I Z A Institute of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 10534

# Estimating Labor Force Joiners and Leavers Using a Heterogeneity Augmented Two-Tier Stochastic Frontier

Tirthatanmoy Das
Solomon W. Polachek

JANUARY 2017

I Z A Institute of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 10534

# Estimating Labor Force Joiners and Leavers Using a Heterogeneity Augmented Two-Tier Stochastic Frontier

**Tirthatanmoy Das**
*University of Central Florida and IZA*

**Solomon W. Polachek**
*State University of New York at Binghamton and IZA*

JANUARY 2017



**IZA – Institute of Labor Economics**

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

# ABSTRACT

## Estimating Labor Force Joiners and Leavers Using a Heterogeneity Augmented Two-Tier Stochastic Frontier*

We derive a non-standard unit root serial correlation formulation for intertemporal adjustments in the labor force participation rate. This leads to a tractable three-error component model, which in contrast to other models embeds heterogeneity into the error structure. Unlike in the typical iid three-error component two-tier stochastic frontier model, our equation's error components are independent but not identically distributed. This leads to a complex nonlinear likelihood function requiring identification through a two-step estimation procedure, which we estimate using Current Population Survey (CPS) data. By transforming the basic equation linking labor force participation to the working age population, this paper devises a new method which can be used to identify labor market joiners and leavers. The method's advantage is its parsimonious data requirements, especially alleviating the need for survey based longitudinal data.

**JEL Classification:** C23, C51, J21
**Keywords:** two-tier stochastic frontier, identification, labor force dynamics

**Corresponding author:**
Solomon W. Polachek
Department of Economics
State University of New York at Binghamton
Binghamton, NY 13902
USA

E-mail: polachek@binghamton.edu

* We thank two anonymous referees as well as Daniel Slottje for valuable comments and suggestions.

## 1. Introduction

How many people join the labor force and how many leave the labor force are important questions regarding labor market dynamics. For the US, the Bureau of Labor Statistics (BLS) computes labor market statistics based on monthly current population surveys (CPS) of about 60,000 eligible households in 2000 geographic sampling units. These monthly data give estimates of labor force participation, employment, and unemployment *levels*. Any two consecutive surveys can be used to compute *net* changes in these statistics. However, these data do not yield information on gross flows. As such, they camouflage how many individuals move in and out of the labor force, or how many move in and out of unemployment. For these latter statistics one must rely on a smaller CPS sample consisting of those households that were interviewed for at least two consecutive months. Unfortunately matched household data obtained from these consecutive follow-up surveys are subject to a number of biases. These biases led to a literature on techniques to overcome them. Essentially the approaches to fix these problems take two forms. One is to correct the data directly and the other is to forecast the data via an estimation model. Neither is perfect, which is why we propose a novel new strategy based on stochastic frontier estimation to identify labor market flows. The beauty of our approach is it alleviates the need to re-survey households, a process which would be costly for countries not currently doing so. Further our approach is fitting for this volume because it stems from techniques Robert Basmann developed to identify key structural parameters, albeit in other economic domains.

Monthly labor market statistics informs social scientists and policy makers about changes in labor force participation, employment, and unemployment, but are completely uninformative regarding how many people actually move in and out of these work-related categories. In short, these data furnish information on *net*, but not *gross* labor flows. However, gross flows can be very important especially if one seeks to understand true labor market dynamics. For example, a zero net change in the labor force is very different when hundreds of thousands simultaneously join and leave compared to when hundreds simultaneously join and leave. Knowing whether it is hundreds, or hundreds of thousands is important for policy purposes.

For the US, the BLS monthly data yields employment and unemployment levels, but one needs data on gross flows to get at how many people are shifting back and forth from one employment status to another.[1] Such US gross flow data on movements in and out of the labor force (and in and out of employment) were published from 1948 until 1952 but ceased because of discrepancies between employment and unemployment levels derived from the flows, and employment and unemployment levels derived from the monthly stock data. For this reason, the BLS had stopped publishing the flow data on a continuous basis in 1952. However, to make the flow data more useful, the BLS devised techniques to reconcile and adjust the flow and stock data. Thus beginning

[1] Other countries have similar data. For example, see Petrongolo and Pissaredes (2008) who use such data for the UK, France, and Spain.

in May 2008, the BLS extended and released the labor force status flows series to include the February 1990 to January 1994 period (http://www.bls.gov/cps/cps_flows.htm). But even these are prone to error.

The main problem with the CPS (and other) gross flow data is that the resulting net changes derived from these data are inconsistent with the stock data. They simply do not produce the same labor market changes one sees using aggregate stock data. There are several type errors. First is simple misreporting. Such response errors misclassify some respondents into the wrong labor force category, for example, working when they are not, or not working when they are. In stock data, these errors tend to cancel, but in flow data they are additive. Second, the CPS uses six of eight matched panels when computing flows. As such, rotation group errors can creep in if the unused panels differ from those panels that are actually used. Third, the CPS does not track individuals who do not complete the survey for various reasons including geographic mobility. Here again, errors result should attrition be nonrandom.

A number of studies attempt to address each of these biases. Fuller and Chua (1985) develop a model for the response error probabilities at a single point in time which requires external identification restrictions because "the number of possible parameters in the response probability matrix exceeds the number of cells available for estimation" (Fuller and Chua, p. 65). Abowd and Zellner (1985) develop an adjustment scheme to take account of missing as well as spurious labor force classifications in the gross flows CPS data. Poterba and Summers (1986) devise a technique to correct for response errors in the CPS gross changes data. Finally, Frazis et al. (2005) describe the procedure the BLS uses to rake the flows data. Taking a very different approach, Barnichon and Nekarda (2012) and Barnichon and Garda (2016) forecast inflows and outflows which can be used to predict net unemployment. Whereas these improvements help, in the end they are approximations and each yields different values. In addition, they require flow data which for many countries are hard to get.

We propose a novel alternative approach with roots back to Robert Basmann's (1985) seminal paper dealing with error structure. To do so we decompose year-to-year labor force fluctuations into changes in population growth and variations in movements in and out of the labor force. The parameterization yields a function with a three component error term. One component is pure random noise. Of the other two, one is positive and depicts the proportion entering the labor force; the other is negative and depicts the proportion leaving the labor market. Each of the latter two components is estimated via a modified version of the two-tiered frontier estimation model (Polachek and Yoon, 1987). Unlike the typical iid three-error component two-tier frontier model, the equation's error components are independent but not identically distributed. This approach emanates from Robert Basmann's (1985) seminal paper which introduces a serial correlation structure based on an intertemporal adjustment mechanism. Of course, that paper builds on his previous pioneering work on estimation and identifiability in structural equations which led to 2SLS (Basmann, R.L. (1957, 1960).

A notable feature of our method, in contrast to other stochastic frontier models, is it embeds heterogeneity directly into the error.[2] As such, our method lets the structure determine the interaction between heterogeneity and the stochastic components. This enables one to identify time-variant stochastic shocks from time-invariant heterogeneity through structural restrictions.[3] The composite error leads to a complex nonlinear likelihood function requiring identification through a two-step estimation procedure. Unlike previous labor market studies, our approach does not require flow data. We rely solely on labor market levels. As such, the method is advantageous because of its parsimonious data requirements compared to previous methods. We apply the approach to males and females of various age groups in each state within the US population. The results track the data well.

Our approach begins with an identity defining the labor force participation rate. Then we take a first-difference and add a random error along with terms depicting labor market joiners and leavers based on a group-specific stochastic process. For tractability, we approximate this characterization with a Taylor series expansion. After verifying the approximation using simulation techniques, we devise an appropriate likelihood function which we estimate using a two-step procedure. We then apply the Mackinnon-Smith approach to check for and correct potential biases introduced by the Taylor series approximation. Finally, we apply a modification of the Jondrow-Lovell-Materov-Schmidt (1982) technique to estimate joiners and leavers for each age-gender-state year group.

## 2. Derivation of an Estimable Equation

We start with a basic identity to estimate labor market flows that defines the stock of labor. To do so, we consider a particular demographic group *i* who's working-age population in period *t* is $P_{it}^*$.[4] If $\lambda_{it}^*$ is the proportion of this working age population in the labor force, then group *i*'s labor force in period *t* ($L_{it}^*$) is

$$L_{it}^* = P_{it}^* \lambda_{it}^* . \tag{1}$$

This expression is a deterministic equation. A rise in either of the right hand side variables ($P_{it}^*$ and $\lambda_{it}^*$) increases the size of the labor force. The effect of one variable on labor force size depends on the magnitude of the other. The data on working-age population $P_{it}^*$ are readily available. However the data on $\lambda_{it}^*$ are not available from any direct sources. For this reason it is often indirectly obtained by calculating the ratio of $L_{it}^*$ and $P_{it}^*$.

However, in many circumstances the true values of $L_{it}^*$ and $P_{it}^*$ are not observed due to measurement errors. In such a situation (1) cannot offer an accurate estimate of the labor force participation rate (LFPR). For this reason, we present (1) as a stochastic equation

[2] For example, in the context of stochastic frontier estimation, Greene (2005) relies on the traditional fixed and random effect linear regression framework to separate heterogeneity from inefficiency.
[3] Our approach also contrasts with studies that address heterogeneity solely through model parameters but do not explicitly specify the role of heterogeneity in the error structure (e.g. Polachek, Das, Thamma-Aprioam, 2015).
[4] A group can mean an economy, a state, a gender, or a race. Using US data we concentrate on state-gender-age-year groups.

$$L_{it} = P_{it}\lambda^*_{it}\eta_{it} \quad (2)$$

where $L_{it}$ and $P_{it}$ represent the observed values of the labor force and working age population, and $\eta_{it}$ is a stochastic component assumed to follow a log-normal distribution, arising from the measurement errors in $L^*_{it}$ and $P^*_{it}$.

The purpose of this paper is to decompose changes in the labor force participation rate arising from those joining and leaving the labor force between two consecutive time periods.[5] Expression (2) is inadequate for this purpose because it represents the relationship between levels. To transform (2) into a growth equation, divide $L_{it}$ by $L_{it-1}$, and then take the logarithm of both sides:

$$Log\left(\frac{L_{it}}{L_{it-1}}\right) = Log\left(\frac{P_{it}}{P_{it-1}}\right) + Log\left(\frac{\lambda^*_{it}}{\lambda^*_{it-1}}\right) + u_{it}\,. \quad (3)$$

The left hand side measures the growth in the labor force. The right hand side has three terms. The first measures the growth in working-age population. This growth depends on new entrants minus deaths between period *t-1* and *t*, and is related to demographic factors. The second term depicts the growth in the LFPR.[6] This growth is determined by group *i*'s net increase in participants (joiners minus leavers) between the two consecutive periods, and arises from changes in people's willingness to work. A policy intervention that alters incentive to work may affect LFPR by bringing some people into the labor market, forcing some people to leave the labor market, or a combination of both. For instance an intervention that enhances incentives to work may induce more people to participate in the labor market hence raising the LFPR. Similarly, a sudden economic downturn may force some people to leave the labor market resulting into a decline in the LFPR. We assume that the factors affecting the size of the working age population (WAP) are different from the factors that affect the LFPR. Later in the estimation section, we test whether the plausibility this assumption causes any bias in the estimates. The third term accounts for errors arising from faulty measurement of $L_{it}$ and $P_{it}$. Let this measurement error be $u_{it} = log(\eta_{it}) - log(\eta_{it-1})$. We assume $u_{it}$ is stochastic and arises from unsystematic errors in measuring the true growth in the labor force (LF) and the WAP. Also, for computational convenience, we assume this component follows a normal distribution with $E(u_{it}) = 0$ and $Var(u_{it}) = \sigma^2_u$.

Let $L^{*N}_{it-1}$ be the number of people *not* in the labor force in period $t-1$, that is $L^{*N}_{it-1} = P^*_{it-1}(1 - \lambda^*_{it-1})$. Assume some changes take place (such as, an incentive enhancing intervention or a change in economic environment) between period *t-1* and *t* that motivates $\frac{\omega_{it}}{(1+\omega_{it})}$ of those out of the labor force to enter the labor market. As such, the number of people entering the labor market is

$$LFI = \frac{\omega_{it}}{(1+\omega_{it})}L^{*N}_{it-1} = \frac{\omega_{it}}{(1+\omega_{it})}(P^*_{it-1} - L^*_{it-1}) \quad (4)$$

where $\omega_{it} \in [0, \infty)$. LFI represents the number of people joining the labor market due to factors

[5] We use annual data, but the analysis also can be done with monthly data.

