in the standard errors; the gap is larger for longer tenured groups - groups with higher job stability. In general, the extent to which the NPH method underestimates the correct standard errors will be correlated with the size of the employment continuation probability.²⁷

As a result, the DNP and NPH approaches to estimating standard errors may lead the researcher to falsely reject the null hypothesis of no change in job stability. Calculating t-statistics for males

²³A similar difficulty occurs when testing across groups, say, groups A and B. The numerators (and denominators) of R_j^A and R_j^B may also be correlated.

²⁴All U.S. continuation probabilities condition on being at least 16 years of age, and not being self-employed.

²⁵The DNP method, refers to the application of the longitudinal standard errors estimator to cross sectional data. Diebold, Neumark, and Polsky (1997) may not have been the first to use the method with cross sectional data, but they were one of the first to provide a detailed explanation of the approach.

²⁶From Equations (33) and (35) one can see that the covariance term will be positive if and only if the covariance between D_{i2} and N_{i2} is also positive.

²⁷Focussing on Equation (14) illustrates this point. Conditioning on a sample distribution for D_{i1} , a larger $E(N_{it+1})$ is associated with a larger first term; a variance term not accounted for by the NPH method.

with 12+ years of tenure illustrates this point. Using either the DNP or NPH methods, one strongly rejects the null hypothesis at the 5% significance level. In fact, my method suggests that the null hypothesis should not be rejected, not even at the 10% level.

6 Conclusion

In this paper, I propose a population cohort approach for estimating labour market transitions. Based on this framework, I propose a non-parametric estimator for the continuation probability and another for its standard errors.

Using the population cohort framework, I also re-examine existing cross sectional methods. I identify the underlying assumptions of the non-parametric approach used in the job stability literature, and propose a consistent estimator for its standard errors.

Finally, I show using CPS data that the choice of method (for constructing standard errors) can make a difference at the inference stage. In particular, that using existing methods can lead to the researcher to conclude that job stability has changed when in fact it has not.

A Appendix

A.1

Proposition 1 *The retention rate of a worker with time-invariant characteristics c who has been employed for s periods at time t can be expressed as $R_t^{s,c} = \frac{E(1[TEN_{it+1}=s+1, X_{it+1}=c])}{E(1[TEN_{it}=s, X_{it}=c])}$.*

proof:

$$R_t^{s,c} = \text{Prob}(TEN_{it+1} = s + 1 | TEN_{it} = s, X_{it} = c) \quad (9)$$

$$= \frac{\text{Prob}(TEN_{it+1} = s + 1, TEN_{it} = s, X_{it} = c)}{\text{Prob}(TEN_{it} = s, X_{it} = c)} \quad (10)$$

and since $TEN_{it+1} = s + 1$ implies $TEN_{it} = s$, one can rewrite $R_t^{s,c}$ as

$$= \frac{\text{Prob}(TEN_{it+1} = s + 1, X_{it} = c)}{\text{Prob}(TEN_{it} = s, X_{it} = c)} \quad (11)$$

$$= \frac{\text{Prob}(TEN_{it+1} = s + 1, X_{it+1} = c)}{\text{Prob}(TEN_{it} = s, X_{it} = c)} \quad (12)$$

$$= \frac{E(1[TEN_{it+1} = s + 1, X_{it+1} = c])}{E(1[TEN_{it} = s, X_{it} = c])} \quad \blacksquare \quad (13)$$

A.2

For ease of exposition, define $N_{it+1} = 1[TEN_{it+1} = s + 1, X_{it+1} = c]$ and $D_{it} = 1[TEN_{it} = s + 1, X_{it} = c]$.

