Economics Letters 155 (2017) 19–23

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

Employment misclassification in survey and administrative reports

Dean R. Hyslop, Wilbur Townsend *

Motu Economic and Public Policy Research, New Zealand

HIGHLIGHTS

- We consider matched survey and administrative reports of employment.
- True employment and misclassification rates are identified from these two reports.
- Both reports have error, but the administrative data is more accurate.
- False positive employment rates are higher than false negative rates.
- Misclassification error substantially affects estimated employment rates.

ARTICLE INFO

Article history:
Received 7 December 2016
Received in revised form 20 February 2017
Accepted 12 March 2017
Available online 14 March 2017

JEL classification:
J6
C18
J21

Keywords:
Unemployment rate
Measurement error
Validation study

ABSTRACT

This paper analyses measurement error in the classification of employment using matched survey and administrative data from New Zealand. We show that the true employment rate and time-invariant error rates can be identified, given access to two measures of employment with independent errors. Empirical identification requires data with time varying employment rates over at least two periods. We find that both measures have error, with the administrative data being substantially more accurate than the survey data, and false positives are much more likely than false negatives in both sources. Allowing for errors substantially affects estimated employment rates.

1. Introduction

Measurement error affects analysis of survey data (Bound et al., 2001), and a substantial literature has found measurement error in employed workers' earnings.[1] However there is little research on the mismeasurement of employment itself. This is surprising given that studies of employment dynamics require assumptions about error in their data;[2] and that studies of earnings error which omit the unemployed may be biased because they select on a non-random subsample.[3] This paper presents a model of measurement error in two measures of employment which is identified under relatively weak conditions. The model is estimated using matched survey and administrative data.

The literature on misclassification in employment typically either relies on a comparison measure that is assumed to be correct, or lacks a comparison measure entirely. Abowd and Zellner (1985) and Poterba and Summers (1986) both analyse job flows from reported employment status using US Current Population Survey (CPS) validation reinterview data, assuming that the reconciled reinterview reports represent the true employment status. Keane and Sauer (2009) estimate a dynamic model of female employment allowing for misclassification in employment which is identified by restricting the longitudinal distribution of employment. Similarly, Feng and Hu (2013) estimate the effect of measurement error on the CPS unemployment rate, assuming that next-period true employment status does not depend on its status 9 months ago, conditional on current status and individual characteristics,

* Corresponding author.
E-mail addresses: dean.hyslop@motu.org.nz (D.R. Hyslop), wilbur.townsend@motu.org.nz (W. Townsend).

[1] See for example Bound and Krueger (1991); Pischke (1995); Kapteyn and Ypma (2007); Abowd and Stinson (2013); Hyslop and Townsend (2016a).

[2] Most studies assume their data lack error, though some have calibrated models to presumed error rates (Poterba and Summers, 1995).

[3] For example, Pischke (1995) found transitory income shocks were under-reported in surveys. The negative correlation between measurement error and transitory shocks could be produced by data in which transitory shocks are not under-reported, if observations with both negative transitory shocks *and* negative measurement error are censored and thus omitted.

http://dx.doi.org/10.1016/j.econlet.2017.03.013
0165-1765/

and that misclassification depends only on the current true status and individual characteristics.[4]

One exception is Chua and Fuller (1987), who specify a multinomial model for two reports with unbiased responses, independent errors and constant error rates over time. Chua and Fuller (1987) use the unreconciled CPS reinterview validation subsamples and estimate false positive and false negative rates of about 2%. The assumption that responses are unbiased is rejected by Feng and Hu's (2013) conclusion that the error-corrected unemployment rate is on average about 2 percentage points higher than the uncorrected rate.

This paper makes two contributions. First, we specify and estimate a model for misclassification of a binary employment variable with two reported measures. We show that the true employment rate and each measure's misreporting rates can be identified when true employment varies over at least two periods, provided that the errors are independent across the two measures, the error rates are constant over time, and the probability of reporting employment in each measure is increasing in true employment. Second, our results contribute to the literature showing that administrative data sources are not error free. We estimate that both reports contain error, with false negative and false positive rates of about 3% and 10%–16% in the survey, and about 1%–4% and less than 3% in the administrative data respectively. Although the administrative data contains less error, like Feng and Hu (2013) we find that allowing for such error increases the estimated employment rate by about 2 percentage points.

2. A model of employment misclassification error

In this section we discuss the model for measurement error using two measures of employment. Let E_t be the binary event that a person is employed in period t, E_t^c be the complementary event that the person is not employed, and let S_t and A_t be the events that the person is reported as employed in survey and administrative data. The probabilities $P(S_t)$, $P(A_t)$ and $P(S_t, A_t)$ can be estimated with sample proportions. Our aim is to estimate the true employment rate ($P(E_t)$), and the false positive ($P(S_t|E_t^c)$, $P(A_t|E_t^c)$) and false negative ($1-P(S_t|E_t)$, $1-P(A_t|E_t)$) error rates associated with each measure.