[6] We use BLS's 16-year old age requirement to define working age population.

other than the change in the working-age population. Also, let the same changes or some other changes force $\frac{v_{it}}{(1+v_{it})}$ per cent of labor force to move out of the labor market. This means that the number of people exiting the labor market is

$$LFO = \frac{v_{it}}{(1+v_{it})} L^*_{it-1} . \tag{5}$$

Based on this flow between period *t−1* and *t*, the net addition to the labor force in period t is

$$LFI - LFO = \frac{\omega_{it}}{(1+\omega_{it})} (P^*_{it-1} - L^*_{it-1}) - \frac{v_{it}}{(1+v_{it})} L^*_{it-1} . \tag{6}$$

Holding $P^*_{it-1}$ constant, the size of labor force in period *t* due only to changes in the LFPR is

$$L'_{it} = L^*_{it-1} + \frac{\omega_{it}}{(1+\omega_{it})} (P^*_{it-1} - L^*_{it-1}) - \frac{v_{it}}{(1+v_{it})} L^*_{it-1} . \tag{7}$$

Dividing both sides by $P^*_{it-1}$ yields

$$\lambda^*_{it} = \lambda^*_{it-1} + \frac{\omega_{it}}{(1+\omega_{it})} (1 - \lambda^*_{it-1}) - \frac{v_{it}}{(1+v_{it})} \lambda^*_{it-1} . \tag{8}$$

Dividing both sides by $\lambda^*_{it-1}$ yields

$$\frac{\lambda^*_{it}}{\lambda^*_{it-1}} = 1 + \frac{\omega_{it}}{(1+\omega_{it})} \frac{(1 - \lambda^*_{it-1})}{\lambda^*_{it-1}} - \frac{v_{it}}{(1+v_{it})} . \tag{9}$$

Taking the log on both sides yields

$$Log\left(\frac{\lambda^*_{it}}{\lambda^*_{it-1}}\right) = Log\left(1 + \frac{\omega_{it}}{(1+\omega_{it})} \frac{(1 - \lambda^*_{it-1})}{\lambda^*_{it-1}} - \frac{v_{it}}{(1+v_{it})}\right). \tag{10}$$

For computational convenience let $\lambda^*_{it-1} = e^{-\theta^*_{it-1}}$ where $\theta^*_{it-1} \in [0, \infty)$. This transforms (10) into the following equation

$$Log\left(\frac{\lambda^*_{it}}{\lambda^*_{it-1}}\right) = Log\left(1 + \frac{\omega_{it}}{(1+\omega_{it})} \left(e^{\theta^*_{it-1}} - 1\right) - \frac{v_{it}}{(1+v_{it})}\right) . \tag{11}$$

We approximate (11) with a first degree Taylor series expansion around the steady state $\theta^*_{it-1} = \theta_{0i}$, $\omega_{it} = 0$, and $v_{it} = 0$. A steady state arises when there are no joiners and no leavers, ($\omega_{it} = 0$, and $v_{it} = 0$), hence a constant LFPR. Non-zero values of $\omega_{it}$, and $v_{it}$ may also constitute steady states. But generally $\omega_{it}$, and $v_{it}$ are small enough to be approximated by zero values. The only alternative is a big $\omega_{it}$, which must be matched with a big $v_{it}$ to maintain the steady state LFPR. This means that there are many joiners and leavers simultaneously, an unlikely scenario.

The approximation yields

$$Log\left(\frac{\lambda_{it}^*}{\lambda_{it-1}^*}\right) = \omega_{it}\left(e^{\theta_{0i}} - 1\right) - v_{it}\ . \quad (12)$$

Substituting the expression in (3) yields

$$Log\left(\frac{L_{it}}{L_{it-1}}\right) = \ Log\left(\frac{P_{it}}{P_{it-1}}\right) + \ \omega_{it}\left(e^{\theta_{0i}} - 1\right) - v_{it} + \ u_{it} \quad (13)$$

For notational simplicity, let $y_{it}=Log(L_{it}/L_{it}-1)$ and $x_{it}=Log(P_{it}/P_{it-1})$so that

$$y_{it} = \ x_{it} + \ \omega_{it}\left(e^{\theta_{0i}} - 1\right) - v_{it} + \ u_{it}\ . \quad (14)$$

We estimate (14), but before doing so we evaluate the validity of the Taylor series approximation (12).

## 3. Evaluating the Approximation

Equation (12) is a first order Taylor series linear approximation of (11). Whereas (12) yields small truncation errors when $\omega$, and $v$ are near zero, it is useful to evaluate (12)'s empirical validity for values of $\omega$, and $v$ typically observed in the labor market. Gauging this entails randomly generating $\omega_{it}$, $v_{it}$, and the steady state LFPR ($e^{-\theta_{0i}}$), which we use in (11) and (12) to compute respective $Log\left(\frac{\lambda_{it}^*}{\lambda_{it-1}^*}\right)$ levels. We then compare the $Log\left(\frac{\lambda_{it}^*}{\lambda_{it-1}^*}\right)$ values obtained from these. First, we test whether $Log\left(\frac{\lambda_{it}^*}{\lambda_{it-1}^*}\right)$ values generated from (12) is a linear approximation of those generated from (11). Second, we test whether the $Log\left(\frac{\lambda_{it}^*}{\lambda_{it-1}^*}\right)$ values generated by (12) are similarly distributed to the values generated by (11). The latter is a crucial requirement in estimating labor market joiners and leavers because, as will be explained later, both joiners and leavers are identified from distributional assumptions. Thus the $Log\left(\frac{\lambda_{it}^*}{\lambda_{it-1}^*}\right)$ distribution should be the same in each equation.

Based on current literature,[7] the maximum proportion of the population joining the labor force is around 8%. The maximum proportion leaving the labor force is about 7%. These data imply an expected value of $\omega$ ($\mu_\omega$) not exceeding 0.08 and an expected value of $v$ ($\mu_v$) not exceeding 0.07. Accordingly, for our simulations we select combinations of $\mu_\omega$ and $\mu_v$ as follows: $\mu_\omega = 0.02, \mu_v = 0.01$; $\mu_\omega = 0.04, \mu_v = 0.02$; $\mu_\omega = 0.05, \mu_v = 0.03$; $\mu_\omega = 0.06, \mu_v = 0.05$; and $\mu_\omega = 0.08, \mu_v =$

[7] These include Barnichon and Garda (2016), Calimita and Sabol (2015), Shimer (2012), as well as BLS data (https://www.bls.gov/web/empsit/cps_flows_recent.pdf).

0.07. Similarly, to illustrate the results do not change, we vary the sample sizes from 100, to 500, 1000, 5000, 10000, and finally to 15000. Figures 1a-1e present the simulations.

### 3.1 The Linear Approximation

The first order Taylor series approximation (12) depicts a linearized version of (11).[8] If this approximation is close, then one must find that the Pearson correlation coefficient (presented as $\rho$ in the figures) between the values generated by (11) and (12) is close to 1. Our simulation exercises confirm this correlation. Figures 1a-1e show that all correlation coefficients are above 0.99 (ranging between 0.991-0.999). The magnitude of the coefficient rises slightly as $\mu_\omega$ and $\mu_v$ decline, but the qualitative results remain robust across different values of $\mu_\omega$ and $\mu_v$, and different sample sizes.

### 3.2 Comparing Distributions

We identify labor force joiners and leavers based on the distributions we assume underlie $\omega_{it}(e^{\theta_{0i}}-1)$ and $v_{it}$ in (11). Given that we estimate (12), we must show that the inherent distributions of both (11) and (12) are the same. To do so we employ the Kolmogorov-Smirnov distance measure to compare the $Log\left(\frac{\lambda_{it}^*}{\lambda_{it-1}^*}\right)$ values simulated from each equation. As illustrated in Figures 1a-1e, we find the Kolmogorov-Smirnov distance statistics range from 0.008 to 0.05. The distances are bigger when number of observations are small (e.g. N=100), but show a declining trend as N grows. The result is also robust across different values of $\mu_\omega$ and $\mu_v$.

Based on both the Pearson correlation and the Kolmogorov-Smirnov test we conclude the Taylor approximation does not alter the magnitude of the distribution of $Log\left(\frac{\lambda_{it}^*}{\lambda_{it-1}^*}\right)$. Thus we go on to consider the empirical specification.[9]

## 4. Empirical Specification

Equation (14) has two known variables ($y_{it}$,$x_{it}$), three unobserved components ($\omega_{it}$,$v_{it}$,$u_{it}$), and a group-specific parameter $\theta_{0i}$. For the sake of exposition, we rewrite (14) as

$$y_{it} = x_{it} + \omega_{it}^* - v_{it} + u_{it} \qquad (15)$$

where $\omega_{it}^* = (e^{\theta_{i0}} - 1)\omega_{it}$. Based on the previously described labor market identities, $\omega_{it}^*$ and $v_{it}$ take on only positive values. As such, $\omega_{it}^* \in [0, \infty)$ and $v_{it} \in [0, \infty)$. The $u_{it}$ term takes any value on the real line so that $u_{it} \in (-\infty, \infty)$. The parameters $\theta_{0i} \in [0, \infty)$ are specific to each group

[8] See Appendix I for the derivation.

[9] The eventual estimation of $\mu_\omega, \mu_v,$ and $\sigma_u$ can be affected even though the magnitudes and distribution of $Log\left(\frac{\lambda_{it}^*}{\lambda_{it-1}^*}\right)$ is unaffected by our Taylor approximation. Later in the empirical section we test for this by applying the Mackinnon-Smith bias correction technique.

$i$. As such, $\theta_{0i}$ represents a group-specific effect, in our case gender, age, and state.

We assume measurement error $u_{it}$ follows an i.i.d normal distribution with zero mean and standard deviation $\sigma_u$. Given that joiner and leaver deviations from the steady state are most likely skewed, we assume $\omega_{it}$, and $v_{it}$ follow i.i.d. exponential distributions with mean $\mu_\omega$ and $\mu_v$. As such, $\omega^*_{it}$ also follows an exponential distribution with mean $\mu_{\omega i} = (e^{\theta_{0i}} - 1)\mu_\omega$. However, because of heterogeneity, the group-specific term, $\omega^*_{it}$ is no longer identically distributed, though it remains statistically independent. To estimate $\widehat{\omega}_{it}$ and $\hat{v}_{it}$ requires values for $\mu_\omega$, $\mu_v$, and $\sigma_u$.

Because one residual is positive ($\omega^*_{it}$) and one is negative ($v_{it}$), specification (15) is known as a two-tiered frontier, but it differs from (Polachek and Yoon, 1987) because $\omega^*_{it}$ contains a group-specific parameter, and thus, as just mentioned, is not identically distributed. This makes estimation more complicated.

Specify the composite error $\varepsilon_{it}$ as

$$\epsilon_{it} = \omega^*_{it} - v_{it} + u_{it} \tag{16}$$

where the distributions of $\omega^*_{it}$, $v_{it}$, and $u_{it}$ are as defined above. The composite error $f(\varepsilon_{it})$ has the density[10]

$$f(\epsilon_{it}) = \frac{G}{a}\left[\Phi(-c) + e^{\frac{a(a-2bc)}{2b^2}}\Phi\left(\frac{bc-a}{b}\right)\right] \tag{17}$$

where

$$G = \frac{1}{\mu_v \mu_{\omega i}} e^{\frac{\sigma_u^2}{2\mu_v^2}} e^{\frac{\epsilon_{it}}{\mu_v}}; a = \left(\frac{1}{\mu_v} + \frac{1}{\mu_{\omega i}}\right); b = \frac{1}{\sigma_u}; c = \left(\frac{\epsilon_{it}}{\sigma_u} + \frac{\sigma_u}{\mu_v}\right)$$

and $\Phi(.)$ represents the CDF of a standard normal distribution. The likelihood function is

$$L = \prod_i \prod_t f(\mu_{\omega i}, \mu_v, \sigma_u | y_{it}, x_{it}) \ . \tag{18}$$

It is computationally impractical to compute all $\mu_{\omega i}$ and other parameters simultaneously, especially if there are a large number of groups ($i$). Thus, we adopt a novel approach.

First, we devise a panel data technique to transform (12) enabling us to identify $\theta_{0i}$. Second, we rewrite the likelihood function (18) to incorporate the computed $\theta_{0i}$ estimates. Third, we estimate this new likelihood function to obtain $\hat{\mu}_\omega, \hat{\mu}_v,$ and $\hat{\sigma}_u$. We then use the Mackinnon-Smith (1998) bias correction technique to overcome inherent potential biases in these parameters coming about because we estimate a Taylor approximation of (11). Utilizing these estimates, we adopt the Jondrow-Lovell- Materov-Schmidt (1982) technique to estimate $\widehat{\omega}^*{}_{it}$ and $\hat{v}_{it}$. The estimates $\widehat{\omega}^*{}_{it}$ allow us to compute $\widehat{\omega}_{it}$. Finally, based on these, we compute the proportion of joiners and leaves.