Proposition 2 *Assuming iid samples for each year that are drawn from a population cohort, independence across years, and $\lim_{n_t, n_{t+1} \rightarrow \infty} \frac{n_t}{n_{t+1}} = 1$, then $\sqrt{n_t}(\hat{R}_t^{s,c} - R_t^{s,c}) \xrightarrow{d} N(0, V)$ where V is*

$$V = \phi_1^2 V(D_{it}) + \phi_2^2 V(N_{it+1}) \quad (14)$$

with

$$\phi_1 = \frac{E(N_{it+1})}{[E(D_{it})]^2}, \quad \phi_2 = \frac{1}{E(D_{it})} \quad (15)$$

proof:

a) (consistency)

Apply the Lindberg-Levy Central Limit Theorem

$$\sum_{i=1}^{n_{t+1}} N_{it+1}/n_{t+1} \xrightarrow{p} E(N_{it+1}) \quad (16)$$

$$\sum_{i=1}^{n_t} D_{it}/n_t \xrightarrow{p} E(D_{it}) \quad (17)$$

and use the result of Proposition 1 \blacksquare

b) (Asymptotic normality)

Let $\hat{N}_{t+1} = n_{t+1}^{-1} \sum_{i=1}^{n_{t+1}} N_{it+1}$, $N_{t+1} = E(N_{it+1})$ and $V_{N_{t+1}} = V(N_{it+1})$, and define \hat{D}_t , D_t and V_{D_t} in a similar fashion.

$$\sqrt{n_t}(\hat{R}_t^{s,c} - R_t^{s,c}) = \sqrt{n_t} \left(\frac{\hat{N}_{t+1}}{\hat{D}_t} - \frac{N_{t+1}}{D_t} \right) \quad (18)$$

$$= \sqrt{n_t} \frac{(\hat{N}_{t+1} - N_{t+1})D_t - (\hat{D}_t - D_t)N_{t+1}}{D_t \hat{D}_t} \quad (19)$$

$$= \frac{\sqrt{\frac{n_t}{n_{t+1}}} \sqrt{n_{t+1}} (\hat{N}_{t+1} - N_{t+1})D_t - \sqrt{n_t} (\hat{D}_t - D_t)N_{t+1}}{D_t^2} + o_p(1) \quad (20)$$

$$= -\phi_1 \sqrt{n_t} (\hat{D}_t - D_t) + \phi_2 \sqrt{n_{t+1}} (\hat{N}_{t+1} - N_{t+1}) + o_p(1) \quad (21)$$

$$\xrightarrow{d} N(0, \phi_1^2 V_{D_t} + \phi_2^2 V_{N_{t+1}}) \quad \blacksquare \quad (22)$$

Replacing the population moments in Equation (22) with the corresponding sample analogs generates a consistent estimator for the asymptotic variance. Taking the square root of the estimated variance will generate the standard errors.

A.3

Proposition 3 *Assume that the composition of a country's population changes from period t to period $t + 1$. Further assume that these compositional changes break (or interrupt) the spell of interest. The retention rate can be expressed as*

$$R_t^{s,c} = \text{adj}_t \frac{E(1[TEN_{it+1} = s + 1, X_{it+1} = c])}{E(1[TEN_{it} = s, X_{it} = c])} \quad (23)$$

where adj_t is the population growth factor.

proof: To ease the presentation, I assume that the change in population is due to the arrival of one new immigrant in year $t + 1$. Similar arguments would hold true for other population changes. Without loss of generality, assume a population of size N in year t , and $N + 1$ in year $t + 1$. Order the year $t + 1$ population so that the new immigrant is last. By Proposition 1, the retention is

$$R_t^{s,c} = \frac{\sum_{i=1}^N 1[TEN_{it+1} = s + 1, X_{it+1} = c]/N}{\sum_{i=1}^N 1[TEN_{it} = s, X_{it} = c]/N} \quad (24)$$