Local identification requires two sets of restrictions. First assume that the false positive and false negative rates are constant over time. For all t:

$$\begin{aligned} P(S_t|E_t) &= P(S|E), \\ P(S_t|E_t^c) &= P(S|E^c), \\ P(A_t|E_t) &= P(A|E), \\ P(A_t|E_t^c) &= P(A|E^c). \end{aligned} \tag{1}$$

This assumption implies that changes in the employment rate are the only source of year-to-year changes in reporting.

Second, assume that the false positive and false negative rates are independent across the two measures of employment.[5] Given Eq. (1),

$$\begin{aligned} P(S, A|E) &= P(S|E) \cdot P(A|E), \\ P(S, A|E^c) &= P(S|E^c) \cdot P(A|E^c). \end{aligned} \tag{2}$$

Given these two assumptions, with T periods there are $3T$ sample moments $\{P(S_t), P(A_t), P(S_t, A_t);\ t = 1, \ldots, T\}$ and $(4+T)$ parameters $\{P(S|E), P(S|E^c), P(A|E), P(A|E^c), P(E_t);\ t = 1, \ldots, T\}$. By applying the law of total probability to each sample moment, we express them in terms of the parameters:

$$\begin{aligned} P(S_t) &= P(S|E) \cdot P(E_t) + P(S|E^c) \cdot [1 - P(E_t)], \\ P(A_t) &= P(A|E) \cdot P(E_t) + P(A|E^c) \cdot [1 - P(E_t)], \\ P(S_t, A_t) &= P(S, A|E) \cdot P(E_t) + P(S, A|E^c) \cdot [1 - P(E_t)] \\ &= P(S|E) \cdot P(A|E) \cdot P(E_t) + P(S|E^c) \cdot P(A|E^c) \\ &\quad \cdot [1 - P(E_t)]. \end{aligned} \tag{3}$$

As specified, the model is locally just-identified when $T = 2$.[6] The Jacobian of the moment conditions is square, with determinant

$$\begin{aligned} \Delta = &\left[P(S|E) - P(S|E^c)\right]^2 \cdot \left[P(A|E) - P(A|E^c)\right]^2 \\ &\cdot [P(E_1) - P(E_2)]^2 \end{aligned} \tag{4}$$

and thus the model will be locally identified if both reports are related to true employment and true employment differs between periods.[7]

In the above model, the predicted moments are invariant to replacing every employment event with a non-employment event: replacing $P(S|E)$ with $P(S|E^c)$, replacing $P(E_t)$ with $1 - P(E_t)$, and so on. Global identification requires some criterion for selecting between these two locally identified estimates. We assume that the probability of reporting employment is greater when employed than not: $P(S|E) \geq P(S|E^c)$ and $P(A|E) \geq P(A|E^c)$.

3. Matched survey and administrative data

Our identification requires us to estimate the proportion of individuals recorded as employed by both measures, and thus requires our two employment measures to be matched. Our primary sample comes from the Survey of Family, Income and Employment (SoFIE), a longitudinal survey collected for 8 annual waves from 2002/2003 until 2009/2010. The administrative data is from the Employer Monthly Schedules (EMS), which each employer must file with New Zealand's Inland Revenue tax department. The EMS lists each worker an employer paid earnings to in each month. Individuals in the SoFIE sample were matched to administrative data sources using name, date of birth and gender.

In the first wave of SoFIE, respondents reported employment activity over the 12 months to the end of the previous month. This determined the calendar months for their 'annual reporting periods'. In subsequent waves, respondents were asked about activity since their previous interview, and their employment activity is allocated to their annual reporting period for that wave. We classify a person as employed in SoFIE in a wave if they report any earnings within that wave's annual reporting period. The EMS employment measure classifies an individual as employed if they received any EMS earnings within their SoFIE annual reporting period.

Our analytical sample consists of the 8-wave balanced panel of individuals aged 20–64 who completed the full SoFIE survey, who had no self-employment activity and no missing employment data or inconsistent annual reference periods over the panel, and could be matched to the EMS data. Table 1 presents summary statistics for this sample, and the excluded sample of working age individuals.

Both of our employment measures have potential error. For example, SoFIE reports will be subject to error from participants' recall of their employment spells, while the EMS data may have errors in the identifiers used to match to SoFIE participants. We

[4] These studies estimate false positive (those non-employed reporting employment) and false negative (those employed reporting non-employment) rates of about 5%–8% and 1%–2% respectively.

[5] If we allow arbitrary correlation between the errors – or, equivalently, attempt to estimate the joint error terms $P(S, A|E)$, $P(S, A|E^c)$ – the model is unidentified. With $T = 3$ periods the model has as many parameters as moment conditions, but the Jacobian of the moment conditions has determinant 0. We do, however, relax the assumption of time-constant error rates.

[6] The identification here requires only that multiple time periods, either repeated cross-sections or longitudinal data, are observed.

[7] Because we only use pooled cross-sectional moments to identify the model, we estimate average error rates across the population. However we will also estimate models that allow for the error rates to vary across observable characteristics.