[10] Appendix II provides the detailed derivation of $f(\varepsilon_{it})$.

Begin by defining the deviation of LFPR from its steady state $\theta_{0i}$

$$\frac{e^{-\theta_{it}^*}}{e^{-\theta_{0i}}} = 1 + \left(e^{\theta_{0i}} - 1\right)\frac{\omega_{it}'}{(1+\omega_{it}')} - \frac{v_{it}'}{(1+v_{it}')} \tag{19}$$

where $\omega'_{it}$, and $v'_{it}$ are the shocks that move the LFPR from its steady state value $\theta_{0i}$. Taking the logarithm of both sides and then approximating yields

$$Log\left(e^{-\theta_{it}^*}\right) = -\theta_{0i} + \left(e^{\theta_{0i}} - 1\right)\omega_{it}' - v_{it}' \,. \tag{20}$$

Allowing for measurement errors in the LFPR estimate, one can rewrite (20) as

$$Log\left(e^{-\theta_{it}}\right) = -\theta_{0i} + \left(e^{\theta_{0i}} - 1\right)\omega_{it}' - v_{it}' + \pi_{it} \tag{21}$$

where $\pi_{it}$ is the random error which is assumed to follow a distribution with $E(\pi_{it}) = 0$. Next, we rewrite (21) as

$$Log\left(e^{-\theta_{it}}\right) = -\theta_{0i} + \left(e^{\theta_{0i}} - 1\right)\mu_\omega - \mu_v + \left(e^{\theta_{0i}} - 1\right)(\omega_{it}' - \mu_\omega) - (v_{it}' - \mu_v) + \pi_{it}$$
$$Log\left(e^{-\theta_{it}}\right) = \beta_i + \zeta_{it} \tag{22}$$

where

$\beta_i = -\theta_{0i} + \left(e^{\theta_{0i}} - 1\right)\mu_\omega - \mu_v$ and $\zeta_{it} = \left(e^{\theta_{0i}} - 1\right)(\omega_{it}' - \mu_\omega) - (v_{it}' - \mu_v) + \pi_{it}$ .

In this form $E(\zeta_{it}) = 0$, which means $\widehat{\beta_\iota}$ can be consistently estimated. To identify $\theta_{0i}$, we utilize a second restriction from (14). It suggests that

$$Log\left(\frac{L_{it}}{L_{it-1}}\right) = \left(e^{\theta_{0i}} - 1\right)\mu_\omega - \mu_v + Log\left(\frac{P_{it}}{P_{it-1}}\right) + \left(e^{\theta_{0i}} - 1\right)(\omega_{it} - \mu_\omega) - (v_{it} - \mu_v) + \eta_{it}$$
$$Log\left(\frac{L_{it}}{L_{it-1}}\right) = \gamma_i + Log\left(\frac{P_{it}}{P_{it-1}}\right) + \delta_{it} \tag{23}$$

where

$\gamma_i = \left(e^{\theta_{0i}} - 1\right)\mu_\omega - \mu_v$ and $\delta_{it} = \left(e^{\theta_{0i}} - 1\right)(\omega_{it} - \mu_\omega) - (v_{it} - \mu_v) + \eta_{it}$ with $E(\delta_{it}) = 0$

This allows one to estimate $\widehat{\gamma_\iota}$. Bringing the estimates of $\widehat{\beta_\iota}$ and $\widehat{\gamma_\iota}$ together yields

$$\hat{\theta}_{0i} = \hat{\gamma}_i - \hat{\beta}_i \,. \tag{24}$$

Identification of $\theta_{0i}$ is crucial because it allows the identification of $\mu_\omega$. Substituting the value of $\theta_{0i}$ implies that $\mu_{\omega i} = (e^{\widehat{\theta_{0\iota}}} - 1)\mu_\omega$. This simplifies the maximum likelihood estimation since now we only have to estimate three parameters $\mu_\omega$, $\mu_v$, and $\sigma_u$. The heterogeneity adjusted modified density and likelihood functions are

$$f(\epsilon_{it}) = \frac{G}{a}\left[\Phi(-c) + e^{\frac{a(a-2bc)}{2b^2}}\Phi\left(\frac{bc-a}{b}\right)\right] \tag{25}$$

where $G = \frac{1}{\mu_v\left(e^{\hat{\theta}_{0i}-1}\right)\mu_\omega} e^{\frac{\sigma_u^2}{2\mu_v^2}} e^{\frac{\epsilon_{it}}{\mu_v}}$; $a = \left(\frac{1}{\mu_v} + \frac{1}{\left(e^{\hat{\theta}_{0i}-1}\right)\mu_\omega}\right)$; $b = \frac{1}{\sigma_u}$; $c = \left(\frac{\epsilon_{it}}{\sigma_u} + \frac{\sigma_u}{\mu_v}\right)$, and

$\Phi(.)$ represents the CDF of a standard normal distribution, and

$$L = \prod_i \prod_t f\left(\mu_\omega, \mu_v, \sigma_u \middle| y_{it}, x_{it}, \hat{\theta}_{0i}\right) . \tag{26}$$

## 5. Correction of the Potential Bias

Specification (12) is a Taylor Series approximation of (11). Although above we showed the truncation error to be negligible and not likely to alter the statistical properties of the composite error, the overall impact on the parameter estimates cannot be predicted a priori. Thus we adopt the MacKinnon and Smith (1998) simulation-based bias correction technique.

Suppose the estimation bias $b(\Omega)$ is a linear function of the parameters, such that

$$b(\Omega) = \alpha' + b'\Omega \tag{27}$$

where $\Omega = (\mu_\omega; \mu_v; \sigma_u)$ represents the parameters, and $\alpha'$ and $b'$ depict their respective intercepts and slopes. Correcting the bias requires specific estimates of $\alpha'$ and $b'$ which are difficult to obtain via analytical methods. For this reason, we apply a simulation-based method.

To do so, we draw random samples from the distributions of the three error terms (11). Because these random draws require specific parameter values, we initially choose the uncorrected parameter estimates. Based on this, we draw 500 samples of random numbers from the ω, $v$, and $u$ distributions based on $\mu_\omega, \mu_v,$ and $\sigma_u$ values obtained by estimating (18). We then use (11) and $\hat{\theta}_{0i}$ to generate the data for the composite error $(\omega_{it}^* - v_{it} + u_{it})$. Next we combine the simulated composite error with the population growth data and obtain the growth in labor force (the outcome variable). We replicate each of the 500 sets of random draws. This process yields 500 different samples of simulated data each with approximately 15,000 observations representing the gender, age, state and time data points. From these, we then estimate parameter values for each sample, thus yielding 500 sets of estimates. Based on the averages of these we then compute the biases by the following formula

$$b(\hat{\Omega}) = E(\Omega_s) - \hat{\Omega} = \bar{\Omega}_s - \hat{\Omega} \tag{28}$$

where $\Omega_s \equiv (\mu_\omega^s; \mu_v^s; \sigma_u^s)$ and $s = [1, 2, 3, \ldots, 500]$; $\bar{\Omega}_s = \frac{1}{500}\sum_{s=1}^{500}\Omega_s$.

Obtaining estimates of $b(\widehat{\Omega})$ alone does not ensure $\alpha'$ and $b'$ are identified because there are two unknowns with one equation. Identification requires another equation with at least another set of $b(\Omega)$ estimates. This means the simulation exercise must be replicated with a different set of $\Omega$. As suggested by MacKinnon and Smith (1998), we take constant bias corrected (CBC) estimates of $\widehat{\Omega}$ to construct a second set of estimates of the biases. These CBC estimates are defined as

$$\widetilde{\Omega} = 2\widehat{\Omega} - \overline{\Omega} \tag{29}$$

where $\widetilde{\Omega}$ depicts the CBC estimates of $\widehat{\Omega}$. We use these estimates to replicate the same simulation procedure again. This yields another set of bias estimates $b(\widetilde{\Omega})$. Utilizing $b(\widehat{\Omega})$, $\widehat{\Omega}$, $b(\widetilde{\Omega})$, and $\widetilde{\Omega}$ we compute $\alpha'$ and $b'$ from the following

$$\alpha' = b(\widehat{\Omega}) - \frac{b(\widehat{\Omega}) - b(\widetilde{\Omega})}{\widehat{\Omega} - \widetilde{\Omega}}\widehat{\Omega}; \tag{30}$$

$$b' = \frac{b(\widehat{\Omega}) - b(\widetilde{\Omega})}{\widehat{\Omega} - \widetilde{\Omega}} . \tag{31}$$

Based on the estimated intercepts, slopes, and the other estimates above, we compute the linear bias corrected estimates using the following

$$\breve{\Omega} = \widehat{\Omega} - \left(I + \hat{B}'\right)^{-1}\hat{b} \tag{32}$$

where $\hat{b} \equiv b(\widehat{\Omega})$ is a $(3 \times 1)$ vector; $\hat{B} \equiv B(\widehat{\Omega})$ is a $(3 \times 3)$ matrix of the derivatives of the bias function and $I$ is an identity matrix. The associated covariance matrix of $\breve{\Omega}$ is

$$V(\breve{\Omega}) = \left(I + \hat{B}'\right)^{-1} V(\widehat{\Omega}) \left(\left(I + \hat{B}'\right)'\right)^{-1}. \tag{33}$$

The bias correction yields a smaller variance of $\breve{\Omega}$ when the slope of the bias function is positive or the variance of $\widehat{\Omega}$ is small. As will be shown in the empirical section, our estimates indicate that the variances of $\widehat{\Omega}$ are indeed small. Because of this small magnitude the standard errors of the bias corrected estimates remain small even though the slopes of the bias functions are negative.

## 6. Estimating Joiners and Leavers

Jondrow, Lovell, Materov, and Schmidt (1982) formulate a method to obtain the conditional mean for inefficiency measures in a one-tier stochastic frontier model. Their formulation has been generalized for a two-tiered model.[11] We apply this technique to estimate $\widehat{\omega_{it}}$ and $\widehat{v_{it}}$ based on our estimate of the composite error $\hat{\varepsilon}_{it}$, as well as $\hat{\mu}_{\omega}$, $\hat{\mu}_{v}$, and $\hat{\sigma}_{\mu}$ obtained from maximizing (26). These expected values are

$$E(\omega_{it}^{*}|\hat{\epsilon}_{it}) = \int_0^{\infty} \omega_{it}^{*} f(\omega_{it}^{*}|\hat{\epsilon}_{it}) d\omega_{it}^{*}$$
$$E(v_{it}|\hat{\epsilon}_{it}) = \int_0^{\infty} v_{it} f(v_{it}|\hat{\epsilon}_{it}) dv_{it} \tag{34}$$

where for notational simplicity we drop the subscripts. We provide the derivation of $f(\omega^{*}{}_{it}|\widehat{\varepsilon_{it}})$ and $f(v_{it}|\widehat{\varepsilon_{it}})$ are provided in Appendix III. We compute $\widehat{\omega}_{it}$ from $\widehat{\omega}_{it} = \frac{\widehat{\omega}_{it}^{*}}{(e^{\widehat{\theta}_{0i}-1})}$. Finally, based on these, we compute the proportion of joiners and leavers as $\widehat{\omega}_{it}/(1+\widehat{\omega}_{it})$ and the proportion of leavers as $\hat{v}_{it}/(1+\hat{v}_{it})$.

## 7. The Data

The data used to estimate (34) are extracted from the Annual Social and Economic Supplement (ASEC) obtained from IPUMS-CPS.[12] ASEC is a CPS survey of approximately 60,000 households surveyed every month since 1962. We use civilian non-institutional population data from 1977 to 2015. The extracted information includes each respondent's labor force status, age, gender, state of residence, sampling weights, and a civilian non-institutional population identifier.[13]

We first classify respondents into four age (16-30, 30-42, 42-60, 60 and above), gender and state groups. This yields 408 ($51 \times 2 \times 4 = 408$) groups. We then use the sampling weight, labor force status, and age to compute the labor force ($L$), and the working-age population ($P$) for each group in every year. This yields a dataset with 15912 observations ($408 \times 39 = 15912$). The variables for the regressions are then constructed as follow:

$$y_{it} = Log(\frac{L_{it}}{L_{it-1}})$$

$$x_{it} = Log(\frac{P_{it}}{P_{it-1}})$$

where $i$ represents the (state-gender-ag) group and $t$ represents the year.