By assuming that the change in population results in breaks in the spell of interest, one can conclude that $1[TEN_{N+1,t+1} = s + 1, X_{N+1,t+1} = c] = 0$. As a result, $R_t^{s,c}$ can be rewritten as

$$= \left(\frac{N + 1}{N} \right) \frac{\sum_{i=1}^{N+1} 1[TEN_{it+1} = s + 1, X_{it+1} = c]/N + 1}{\sum_{i=1}^N 1[TEN_{it} = s, X_{it} = c]/N} \quad (25)$$

$$\equiv \text{adj}_t \frac{E(1[TEN_{it+1} = s + 1, X_{it+1} = c])}{E(1[TEN_{it} = s, X_{it} = c])} \quad \blacksquare \quad (26)$$

A.4

Proposition 4 *Given repeated cross sections where the base weights sum up to the target population, then*

$$\tilde{Q}_t^{s,c} = \hat{adj}_t \left(\frac{\sum_{i=1}^{n_{t+1}} nw_{it+1} 1[TEN_{it+1} = s+1, X_{it+1} = c]/n_{t+1}}{\sum_{i=1}^{n_t} nw_{it} 1[TEN_{it} = s, X_{it} = c]/n_t} \right) \quad (27)$$

proof:

The existing cross sectional estimator is presented as a ratio of two population counts. It takes form

$$Q_t^{s,c} = \frac{\tilde{n}_{t+1}^{s+1,c}}{\tilde{n}_t^{s,c}} \quad (28)$$

where $\tilde{n}_t^{s,c}$ is the sum of the base weights of all individuals with characteristics c who have been employed speriods as of period t .

For ease of exposition, define $N_{it+1} = 1[TEN_{it+1} = s+1, X_{it+1} = c]$ and $D_{it} = 1[TEN_{it} = s+1, X_{it} = c]$. One can rewrite this estimator as

$$\tilde{Q}_t^{s,c} = \frac{\sum_{i=1}^{n_{t+1}} bw_{it+1} N_{it+1}}{\sum_{i=1}^{n_t} bw_{it} D_{it}} \quad (29)$$

$$= \frac{\sum_{i=1}^{n_{t+1}} bw_{it+1}}{\sum_{i=1}^{n_t} bw_{it}} \cdot \frac{\sum_{i=1}^{n_{t+1}} \frac{bw_{it+1}}{\sum_{i=1}^{n_{t+1}} bw_{it+1}/n_{t+1}} N_{it+1}/n_{t+1}}{\sum_{i=1}^{n_t} \frac{bw_{it}}{\sum_{i=1}^{n_t} bw_{it}/n_t} D_{it}/n_t} \quad (30)$$

$$= \frac{\sum_{i=1}^{n_{t+1}} bw_{it+1}}{\sum_{i=1}^{n_t} bw_{it}} \cdot \frac{\sum_{i=1}^{n_{t+1}} nw_{it+1} N_{it+1}/n_{t+1}}{\sum_{i=1}^{n_t} nw_{it} D_{it}/n_t} \quad (31)$$

$$= \hat{adj}_t \frac{\sum_{i=1}^{n_{t+1}} nw_{it+1} N_{it+1}/n_{t+1}}{\sum_{i=1}^{n_t} nw_{it} D_{it}/n_t} \quad \blacksquare \quad (32)$$

A.5

Proposition 5 *Assuming iid samples for each year, samples of equal size, independence across years, and no change in population, then $\sqrt{n}((\hat{R}_j - \hat{R}_1) - (R_j - R_1)) \xrightarrow{d} N(0, V)$ where V depends on j , an integer greater than or equal to 2. Case 1: $j = 2$*

$$V = \phi_1^2 V(D_{i1}) + \phi_2^2 V(N_{i2}) + \phi_3^2 V(D_{i2}) + \phi_4^2 V(N_{i3}) + 2\phi_2\phi_3\mu Cov(D_{i2}, N_{i2}) \quad (33)$$

