

A REFINEMENT OF THE REGRESSION COMPOSITE ESTIMATOR IN THE LABOUR FORCE SURVEY FOR CHANGE ESTIMATES

Jean-François Beaumont¹ and Cynthia Bocci²

ABSTRACT

The Labour Force Survey is a stratified multi-stage monthly survey with a rotating sample design. Its main objective is to provide estimates of employment and unemployment rates. Estimates are obtained by composite regression which uses covariates defined with previous month variables of interest. For the new entrants in the survey, these covariates are missing. This problem is currently dealt with by imputing the missing values and by modifying the values of non-missing units in order to maintain the design-consistency property of the resulting estimator. We propose a new way of dealing with missing values in the covariates using calibrated imputation.

KEY WORDS: Calibrated imputation, Composite regression, Missing covariates.

RÉSUMÉ

L'enquête sur la population active est une enquête mensuelle, stratifiée à plusieurs degrés avec renouvellement de l'échantillon. L'objectif principal de l'enquête est de produire des estimations de taux d'emploi et de chômage. Les estimations sont obtenues par la méthode de la régression composite qui utilise des variables auxiliaires définies à partir de variables d'intérêt du mois précédent. Pour le groupe de renouvellement naissant, ces variables auxiliaires sont manquantes. Présentement, ce problème est réglé en imputant les valeurs manquantes et en modifiant les valeurs des unités non-manquantes de telle sorte que l'estimateur qui en résulte maintienne la propriété de convergence par rapport au plan. On propose une nouvelle façon de traiter les valeurs manquantes dans les variables auxiliaires en utilisant l'imputation calée.

MOTS CLÉS : Imputation calée; régression composite; variables auxiliaires manquantes.

1. INTRODUCTION

The Labour Force Survey (LFS) is a stratified multi-stage monthly survey with a rotating sample design. Each month 1/6 of the LFS sample dwellings is replaced by new dwellings. These new dwellings form the birth rotation group. Each new dwelling stays in the sample for six consecutive months. Information is collected for every person in the dwelling.

The main objective of the LFS is to provide estimates of employment and unemployment rates at the provincial and regional levels. Estimation is obtained by composite regression, which uses some covariates defined using information collected in the previous month. For the new entrants in the survey (i.e. those in the birth rotation) the previous month covariates are missing. This problem is currently dealt with by imputing the missing values and by modifying the values of non-missing units in order to maintain the design-consistency property of the estimator. Essentially, the LFS regression composite estimator is a regression estimator with missing values in the covariates and some estimated benchmarks.

¹ Jean-François Beaumont, Statistics Canada, Business Survey Methods Division, RHC 11th floor, Tunney's Pasture, Ottawa, Ontario K1A 0T6, Jean-Francois.Beaumont@statcan.ca

² Cynthia Bocci, Statistics Canada, Business Survey Methods Division, RHC 11th floor, Tunney's Pasture, Ottawa, Ontario K1A 0T6, Cynthia.Bocci@statcan.ca

Details of the LFS regression composite estimator can be found in the special section of the June 2001 issue of *Survey Methodology* (Singh, Kennedy and Wu (2001); Fuller and Rao (2001); Gambino, Kennedy and Singh (2001)).

We propose a new way of dealing with missing values in the covariates using the idea of calibrated imputation (Beaumont (2005)). One advantage of the proposed method is that it is not necessary to modify non-missing values to maintain the design-consistency property. This should help preserve the relationship between the covariates and the variables of interest and reduce the variance. Also, the proposed estimator does not depend on the choice of an arbitrary tuning constant as in the current method. An empirical investigation shows that the proposed method seems promising for estimates of month-to-month change.

Section 2 describes the composite regression estimation method currently in place for the LFS and the proposed refinement. This is followed by a study and results presented in the section 3.

2. COMPOSITE ESTIMATION REGRESSION METHOD

Composite regression estimation is a modified regression method of estimation applicable to periodic surveys with rotating sample design whereby the resulting estimator uses information from previous periods. This modified regression involves not only the usual auxiliary variables observed at time t but also auxiliary variables from the previous time period $t-1$, called composite auxiliary variables. The idea is to improve the estimate for time t by incorporating information from time $t-1$. Since the composite regression estimator is a function of the previous period composite estimator, it is a recursive estimator. This recursive process is one where the composite auxiliary variables have random benchmarks determined by setting the weighted sum of a composite auxiliary variable equal to last month's estimate. The composite start date is the first time period from which this recursive process begins. In this context, composite regression differs from generalized regression in that the matrix of auxiliary variables is augmented with variables whose values depend on the previous period and are missing for new entrants.