---

[11] A derivation of this process is provided in Appendix III. An alternative derivation is also available in Kumbhakar and Parmeter (2009)

[12] IPUMS-CPS is an integrated set of data collected over 50 years (1962 onward) of the Current Population Survey (CPS).

[13] For year 2014, ASEC was redesigned. We only keep the respondents present in original 5/8 sample to keep the data comparable across years.

## 8. Estimation

We utilize the ASEC data on $y_{it} = Log(\frac{L_{it}}{L_{it-1}})$ and $x_{it} = Log(\frac{P_{it}}{P_{it-1}})$ in a multistep approach to predict the annual number of incoming labor force joiners and outgoing labor force leavers for specific gender and age segments of the population. As explained, we first utilize the data in a panel setting to estimate group i's steady state LFPR ($e^{-\hat{\theta}_{0i}}$) from (24). Next, we incorporate estimates of $\theta_{0i}$ into likelihood function (26) to estimate the parameters we use to apply the Mackinnon-Smith bias correction technique to yield bias corrected estimates. From this we obtain estimates of joiners and leavers by utilizing the modified Jondrow, Lovell, Materov, and Schmidt (1982) formulation. Augmenting annual labor force participation rates by the predicted number of joiners and leavers year by year yields predicted annual labor force participation rates which we track to BLS data.

### 8.1 Steady State Labor Force Participation

Steady state labor force participation rates $e^{-\hat{\theta}_{0i}}$ are obtained from (24) based on estimating (22) and (23). We obtain one coefficient for each group $i$, in our case 408 coefficients reflecting the four age, two gender, and 51 states categories. Table 1 summarizes these steady state estimates. A number of important results emerge. First, on average 64.4 percent of the working age population participate in labor market in the steady state. Second, men's steady state labor force participation rate (73.6%) exceeds women's (55.9%). Third, the steady state labor force participation rate for the 30-60 year olds is higher than steady state labor force participation rate for 16-30 year olds. The participation rate of the oldest age group (60 and above) is the lowest among all.

These values make intuitive sense, because typically labor force participation rises and then falls over the lifecycle, and men have higher participation rates than women. Similarly, Kiefer and Neumann (2006) also find that men's steady state labor force participation exceeds women's. Kiefer and Neumann estimate for men (22-59 years) is 92% which is close to our estimates of 90% percent (for men 30-60 years of age). However, our estimate for women is 70% while theirs is 30%, but we believe our estimates to be more realistic given current women's participation.

### 8.2 Maximum Likelihood Estimation

Incorporating $\hat{\theta}_{0i}$ into (18) yields likelihood function (26), thus providing population-wide $\hat{\mu}_{\omega}, \hat{\mu}_{v}$, and $\hat{\sigma}_{u}$. However, as reported in table 2 (left-most column) this estimation yields biased results namely 0.041, 0.034, and 0.035 respectively.[14] To employ the MacKinnon-Smith constant

[14] This estimate imposes the structural constraint that the $x_{it}$coefficient equals one, meaning that the rate of change in population growth equals the rate of change in labor force participation growth, ceteris paribus. Relaxing this constraint yields a 1.04 $x_{it}$ coefficient, and virtually no change in $\mu_{\omega}$, $\mu_{v}, \sigma_{\mu}$ which remains at 0.040, 0.034, and 0.035 respectively. This corroboration substantiates our assumption that factors affecting the size of the working age

bias correction approach, we utilize these estimates to draw random values of $\omega_{it}$, $v_{it}$, and $u_{it}$ for each of the 15,504 observations, assuming two exponential and a normal distribution. From these we simulate $y_{it}$ and estimate (26) to obtain $\mu_{\omega}$, $\mu_{v}$, and $\sigma_{u}$. We replicate this simulation and estimation 500 times, after which we pick between 100 and 500 from which we take average $\mu_{\omega}$, $\mu_{v}$, and $\sigma_{u}$ values (columns 2-6, top of table 2). Clearly 500 replications are sufficient given that these average values are virtually identical independent of the number of simulations. The difference between these values and the original estimates (0.007, 0.0031, -.0008) is an initial indication of the bias $b(\widehat{\Omega})$. Following MacKinnon-Smith's constant bias correction approach, we add these biases to the original estimates to obtain the CBC (0.0476, 0.037, and 0.034 in the lower left-hand of table 2), which we use to replicate the simulation a second time. The results for the new estimates are given in columns 2-6 (bottom of table). These are used in conjunction with the CBC estimates to obtain another set of biases $b(\widetilde{\Omega})$ which along with $b(\widehat{\Omega})$ are used to solve (30) and (31), thus yielding the linear bias corrected and the covariance matrix based on (32) and (33). These linear bias corrected estimates $\breve{\Omega}$ along with their standard errors are given in table 3 (bottom two rows). Also reported in the original estimates (row 1) and the slope of the bias correction function (row 3). The results suggest that the biases in the estimates of $\mu_{v}$, and $\sigma_{u}$ are relatively small (12 percent and -3 percent respectively), whereas the bias in the estimate of $\mu_{\omega}$ is more significant (approximately 22 percent). The slope of the bias function is negative, which increases the variance of the bias corrected estimates. As such, the standard errors of the bias corrected estimates are slightly larger than that of the uncorrected ones. Nevertheless, in terms of the magnitude these standard errors remain fairly small giving a strong amount of confidence to our estimates.

### 8.3 Estimating Labor Force Joiners and Leavers

Applying the modified Jondrow, Lovell, Meterov, and Schmidt (1982) formula (34) along with the bias corrected estimates $\hat{\mu}_{\omega}$, $\hat{\mu}_{v}$, and $\hat{\sigma}_{u}$ enables us to compute the joiners (as proportion of persons not in labor force) and leavers (as proportion of labor force). Table 4 presents summary statistics by gender and age group. On average 4.6 percent of those outside the labor market in the previous year join the labor market in current year. Also, approximately 3.4 percent of the people in labor force in previous year leave the labor market before the current year. Typically males have a higher rate of joining than females except during the post child bearing years when women reenter the labor market after having left. As expected the proportion of joiners on average exceeds the proportions of leavers.

Table 4 also indicates women are less likely to join the labor market (4.7 percent for men vs. 4.5 percent for women) and more likely to leave (3.7 percent vs. 3.1 percent for men) than men. In addition, older individuals (60 years and above) are less likely to join (2.5 percent) and more likely to leave (6.2 percent). Those in 30-42 and 42-60 age groups are very similar in labor market entry

---

population $x_{it}$ are different than the factors affecting the labor force participation rate, otherwise the coefficient would have changed because of endogeneity biases in that $x_{it}$ would be correlated with the composite error.

and exit behavior. The youngest group 16-30 has the same likelihood of entering the labor market as the 30-60 year olds. However, 16-30 year olds are slightly more likely to leave the labor market than 30-60 year olds.

## 9. Comparison with actual data

It is instructive to compare our estimates with actual data. Our sample contains 15,504[15] observations (408 age-gender-sate groups over 38 years). For each observation we can compute the labor force participation rate in either of two ways. First, directly from data: The CPS contains annual information on the working age population and the number in the labor force. The ratio is the labor force participation rate. Second, the labor force participation rate in any one year can be computed as the labor force participation rate in the previous year plus the difference in the proportions between joiners and leavers in the interim. Given the proportion of joiners and leavers just computed, we can augment any one year's labor force participation rate to obtain estimates of the labor force participation in the future. We employ both methods and compared the results. Obviously, the closer the latter is to the former, the better our estimates.

Table 5 (column 1) presents the overall as well as group specific averages of labor force participation rates observed in the data (actual LFPR). Column 2 contains our predicted labor force participation rates.[16] Comparing the two columns indicates how accurately the two track each other. This is true at the aggregate level and across gender and age groups. The correlation coefficient between actual and predicted LFPR is 0.997. The distribution of actual and predicted LFPR are virtually indistinguishable (Figure 2). The trends in actual and predicted LFPRs (Figure 3) also confirm this close resemblance, but with one pitfall. The estimates are slightly lower than actual, particularly for men. One reason is aggregation. Our approach uses $\hat{\mu}_{\omega}$, $\hat{\mu}_{v}$, and $\hat{\sigma}_{u}$ based on population-wide estimates of eq (14) using state-year-gender-age observations. However, these estimates can be imprecise if labor force participation functions are heterogeneous across specific population segments, for example, gender and age. To account for this possibility, we re-estimate (14) separately by four age groups and gender. Predictions based on these estimates are given in Figure 4. As can be seen, they track the data more accurately. Of course, the lesson is the importance of heterogeneity. Whereas we initially used group-specific steady state labor force participation rates to account for heterogeneity, we now compute age and gender specific estimates to discern ancillary population differences.[17]

To assess the predictive power of our model, we construct out-of-sample predictions. To do so, we re-estimate the model using 1977-2012 (rather than 1977-2015) data, and predict LFPRs for 2013-

[15] The number of observations in our sample is 408 fewer than in the initial data because of first-differencing.

[16] The predicted LFPR is calculated by adjusting the lag value of actual LFPR by our estimates of steady state LFPR, joiners, and leavers. The formula is given in the (9).

[17] Polachek, Das and Apiroam-Thamma (2015) show the importance of individual-specific heterogeneity by illustrating inherent biases in estimating population-wide persistence of permanent and transitory shocks. Studies that examine persistent and transitory effects within the context of frontier estimation models are Ahn and Sickles (2000), Colombi (2010), Tsionas and Kumbhakar (2014), and Filippini and Greene (2016).

2015. This computation first entails estimating bias corrected $\hat{\mu}_{\omega}, \hat{\mu}_{v}$ and $\hat{\sigma}_{u}$ from the 1977-2012 data. Then, second, it involves combining these estimates with the LFPR growth data (i.e. the composite error $\epsilon_{it} = y_{it} - x_{it}$) to obtain the estimates of joiners and leavers from 1978-2015. We use these to construct both in-sample and out-of-sample predictions of the 2012-2015 LFPRs (Figure 5). Both the in-sample and out-of-sample LFPR predictions track actual LFPRs well. Further, they closely match the predicted LFPRs in Figure 3. These similarities, especially for 2013, 2014 and 2015, add further credibility to our approach.

Using the same data one can plot the net year-to-year changes in the labor force. Figure 6a-6b plot these for both men and women. Our predicted labor force measures clearly track the net changes in the actual labor force. The similarity is noteworthy because net changes based on CPS stock data often do not conform to the net changes computed from the gross flow data (Frazis et. al., 2005). Thus our approach is noteworthy as a good alternative to using gross flow data.

## 10. Conclusion

Nowadays no one can get an economics paper published without seriously considering his/her identification strategy. Clearly Robert Basmann was an original pioneer in this area. His early classic articles (1957, 1960) illustrated how to identify coefficients in multi-equation systems. This discovery later paved the way to consider identification designs in other regimes. In Robert's honor this paper deals with identification. The question we address is how can one identify the proportion of people annually entering and exiting the labor force when one only has yearly data on population size and a measure of the number those in the labor force. First differencing the labor force data only yields net changes, but gross changes, that is the number of workers entering the number of workers exiting, are often better tools to gauge the economy. For the US, the BLS makes available matched panel data, but these are subject to large errors. Now there emerged a small literature on how to make corrections, but these corrections are imperfect. Further, even if accurate, many countries simply do not and cannot follow a given panel even for short time periods. But identifying both the number of labor force joiners and leavers is important. With strong roots to the Robert Basmann's pioneering work, we devise a way to identify these labor market flows from purely cross-sectional data. The approach relies on a time-varying unobserved labor market variable which follows a nonstandard unit root process. This formulation is in the spirit of Robert Basmann (1985) who identifies a covariance structure based on a serially correlated error that emerges from consumer theory. Whereas his approach leads to cross-equation restrictions, ours involves a panel data methodology in addition to distributional assumptions. Finally, we adopt a several step estimation procedure from which we are able to accurately compute gross annual flows of joiners and leavers to and from the labor market. From these we build up our estimates to compute labor force participation rates for various demographic groups as well as the changes over time. We find our estimates track the data particularly well.

Whereas we concentrate on labor force participation, our approach can equally be used to estimate transitions to and from employment to track changes in unemployment rates. The advantage of our approach is it utilizes readily available cross-sectional data, thereby alleviating the need for more costly and difficult to obtain panel data which are typically more prone to errors.