Case 2: $j \geq 3$

$$V = \phi_1^2 V(D_{i1}) + \phi_2^2 V(N_{i2}) + \phi_3^2 V(D_{ij}) + \phi_4^2 V(N_{ij+1}) \quad (34)$$

with

$$\phi_1 = \frac{E(N_{i2})}{[E(D_{i1})]^2}, \quad \phi_2 = \frac{1}{E(D_{i1})}, \quad \phi_3 = \frac{E(N_{ij+1})}{[E(D_{ij})]^2}, \quad \phi_4 = \frac{1}{E(D_{ij})} \quad (35)$$

and μ is the probability that a random chosen person in the population from which D_{it} is drawn, is also part of the population from which N_{it} is drawn.

proof: For ease of notation let $\hat{N}_j = n_j^{-1} \sum_{i=1}^{n_j} N_{ij}$, $N_j = E(N_{ij})$ and $V_{N_j} = V(N_{ij})$, and define \hat{D}_j , D_j and V_{D_j} in a similar fashion. Finally, let $C_2 = Cov(D_{i2}, N_{i2})$

Case 1: $j = 2$

$$\begin{aligned} & \sqrt{n}((\hat{R}_2 - \hat{R}_1) - (R_2 - R_1)) \\ &= \sqrt{n} \left(\left(\frac{\hat{N}_3}{\hat{D}_2} - \frac{N_3}{D_2} \right) - \left(\frac{\hat{N}_2}{\hat{D}_1} - \frac{N_2}{D_1} \right) \right) \end{aligned} \quad (36)$$

$$= \sqrt{n} \frac{(\hat{N}_3 - N_3)D_2 - (\hat{D}_2 - D_2)N_3}{D_2 \hat{D}_2} - \sqrt{n} \frac{(\hat{N}_2 - N_2)D_1 - (\hat{D}_1 - D_1)N_2}{D_1 \hat{D}_1} \quad (37)$$

$$= \sqrt{n} \frac{(\hat{N}_3 - N_3)D_2 - (\hat{D}_2 - D_2)N_3}{D_2^2} - \sqrt{n} \frac{(\hat{N}_2 - N_2)D_1 - (\hat{D}_1 - D_1)N_2}{D_1^2} + o_p(1) \quad (38)$$

$$= \phi_1 \sqrt{n}(\hat{D}_1 - D_1) - \phi_2 \sqrt{n}(\hat{N}_2 - N_2) - \phi_3 \sqrt{n}(\hat{D}_2 - D_2) + \phi_4 \sqrt{n}(\hat{N}_3 - N_3) + o_p(1) \quad (39)$$

$$\xrightarrow{d} N(0, \phi_1^2 V_{D_1} + \phi_2^2 V_{N_2} + \phi_3^2 V_{D_2} + \phi_4^2 V_{N_3} + 2\phi_2 \phi_3 \mu C_2) \quad (40)$$

Case 2: $j \geq 3$. The proof is similar to Case 1, with one exception. Since the four components of the test statistics, i.e. \hat{N}_{j+1} , \hat{D}_j , \hat{N}_2 and \hat{D}_1 are functions of different yearly samples when $j \geq 3$, the covariance term is zero. ■

Replacing the population moments with corresponding sample analogs generates a consistent estimator for each asymptotic variance.