The main focus of this paper is the construction of the composite auxiliary variables for the LFS. Section 2.1 provides the details of this method as it is currently applied to the LFS. Section 2.2 suggests a refinement to the current estimation strategy in the LFS by proposing a new definition for the composite auxiliary variables.

2.1 Composite Regression Estimation Method in the LFS

Composite regression estimation lends itself very nicely to the LFS given the rotating panel design of the survey. Five-sixths of the sample at time t will also have been in the sample at time $t-1$. This part of the sample which is common from one month to the next is referred to as the overlap sample. Recall that the remaining 1/6 of the sample is known as the birth rotation or birth panel.

Presently in the LFS, the matrix of regression covariates is comprised of columns of demographic auxiliary variables and 25 composite auxiliary variables for each row k representing an individual. In practice, there are column constraints such that the weighted sum of the demographic auxiliary variable over all k is equal to the demographic census projection whereas the weighted sum of a composite auxiliary variable is set equal to last month's estimate. The latter sum is referred to as a composite control total. Under these constraints, the regression weights are constructed to define the estimator for the current period. These weights have thus been calibrated.

Of particular interest in this subsection is the definition of the composite auxiliary variable currently being used in the LFS. Let

- S_t = the LFS sample at month t
- A_t = the overlap (matching) sample for month t
- B_t = the birth rotation sample for month t

Consider only those in S_t who are 15 years or older which results in a sample size of n . Denote x_{ctk} the composite auxiliary variable c in month t for person k , where $c=1, \dots, 25$ and $k=1, \dots, n$. Ideally, x_{ctk} would be equal to $y_{c,t-1,k}$, the indicator variable for a particular characteristic of interest. However, $y_{c,t-1,k}$ is not available for those in the birth

rotation so that these new entrants to the survey have missing values for the composite auxiliary variables. Fix a composite auxiliary variable c so that x_{ctk} is expressed as $x_{\bullet tk}$. To deal with these missing values, the LFS currently defines a composite auxiliary variable as

$$x_{\bullet tk} = (1 - \alpha)MR1_{tk} + \alpha MR2_{tk}, \quad \text{with } \alpha = \frac{2}{3}, \quad (1)$$

where

$$MR1_{tk} = \begin{cases} y_{t-1,k} & k \in A_t \\ \hat{\mu}_{t-1} & k \in B_t \end{cases}, \quad MR2_{tk} = \begin{cases} y_{t-1,k} + (y_{t-1,k} - y_{tk})(\theta_{tk}^{-1} - 1) & k \in A_t \\ y_{tk} & k \in B_t \end{cases}, \quad (2)$$

and $\hat{\mu}_{t-1}$ is the composite estimator at time $t-1$ of the proportion of people aged 15 and over with the specified characteristic of interest and $\theta_{tk} = P(k \in A_t | S_t)$. In the LFS, the value of θ_{tk} is constant over t and k and is equal to $5/6$. The choice of $\alpha = \frac{2}{3}$ is taken to satisfy the estimation objectives of the LFS and was studied in great detail by Chen and Liu (2002). It offers a compromise between variables $MR1$ and $MR2$ that has previously been shown empirically to be adequate for level-driven and change-driven estimators respectively.

Notice that $MR1$ uses mean imputation to impute the missing values for those in the birth rotation and the previous month value for those in the overlap sample. In contrast, $MR2$ uses carry backward imputation for those in the birth sample and modifies the previous month values of those in the overlap sample to maintain design consistency property. The forms of $MR1$ and $MR2$ are themselves derived so that given the sample, the Horvitz-Thompson estimator of the total for $x_{\bullet tk}$ is unbiased for the Horvitz-Thompson estimator of the total of $y_{t-1,k}$. Note that the integrated method of weighting of LeMaître and Dufour (1987) is used to insure that all members of the same household have a common weight. As a result, the matrix of regression covariates is slightly modified.