**Table 1: Average steady state labor force participation rate estimates ($e^{-\hat{\theta}_{0i}}$), by age and gender 1978-2015**

| Age group | Men | Women | Total |
|---|---|---|---|
| 16-30 | 0.744 | 0.636 | 0.690 |
| 30-42 | 0.932 | 0.722 | 0.825 |
| 42-60 | 0.872 | 0.681 | 0.774 |
| 60 and above | 0.299 | 0.172 | 0.227 |
| | | | |
| Total | 0.736 | 0.559 | 0.644 |

Source: Based on our estimates obtained from IPUMS-CPS ASEC survey data (1977-2015).
Notes: The numbers represent the steady state proportion of working age population in the respective groups.

**Table 2: Maximum likelihood simulation based estimates of $\mu_\omega, \mu_v$, and $\sigma_u$**

| | Number of replication | | | | |
|---|---|---|---|---|---|
| Simulations | N=100 | N=200 | N=300 | N=400 | N=500 |
| Based on original estimates | | | | | |
| $\widehat{\Omega} \equiv [\hat{\mu}_\omega = 0.041, \hat{\mu}_v = 0.034, \hat{\sigma}_u = 0.035]$ | | | | | |
| Average of estimates from simulated data | | | | | |
| $\mu_\omega$ | 0.0336 | 0.0336 | 0.0335 | 0.0335 | 0.0335 |
| $\mu_v$ | 0.0307 | 0.0307 | 0.0307 | 0.0307 | 0.0307 |
| $\sigma_u$ | 0.0359 | 0.0359 | 0.0359 | 0.0360 | 0.0359 |
| | | | | | |
| Based on bias-corrected estimates (CBC) | | | | | |
| $\widetilde{\Omega} \equiv [\tilde{\mu}_\omega = 0.048, \tilde{\mu}_v = 0.037, \tilde{\sigma}_u = 0.034]$ | | | | | |
| Average of estimates from simulated data | | | | | |
| $\mu_\omega$ | 0.0385 | 0.0385 | 0.0384 | 0.0384 | 0.0385 |
| $\mu_v$ | 0.0333 | 0.0333 | 0.0333 | 0.0333 | 0.0332 |
| $\sigma_u$ | 0.0353 | 0.0354 | 0.0354 | 0.0354 | 0.0354 |

Source: The original estimates are MLE based estimates obtained from IPUMS-CPS ASEC survey data (1977-2015). The averages in column 2 to column 6 are based on the results from the simulation exercises.

Notes: The numbers represent the average of the estimates obtained from two simulation exercises. $\widehat{\Omega}$ represents the vector of original estimates (before bias correction), whereas $\widetilde{\Omega}$ represents the constant biased corrected estimate (MacKinnon and Smith (1998)) based on the original estimates of $\widehat{\Omega}$.

**Table 3: Uncorrected and bias corrected estimates and slope of the bias function**

| | $\mu_\omega$ | $\mu_v$ | $\sigma_u$ |
|---|---|---|---|
| Estimate, original | 0.0406 | 0.0338 | 0.0351 |
| Standard error | *0.0014* | *0.0005* | *0.0008* |
| | | | |
| Slope of bias function | -0.3003 | -0.1915 | -0.2746 |
| | | | |
| Estimate, bias corrected (LBC) | 0.0507 | 0.0377 | 0.0341 |
| Standard error | *0.0020* | *0.0006* | *0.0011* |

Source: The original estimates are obtained from IPUMS-CPS ASEC survey data (1977-2015).
Notes: The bias correction is conducted based on MacKinnon and Smith (1998) simulation based linear bias correction method. Slope represents the estimated slope of the linear bias function.

**Table 4: Estimates of joiners and leavers (averages), by age and gender.**

| | Joiners (% of NLF) | | | Leavers (% of LF) | | |
|---|---|---|---|---|---|---|
| Age group | Men | Women | Total | Men | Women | Total |
| 16-30 | 0.053 | 0.051 | 0.052 | 0.029 | 0.032 | 0.031 |
| 30-42 | 0.050 | 0.055 | 0.052 | 0.021 | 0.026 | 0.023 |
| 42-60 | 0.051 | 0.052 | 0.052 | 0.022 | 0.027 | 0.025 |
| 60 and above | 0.031 | 0.020 | 0.025 | 0.058 | 0.065 | 0.062 |
| Total | 0.047 | 0.045 | 0.046 | 0.031 | 0.037 | 0.034 |

Source: Based on our estimates obtained from IPUMS-CPS ASEC survey data (1977-2015).
Notes: Joiners represents the proportion of people not in the labor force (NLF) in previous year who join the labor force in current year. Leavers represents the proportion of people in the labor force (LF) in previous year who leave the labor force in current year.

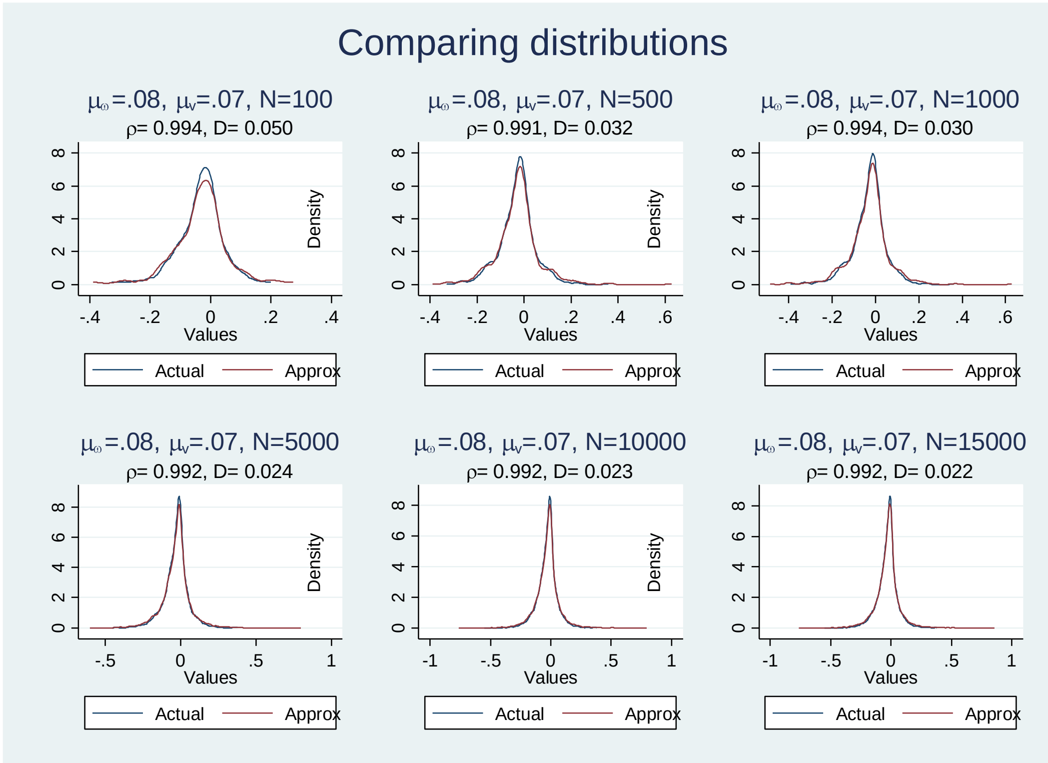
**Table 5: Average actual and predicted labor force participation rate between 1978-2015, by age and gender**

| | Average LFPR, 1978-2015 | |
|---|---|---|
| | **Actual** | **Predicted** |
| Men | 0.716 | 0.707 |
| Women | 0.572 | 0.570 |
| Total | 0.644 | 0.638 |
| | | |
| *Age group* | | |
| 16-30 | 0.701 | 0.696 |
| 30-42 | 0.838 | 0.827 |
| 42-60 | 0.788 | 0.780 |
| 60 and above | 0.250 | 0.251 |
| Total | 0.644 | 0.638 |
| | | |
| Correlation (actual, predicted) | 0.9973 | |

Source: Both actual and predicted values are based on our estimates obtained from IPUMS-CPS ASEC survey data (1977-2015).

Notes: The numbers represent the proportion of working age population in the respective groups. Actual labor force is computed by averaging groups specific LFPRs observed in the data. The predicted LFPR is obtained by adjusting previous year's actual LFPR by our estimates of joiners and leavers.

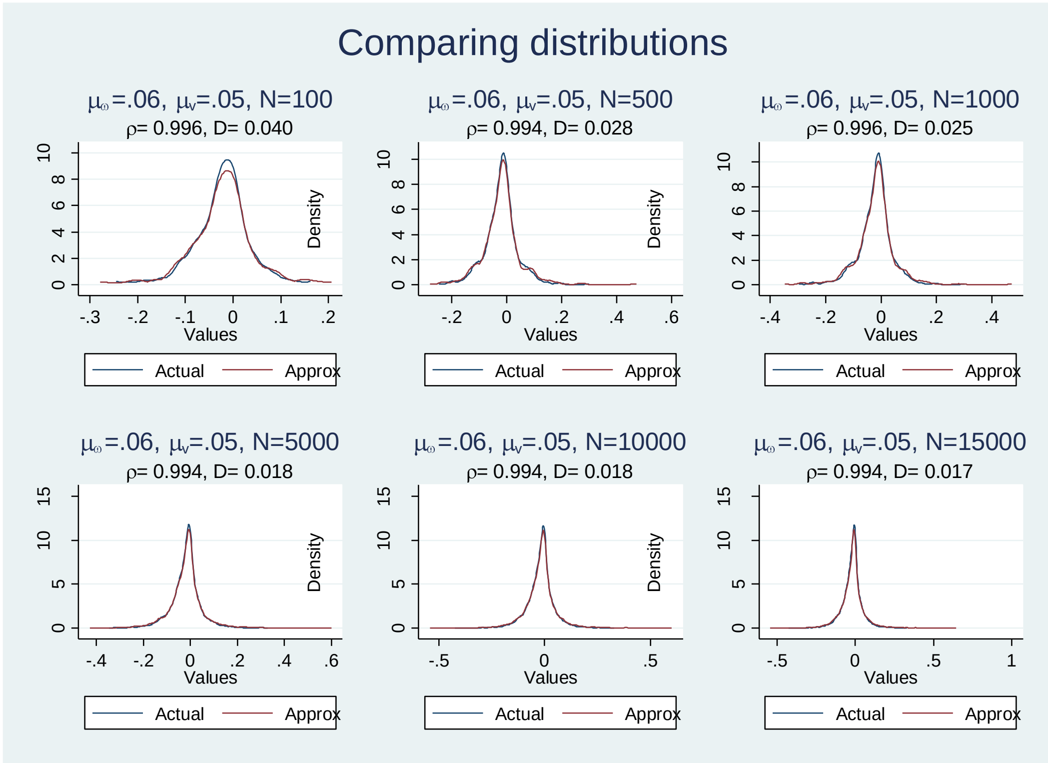
Figure 1a: Comparing the actual and the Taylor's series approximation of the error (Eq 11 vs. Eq 12)




Source: Based on our simulation exercises. See text for details.

Notes: The values for $\mu_\omega$ and $\mu_v$ are our choices of parameter values for the simulations. N and $\rho$ represent the sample size and the correlation coefficient between the actual and the approximated series. D represents the Kolmogorov-Smirnov distance statistic between the distribution of actual and the approximated series.

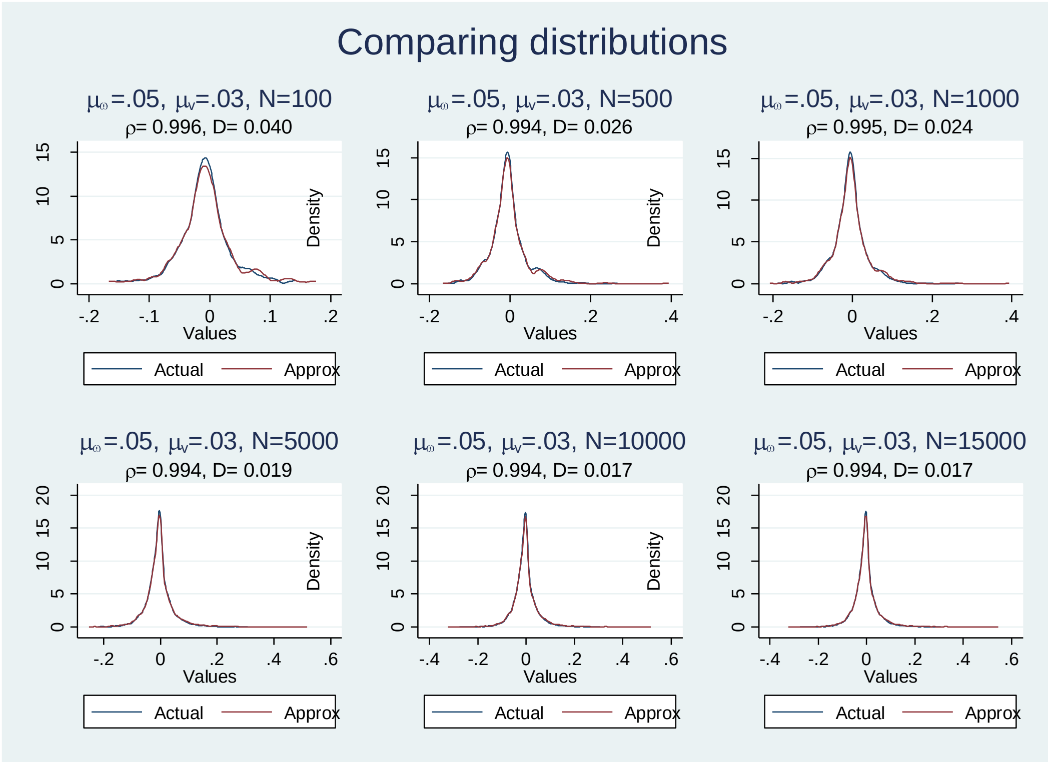
Figure 1b: Comparing the actual and the Taylor's series approximation of the error (Eq 11 vs. Eq 12)




Source: Based on our simulation exercises. See text for details.