References

- BAKER, M. (1992): “Unemployment Duration: Compositional Effects and Cyclical Variability,” *American Economic Review*, 82(1), 313–321.
- BROCHU, P. R. (2006): “An Exploration in Job Stability,” PhD Thesis, University of British Columbia.
- (2008): “Rising Job Stability in the 1990s: the Impact of Compositional Change,” Unpublished Manuscript, University of Ottawa.
- DEATON, A. (1985): “Panel Data from Time Series of Cross-Sections,” *Journal of Econometrics*, 30, 109–126.
- DIEBOLD, F. X., D. NEUMARK, AND D. POLSKY (1997): “Job Stability in the United States,” *Journal of Labor Economics*, 15(2), 206–233.
- GÜELL, M., AND L. HU (2006): “Estimating the Probability of Leaving Unemployment Using Uncompleted Spells from Repeated Cross-section Data,” *Journal of Econometrics*, 133(1), 307–341.
- HEISZ, A. (2005): “The Evolution of Job Stability in Canada: Trends and Comparisons with U.S. Results,” *Canadian Journal of Economics*, 38(1), 105–127.
- JONES, S. R. G., AND W. C. RIDDELL (1999): “The Measurement of Unemployment: An Empirical Approach,” *Econometrica*, 67(1), 147–162, Notes and Comments.
- MOFFITT, R. (1993): “Identification and Estimation of Dynamic Models with a Time Series of Repeated Cross-section,” *Journal of Econometrics*, 59, 99–123.
- NEUMARK, D., D. POLSKY, AND D. HANSEN (1999): “Has Job Stability Declined Yet? New Evidence for the 1990s,” *Journal of Labor Economics*, 17(4), S29–S64.
- SWINNERTON, K. A., AND H. WIAL (1995): “Is Job Stability Declining in the U.S. Economy?,” *Industrial and Labor Relations Review*, 48(2), 293–304.
- VERBEEK, M., AND F. VELLA (2004): “Estimating Dynamic Models from Repeated Cross Sections,” *Journal of Econometrics*, 127, 83–102.

Table 1: 1-year Employment-Continuation Probabilities: Canada, 2000

Group Specification	Proposed Method (\tilde{R}_1)	Existing Method (\tilde{Q}_1)	Population Growth Factor (\hat{adj}_1)
Overall	0.8040 (0.0028)	0.8101 (0.0029)	1.0076
Male	0.7975 (0.0050)	0.8036 (0.0051)	1.0076
Female	0.8108 (0.0052)	0.8170 (0.0053)	1.0076
Tenure less than 1 year	0.5517 (0.0070)	0.5559 (0.0071)	1.0076

Table 2: U.S. 4-year Male Employment-Continuation Probabilities - Time Differentials

Tenure Group Specification	1996	2000	Difference	Standard Errors Method
0-2	0.4775	0.4999	0.0224 (0.0076)** (0.0106) ** (0.0116) ** (0.0118) *	DNP NPH proposed proposed (no covariance term)
3-6	0.4522	0.4658	0.0136 (0.0092) (0.0133) (0.0148) (0.0151)	DNP NPH proposed proposed (no covariance term)
7-11	0.7069	0.6554	-0.0515 (0.0111)** (0.0197)** (0.0250)** (0.0244)**	DNP NPH proposed proposed (no covariance term)
12+	0.7288	0.6927	-0.0362 (0.0083)** (0.0149)** (0.0227) (0.0194)*	DNP NPH proposed proposed (no covariance term)
total	0.5798	0.5716	-0.0081 (0.0045)* (0.0060) (0.0083) (0.0073)	DNP NPH proposed proposed (no covariance term)

** The estimated difference is significant at the 5% level

* The estimated difference is significant at the 10% level

Table 3: U.S. 4-year Female Employment-Continuation Probabilities - Time Differentials

Tenure Group Specification	1996	2000	Difference	Standard Errors Method
0-2	0.4294	0.4785	0.0491 (0.0074)** (0.0097)** (0.0111)** (0.0113)**	DNP NPH proposed proposed (no covariance term)
3-6	0.4370	0.4253	-0.0117 (0.0092) (0.0125) (0.0148) (0.0150)	DNP NPH proposed proposed (no covariance term)
7-11	0.6630	0.6134	-0.0496 (0.0106)** (0.0206)** (0.0270)* (0.0272)*	DNP NPH proposed proposed (no covariance term)
12+	0.7264	0.6699	-0.0565 (0.0098)** (0.0172)** (0.0268)** (0.0231)**	DNP NPH proposed proposed (no covariance term)
total	0.5355	0.5322	-0.0033 (0.0047) (0.0060) (0.0085) (0.0075)	DNP NPH proposed proposed (no covariance term)

** The estimated difference is significant at the 5% level

* The estimated difference is significant at the 10% level