2.2 Proposed Refinement of Current Composite Regression Estimation in the LFS

The proposed refinement simply involves redefining the composite auxiliary variable $x_{\bullet tk}$ given in (1). Consider the composite auxiliary variable

$$MRR_{tk} = \begin{cases} y_{t-1,k} & k \in A_t \\ y_{tk} + \frac{1 - \theta_{tk}}{\theta_{tk}} \frac{\sum_{l \in A_t \cap P_k} w_l (y_{t-1,l} - y_{tl})}{\sum_{l \in B_t \cap P_k} w_l} & k \in B_t \end{cases} \quad (3)$$

where w_l is a weight before calibration and P_k is the province of residence corresponding to individual k . Now define $x_{\bullet tk} = MRR_{tk}$.

The motivation behind the definition (3) comes from calibrated imputation (Beaumont 2005). For the birth sample, the values for the preceding month, $y_{t-1,k}$, are missing and must be imputed. These values are imputed in such a way so as to minimise the sum of weighted squared differences between $x_{\bullet tk}$ and y_{tk} under calibration constraints. Essentially, the idea consists of imputing the missing $y_{t-1,k}$ by the preliminary imputed values y_{tk} and then calibrating these imputed values to obey appropriate constraints. From this perspective, it is not necessary to modify non-missing values to maintain the design-consistency property as is done in $MR2$. An attractive feature of MRR over the composite estimator given in (1) is that it does not involve an arbitrary choice of a tuning constant α .

3. EMPIRICAL STUDY AND RESULTS

In order to evaluate the performance of the proposed composite estimator described in Section 2.2, we compare its estimated variance to that obtained by 3 other composite estimators relative to the estimated variance of the generalized

regression estimator (GREG). The measure used to evaluate each of the four composite estimation schemes is the relative efficiency (RE), defined as

$$RE = 100 * Var(GREG) / Var(composite estimator)$$

All the estimated variances are calculated by the Jackknife method. The four different composite auxiliary variables used in this study are described below.

MR1MR2: The *current* composite estimation method using the composite auxiliary variables *MR1* and *MR2* in the linear combination $(1 - \alpha) MR1 + \alpha MR2$, with $\alpha = \frac{2}{3}$.

MR1: The composite estimation method using the composite auxiliary variable *MR1* only-equivalent to setting $\alpha = 0$ in the linear combination

MR2: The composite estimation method using the composite auxiliary variable *MR2* only-equivalent to setting $\alpha = 1$ in the linear combination

MRR: The proposed refinement to the current composite estimation method using the composite auxiliary variable *MRR*

The relative efficiencies are calculated for several sets of estimates including both level and month-to-month change estimates. Specifically, we obtain relative efficiencies for estimates of employment and unemployment totals by province, estimates of totals of labour force, employment, unemployment and unemployment rate by sex at the Canada level and estimates of employment total by class of worker, sector and industry at the Canada level.

In order to avoid seasonal effects, the *average* relative efficiency over the 12 month period of LFS data from July 2000 to June 2001 is calculated, with the first month of compositing starting from February 1998. This time period and composite start date were chosen to mirror the study conducted by Chen and Liu (2002) who investigated the choice of alpha in the composite regression estimation method currently used in the LFS as described in Section 2.1. The LFS data used in this study was based on 1996 census projections.

Tables of relative efficiencies for estimates of employment and unemployment totals by province are shown below. In Table 1A, it is clear that the current composite estimation method (MR1MR2) outperforms the others for the level estimates of employment. It also performs adequately as compared with the other methods for the estimates of unemployment. Neither MRR nor surprisingly MR1 perform as well as expected for the level estimates. In Table 1B, the provincial employment estimates of month-to-month change are much more efficient than the provincial unemployment estimates with the exception of the MR1 method. Furthermore, the proposed method (MRR) has overall higher relative efficiency for the month-to-month change estimates of provincial employment. MRR also performs well comparatively in the estimation of month-to-month change in unemployment.

Although not shown here, similar results hold for the other estimates. Specifically, the relative efficiencies of month-to-month change estimates by labour force status and sex are higher than those of the level estimates with the exception of the MR1 method. Once again the proposed method generally outperforms the other methods for estimates of month-to-month change. Similar results hold for the remaining estimates.

Table 1A: Average relative efficiency (RE) of LEVEL estimates of employment and unemployment by province for 4 estimation schemes. RE averaged over a 12 month period from July 2000 to June 2001 with composite start date of February 1998. RE rounded to the nearest tenth..