Notes: The values for $\mu_\omega$ and $\mu_v$ are our choices of parameter values for the simulations. N, $\rho$, represent the sample size, and the correlation coefficient between the actual and the approximated series. D represents the Kolmogorov-Smirnov distance statistic between the distribution of actual and the approximated series.

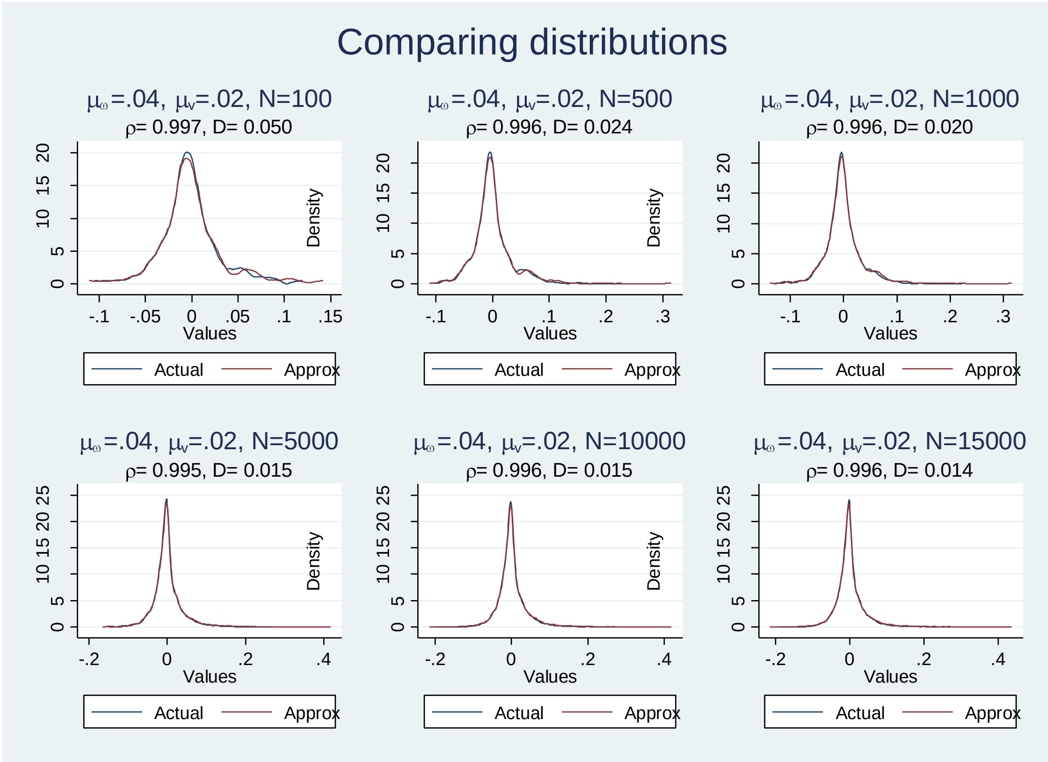
Figure 1c: Comparing the actual and the Taylor's series approximation of the error (Eq 11 vs. Eq 12)




Source: Based on our simulation exercises. See text for details.

Notes: The values for $\mu_\omega$ and $\mu_v$ are our choices of parameter values for the simulations. N, $\rho$, represent the sample size, and the correlation coefficient between the actual and the approximated series. D represents the Kolmogorov-Smirnov distance statistic between the distribution of actual and the approximated series.

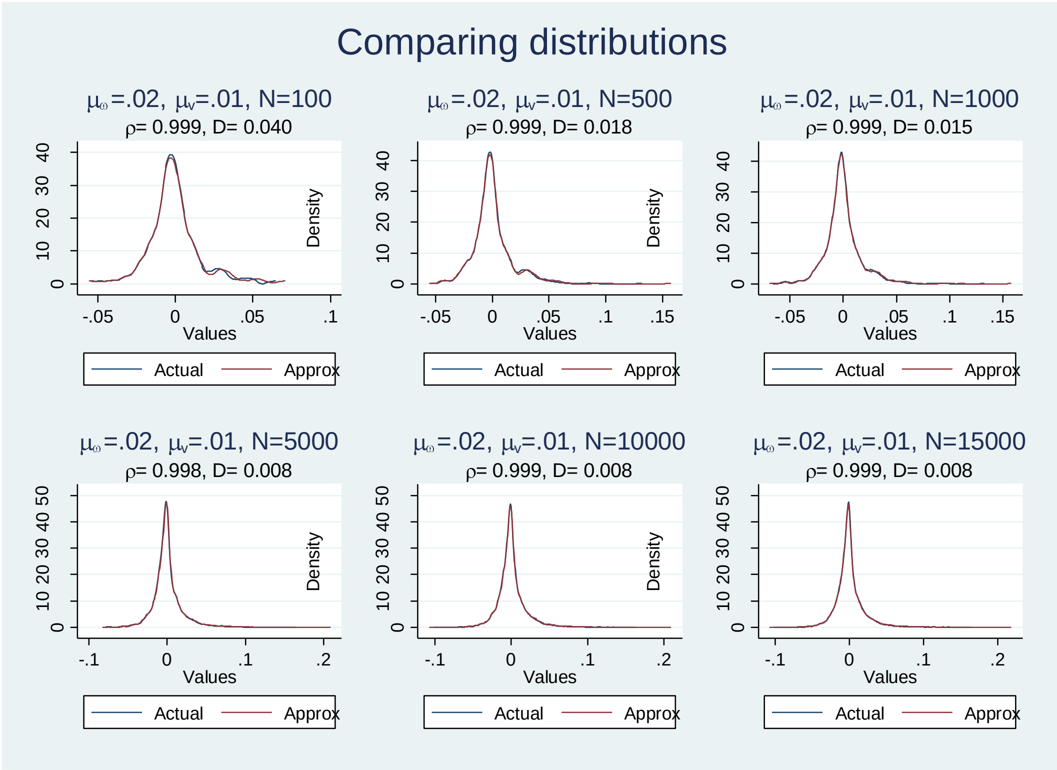
Figure 1d: Comparing the actual and the Taylor's series approximation of the error (Eq 11 vs. Eq 12)




Source: Based on our simulation exercises. See text for details.

Notes: The values for $\mu_{\omega}$ and $\mu_{v}$ are our choices of parameter values for the simulations. N, $\rho$, represent the sample size, and the correlation coefficient between the actual and the approximated series. D represents the Kolmogorov-Smirnov distance statistic between the distribution of actual and the approximated series.

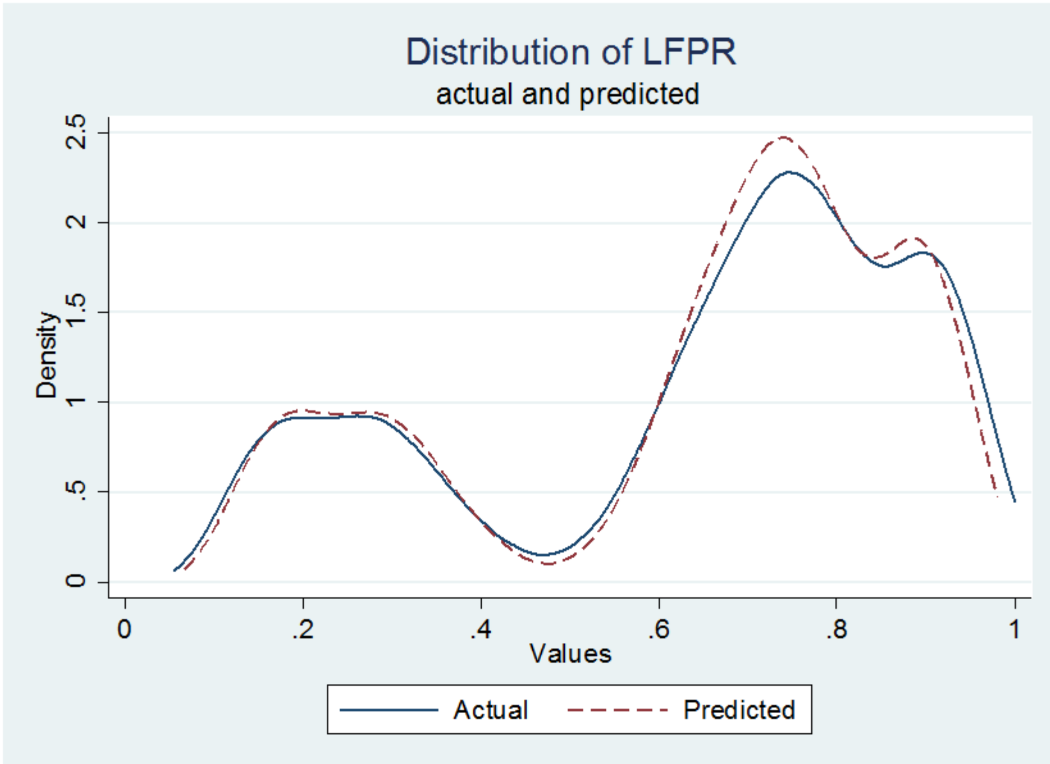
Figure 1e: Comparing the actual and the Taylor's series approximation of the error (Eq 11 vs. Eq 12)




Source: Based on our simulation exercises. See text for details.

Notes: The values for $\mu_\omega$ and $\mu_v$ are our choices of parameter values for the simulations. N, $\rho$, represent the sample size, and the correlation coefficient between the actual and the approximated series. D represents the Kolmogorov-Smirnov distance statistic between the distribution of actual and the approximated series.

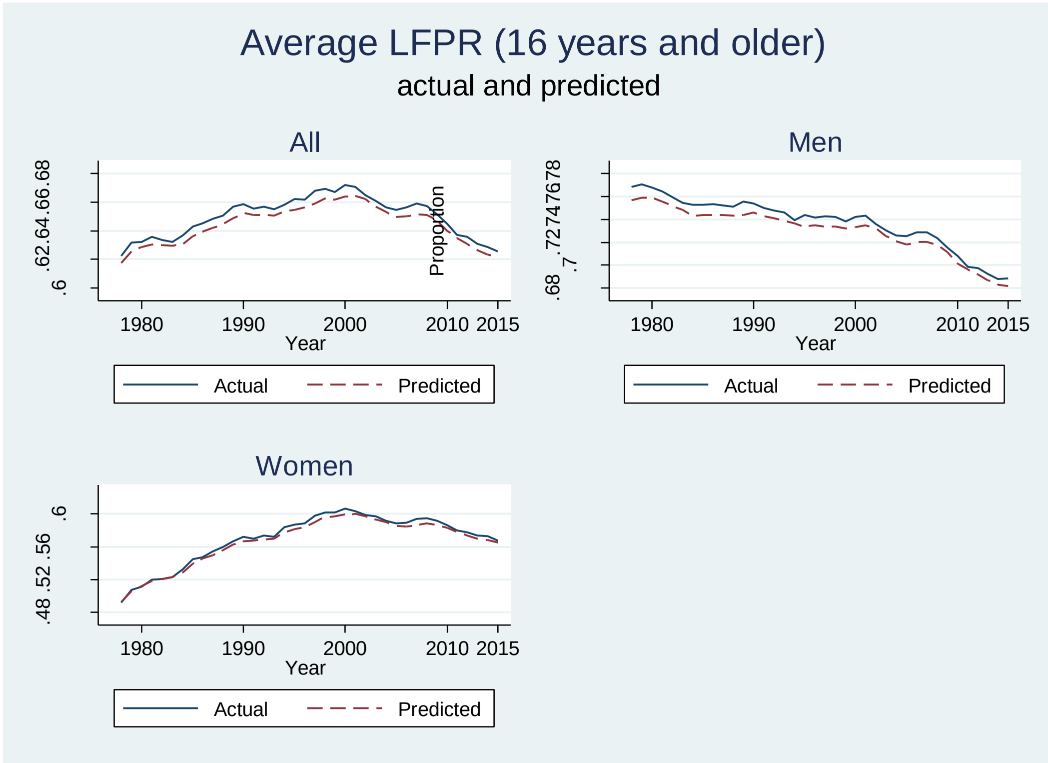
Figure 2: Comparing the distribution of actual and predicted labor force participation rate




Source: Both actual and predicted values are based on our estimates obtained from IPUMS-CPS ASEC survey data (1977-2015).

Note: Actual labor force values are the observed LFPRs. The predicted LFPR is obtained by adjusting previous year's actual LFPR by our estimates of joiners and leavers.

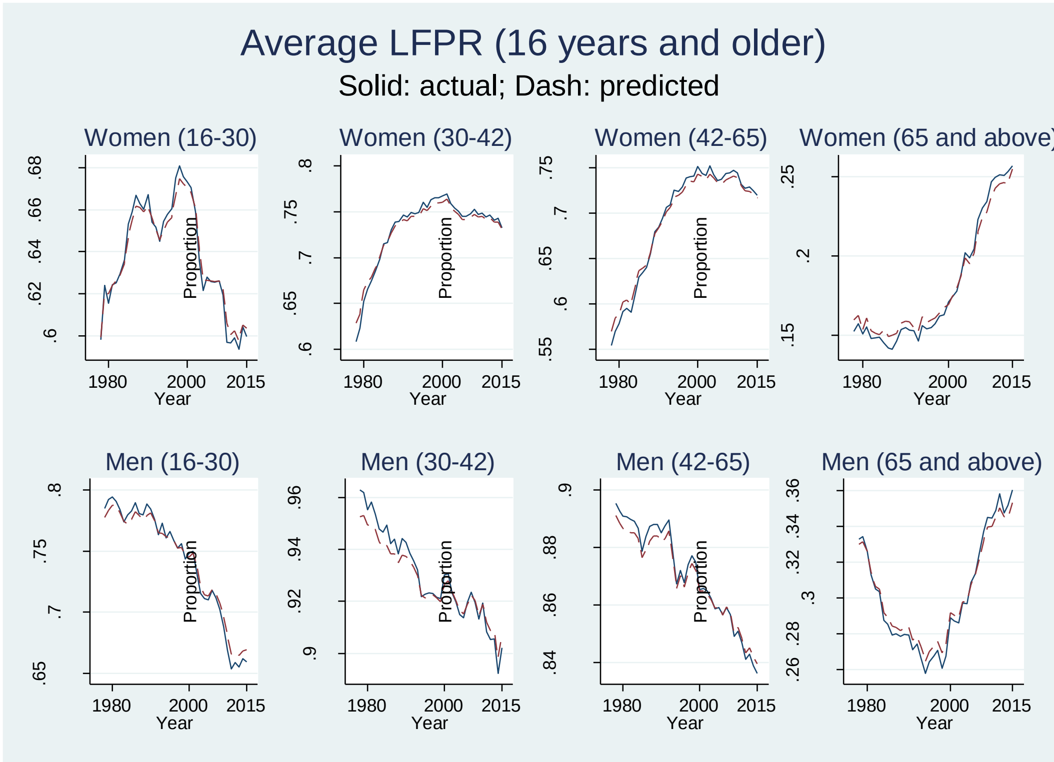
Figure 3: Comparing the actual and predicted labor force participation rate (1978-2015) by gender




Source: Both actual and predicted values are based on our estimates obtained from IPUMS-CPS ASEC survey data (1977-2015).

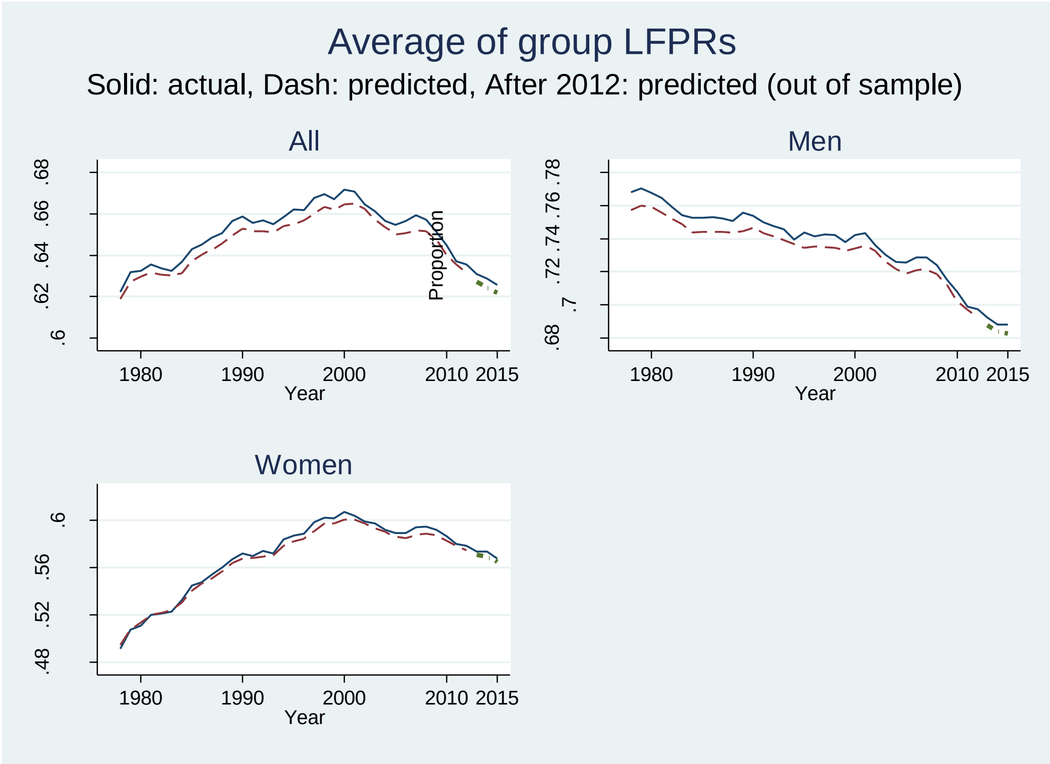
Notes: The numbers represent the proportion of working age population. Actual LFPR is computed by averaging the group-specific observed LFPR in the data. The predicted LFPR is obtained by adjusting previous year's actual LFPR by our estimates of joiners and leavers.

Figure 4: Comparing the actual and predicted labor force participation rate (1978-2015) by gender and age




Source: Both actual and predicted values are based on our estimates obtained from IPUMS-CPS ASEC survey data (1977-2015).
Notes: The numbers represent the proportion of working age population. Actual LFPR is computed by averaging the group-specific observed LFPR in the data. The predicted LFPR is obtained by adjusting previous year's actual LFPR by our estimates of the proportion of joiners and leavers.

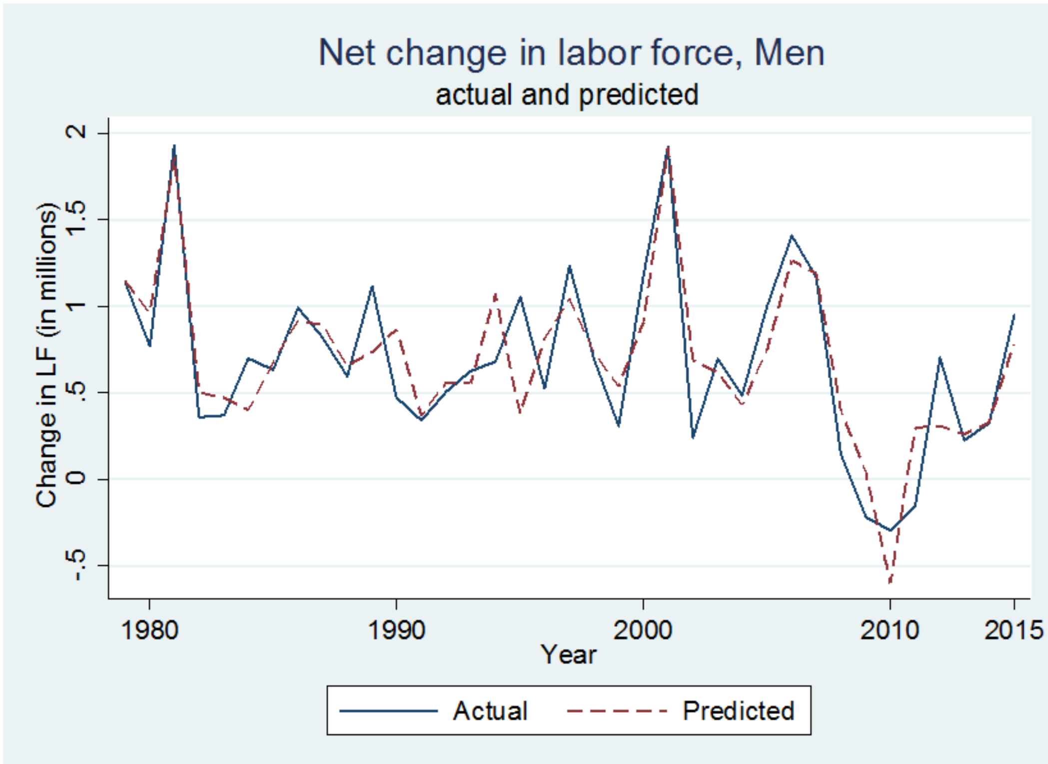
Figure 5: Out of sample predictions




Source: Both actual and predicted values are based on the data from IPUMS-CPS ASEC survey data (1977-2015).

Note: The parameter estimates are based on 1977-2012 data. The predicted LFPRs for 2013, 2014, 2015 are out-of-sample estimates. These values are calculated by applying a modified Jondrow et. al. (1982) formula described in the text.

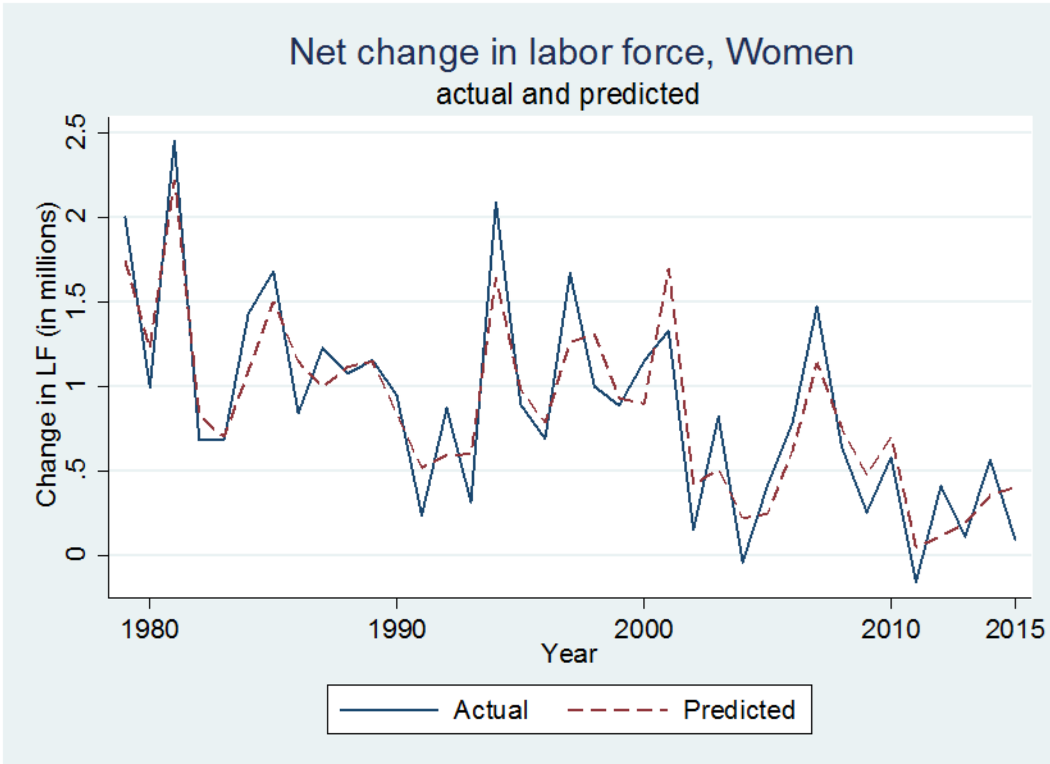
Figure 6a: Comparing net change in actual and predicted labor force, 1978-2015 (Men)




Source: Both actual and predicted values are based on our estimates obtained from IPUMS-CPS ASEC survey data (1977-2015).