Labour Force Status	Province	Method			
		MRR	MR1MR2 $\alpha = .67$	MR1	MR2
Employment	NF	111.4	123.2	103.1	117.7
	PEI	101.1	120.0	104.7	116.6
	NS	101.3	123.8	104.7	114.2
	NB	107.0	124.0	106.0	118.3
	QC	103.6	121.8	103.0	117.5
	ONT	100.9	126.0	106.7	118.8
	MAN	108.7	130.3	107.7	124.2
	SASK	105.1	122.0	103.2	118.9
	ALB	106.0	121.9	106.7	117.4
	BC	105.8	124.4	106.6	118.3
	CANADA	102.9	124.0	105.4	118.0
	Unemployment	NF	98.9	105.0	106.1
	PEI	106.4	116.2	108.1	114.0
	NS	94.2	103.1	106.2	100.9
	NB	83.5	97.0	104.2	92.7
	QC	96.9	104.9	104.8	102.7
	ONT	93.9	103.0	104.3	110.4
	MAN	98.9	105.6	103.4	103.2
	SASK	100.1	103.0	98.4	102.8
	ALB	93.5	102.9	104.0	100.3
	BC	96.0	103.8	105.9	100.7
	CANADA	95.0	103.7	104.6	101.0

Table 1B: Average relative efficiency (RE) of month-to-month CHANGE estimates of employment and unemployment by province for 4 estimation schemes. RE averaged over a 12 month period from July 2000 to June 2001 with composite start date of February 1998. RE rounded to the nearest tenth.

Labour Force Status	Province	Method			
		MRR	MR1MR2 $\alpha = .67$	MR1	MR2
Employment	NF	157.8	144.6	90.5	152.4
	PEI	148.0	140.9	93.9	147.0
	NS	163.3	149.0	92.0	158.3
	NB	161.7	151.3	98.1	157.9
	QC	162.0	151.5	95.5	159.5
	ONT	165.4	154.2	96.8	162.2
	MAN	150.5	139.6	92.1	146.8
	SASK	165.2	149.1	95.8	159.1
	ALB	151.3	142.5	99.7	147.9
	BC	164.9	152.6	104.1	159.9
	CANADA	161.9	151.1	97.2	158.7
	Unemployment	NF	108.4	106.4	103.5
	PEI	126.9	123.7	104.7	125.9
	NS	107.9	107.8	107.0	107.1
	NB	100.3	105.5	107.5	102.3
	QC	110.1	109.7	104.3	109.5
	ONT	107.2	108.0	105.5	106.5
	MAN	111.3	110.4	105.0	108.6
	SASK	110.8	106.9	99.9	108.4
	ALB	105.0	106.6	104.6	105.3
	BC	108.4	108.5	106.7	106.8
	CANADA	108.0	108.5	104.9	107.4

4. CONCLUSIONS

The empirical investigation carried out in Section 3 shows that the proposed method seems promising with respect to the relative efficiency gains for month-to-month change estimates but not so for level estimates. The LFS, however, does produce level estimates; and it appears that the current method satisfies both requirements adequately.

An area of future study to refine the composite estimation method yet further might be to consider a linear combination of the composite auxiliary variables MRR and $MR1$. This may marginally improve the efficiency of some estimates although it would again involve an arbitrary tuning constant.

A better way to refine the proposed calibrated imputation approach would be to start with better preliminary imputed values in MRR as defined in equation (3). For example, preliminary imputed values could be obtained using regression imputation. This could be investigated in a future study.

REFERENCES

- Beaumont, J.F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach, *Journal of the Royal Statistical Society, Series B*, Vol.67, issue 3, pp. 445-458.
- Chen, Edward J. and Liu, T. P.(2002)., Choices of alpha value in regression composite estimation for the Canadian Labour Force Survey: Impacts and evaluation, Internal document, Statistics Canada, Household Survey Methods Division.
- Fuller, Wayne A. and Rao, J.N.K.(2001). A regression composite estimator with application to the Canadian Labour Force Survey, *Survey Methodology*, Vol.27, No.1, pp.45-51.
- Gambino, J., Kennedy, B. and Singh, M.P.(2001). Regression composite estimation for the Canadian Labour Force Survey: Evaluation and implementation, *Survey Methodology*, Vol.27, No.1, pp.65-74.
- Lemaître, G.E. and Dufour, J.(1987). An integrated method for weighting persons and families, *Survey Methodology*, Vol.13, pp.199-207.
- Singh, A.C., Kennedy, B., Wu (2001). Regression composite estimation for the Canadian Labour Force Survey with a rotating panel design, *Survey Methodology*, Vol.27, No.1, pp.33-44.