Note: The net change in actual labor force is computed by subtracting previous year's labor force from the current year's labor force. The change in predicted labor force is computed by subtracting the previous year's predicted labor force from the current year's predicted labor force. The predicted labor force is obtained by multiplying the average predicted LFPR by the total working age population in the same year.

Figure 6b: Comparing net change in actual and predicted labor force, 1978-2015 (Women)




Source: Both actual and predicted values are based on our estimates obtained from IPUMS-CPS ASEC survey data (1977-2015).

Note: The net change in actual labor force is computed by subtracting previous year's labor force from the current year's labor force. The change in predicted labor force is computed by subtracting the previous year's predicted labor force from the current year's predicted labor force. The predicted labor force is obtained by multiplying the average predicted LFPR by the total working age population in the same year.

## References


Abowd, J. M., & Zellner, A. (1985). Estimating gross labor-force flows. *Journal of Business & Economic Statistics*, *3*(3), 254-283.

Ahn, Seung C. & Robin C. Sickles (2000) Estimation of long-run inefficiency levels: a dynamic frontier approach, *Econometric Reviews*, 19(4): 461-492.

Barnichon, R., & Garda, P. (2016). Forecasting unemployment across countries: The ins and outs. *European Economic Review*, *84*, 165-183.

Barnichon, R., & Nekarda, C. J. (2012). The ins and outs of forecasting unemployment: Using labor force flows to forecast the labor market. *Brookings Papers on Economic Activity*, *2012*(2), 83-131.

Basmann, R. L. (1957). A generalized classical method of linear estimation of coefficients in a structural equation. *Econometrica: Journal of the Econometric Society*, 77-83.

Basmann, R. L. (1960). On finite sample distributions of generalized classical linear identifiability test statistics. *Journal of the American Statistical Association*, *55*(292), 650-659.

Basmann, R. L. (1985). A theory of serial correlation of stochastic taste changers in direct utility functions. *Econometric Theory*, *1*(02), 192-210.

Basmann, R. L. (2006). "Four Entrenched Notions Post Walrasians Should Avoid," in Colander, D. (Ed.). *Post Walrasian macroeconomics: Beyond the dynamic stochastic general equilibrium model*. Cambridge University Press, pp 277-86.

Bureau of Labor Statistics (2015) "Research series on labor force status flows from the Current Population Survey," https://www.bls.gov/cps/cps_flows.htm.

Colombi, Roberto (2010) "A skew normal stochastic frontier model for panel data," *Proceedings of the 45$^{th}$ Scientific meeting of the Italian Statistical Society*, also available at http://new.sis-statistica.org/wp-content/uploads/2013/09/RS10-A-Skew-Normal-Stochastic-Frontier-Model-for-Panel-Data.pdf.

Filippini, Massimo & William Greene. (2016). "Persistnet and transient productive inefficiency: a maximum likelihood approach," *Journal of Productivity Analysis* 45(2):187-196.

Flood, S. K., Ruggles , M, S, & Warren, J. R. (2015). Integrated Public Use Microdata Series, Current Population Survey: Version 4.0. [Machine-readable database]. Minneapolis: University of Minnesota.

Frazis, H. J., Robison, E. L., Evans, T. D., & Duff, M. A. (2005). Estimating gross flows consistent with stocks in the CPS. *Monthly Labor Review*, *128*, 3-9.

Fuller, W. A., & Chua, T. C. (1985, June). Gross change estimation in the presence of response error. In *Proceedings of the Conference on Gross Flows in Labor Force Statistics* (pp. 65-77). US Census Bureau and US Bureau of Labor Statistics Washington, DC.

Greene, W. (2005). Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. *Journal of Econometrics*, *126*(2), 269-303.

Jondrow, J., Lovell, C. K., Materov, I. S., & Schmidt, P. (1982). On the estimation of technical inefficiency in the stochastic frontier production function model. *Journal of econometrics*, *19*(2-3), 233-238.

Kiefer, Nicholas M., and George R. Neumann. (2006). *Search models and applied labor economics*. Cambridge University Press.

Kumbhakar, S.C & Parmeter, C. F. (2009). The effects of match uncertainty and bargaining on labor market outcomes: evidence from firm and worker specific estimates. *Journal of Productivity Analysis*, 31(1), 1-14.

MacKinnon, J. G., & Smith, A. A. (1998). Approximate bias correction in econometrics. *Journal of Econometrics*, *85*(2), 205-230.

Petrongolo, B. & Pissarides , C.A. (2008). The ins and outs of European unemployment. *American Economic Review,* 98(2), 256-62.

Polachek, S. W., Das, T., & Thamma-Apiroam, R. (2015). Micro- and Macroeconomic Implications of Heterogeneity in the Production of Human Capital, *Journal of Political Economy*, 123(6), 1410-1455.

Polachek, S. W., & Yoon, B. J. (1987). A two-tiered earnings frontier estimation of employer and employee information in the labor market. *The Review of Economics and Statistics*, 296-302.

Poterba, J. M., & Summers, L. H. (1986). Reporting errors and labor market dynamics. *Econometrica: Journal of the Econometric Society*, 1319-1338.

Steven Sabol & Reid Calamita. (2015). "The ins and outs of labor market forecasting," Working Paper.

Shimer, Robert. (2012). "Reassessing the ins and outs of unemployment," *Review of Economic Dynamics* 15:127–14.

Tsionas, Efthymios G. & Subal C. Kumbhakar. (2014). "Firm heterogeneity persistent and transient technical inefficiency: a generalized true random-effects model," *Journal of Applied Econometrics* 29(1):110–132.

**Appendix I**

The function to be approximated is

$$f(\theta^*_{it-1}, \omega_{it}, v_{it}) = Log\left(1 + \frac{\omega_{it}}{(1+\omega_{it})}\left(e^{\theta^*_{it-1}} - 1\right) - \frac{v_{it}}{(1+v_{it})}\right)$$

A first order Taylor's series approximation of $f(.)$ around $\theta^*_{it-1} = \theta_{0i},, \omega_{it} = 0, v_{it} = 0$ is

$$\begin{aligned} f(\theta^*_{it-1}, \omega_{it}, v_{it}) & \\ &= f(\theta^*_{it-1} = \theta_{0i},, \omega_{it} = 0, v_{it} = 0) \\ &+ f_{\theta^*_{it-1}}(\theta^*_{it-1} = \theta_{0i},, \omega_{it} = 0, v_{it} = 0)(\theta^*_{it-1} - \theta_{0i}) \\ &+ f_{\omega_{it}}(\theta^*_{it-1} = \theta_{0i},, \omega_{it} = 0, v_{it} = 0)(\omega_{it} - 0) \\ &+ f_{v_{it}}(\theta^*_{it-1} = \theta_{0i},, \omega_{it} = 0, v_{it} = 0)(v_{it} - 0) \qquad (A1.2) \end{aligned}$$

The value of $f(\theta^*_{it-1} = \theta_{0i},, \omega_{it} = 0, v_{it} = 0) = Log(1) = 0$

The individual derivatives are

$$\begin{aligned} f_{\theta^*_{it-1}}(\theta^*_{it-1} = \theta_{0i},, \omega_{it} = 0, v_{it} = 0) & \\ &= \frac{1}{\left(1 + \frac{\omega_{it}}{(1+\omega_{it})}\left(e^{\theta^*_{it-1}} - 1\right) - \frac{v_{it}}{(1+v_{it})}\right)} e^{\theta^*_{it-1}} \frac{\omega_{it}}{(1+\omega_{it})} = 0 \ \ [\because \ \omega_{it} = 0\,] \end{aligned}$$

$$\begin{aligned} f_{\omega_{it}}(\theta^*_{it-1} = \theta_{0i},, \omega_{it} = 0, v_{it} = 0) & \\ &= \frac{1}{\left(1 + \frac{\omega_{it}}{(1+\omega_{it})}\left(e^{\theta^*_{it-1}} - 1\right) - \frac{v_{it}}{(1+v_{it})}\right)} \left(e^{\theta^*_{it-1}} - 1\right)\left[\frac{1}{(1+\omega_{it})} \right. \\ &\left. - \frac{\omega_{it}}{(1+\omega_{it})^2}\right] = \left(e^{\theta_{0i}} - 1\right) \end{aligned}$$

$$\begin{aligned} f_{v_{it}}(\theta^*_{it-1} = \theta_{0i},, \omega_{it} = 0, v_{it} = 0) & \\ &= -\frac{1}{\left(1 + \frac{\omega_{it}}{(1+\omega_{it})}\left(e^{\theta^*_{it-1}} - 1\right) - \frac{v_{it}}{(1+v_{it})}\right)}\left[\frac{1}{(1+v_{it})} - \frac{v_{it}}{(1+v_{it})^2}\right] = -1 \end{aligned}$$

Substituting these values in $(A1.2)$ yields

$$f(\theta^*_{it-1}, \omega_{it}, v_{it}) \approx 0 + 0 \times (\theta^*_{it-1} - \theta_{0i}) + \left(e^{\theta_{0i}} - 1\right)(\omega_{it} - 0) - 1 \times (v_{it} - 0)$$

After simplifying we obtain the following approximated function

$$f(\theta^*_{it-1}, \omega_{it}, v_{it}) \approx \left(e^{\theta_{0i}} - 1\right)\omega_{it} - v_{it}$$

## Appendix II

Consider three statistically independent random variables $\omega_{it}, v_{it}, u_{it}$. Suppose they follow the following probability distributions (suppressing the subscripts):

$$\omega \sim Exp(\mu_\omega); f(\omega) = \frac{1}{\mu_\omega} e^{-\frac{\omega}{\mu_\omega}}, where\ \omega \in [0, \infty) \qquad (A2.1a)$$

$$v \sim Exp(\mu_v); f(v) = \frac{1}{\mu_v} e^{-\frac{v}{\mu_v}}, where\ v \in [0, \infty) \qquad (A2.1b)$$

$$u \sim N(0, \sigma_u);\ f(u) = \frac{1}{\sigma_u\sqrt{2\pi}} e^{-\frac{u^2}{2\sigma_u^2}}, where\ u \in (-\infty, \infty) \qquad (A2.1c)$$

Define the composite error $\epsilon$ as

$$\epsilon = u + \ \omega - v \qquad (A2.2)$$

The goal is to obtain the density of the composite error $\epsilon$ i.e. $f(\epsilon)$. It is the combination of the three random variables. We derive this density in two steps. In the first we define another random variable $z$ such that

$$z = u - v \qquad (A2.3)$$

First we will find the density of $z$ i.e. $f(z)$. An algebraic manipulation of yields

$$u = z + v \qquad (A2.4)$$

Since $u$ and $v$ are statistically independent, their joint density is the product of their marginal densities

$$f(z + v, v) = f(u)f(v) = f_u(z + v)f_v(v) \qquad (A2.5)$$

To obtain the density of $z$, one needs to integrate out $v$, i.e.

$$f_z(z) = \int_0^\infty f_u(z + v)f_v(v)dv = \frac{1}{\mu_v\sigma_u\sqrt{2\pi}} \int_0^\infty e^{-\frac{v}{\mu_v}} e^{-\frac{(z+v)^2}{2\sigma_u^2}} dv \qquad (A2.6)$$

The indefinite integral is

$$f_z(z) = \int_0^\infty f_u(z + v)f_v(v)dv = \frac{1}{\mu_v\sigma_u\sqrt{2\pi}} \sqrt{\frac{\pi}{2}} \sigma_u e^{\frac{\sigma_u^2}{2\mu_v^2} + \frac{z}{\mu_v}} \mathrm{erf}\left(\frac{\mu_v(v + z) + \sigma_u^2}{\mu_v\sigma_u\sqrt{2}}\right) \qquad (A2.7)$$

where erf(.) represents the error function. Evaluating (A2.7) for $v = 0, \infty$ yields

$$f(z) = \frac{1}{2\mu_v} e^{\frac{\sigma_u^2}{2\mu_v^2}+\frac{z}{\mu_v}} \left\{1 - \operatorname{erf}\left[\frac{1}{\sqrt{2}}\left(\frac{z}{\sigma_u} + \frac{\sigma_u}{\mu_v}\right)\right]\right\}$$

$$f(z) = \frac{1}{2\mu_v} e^{\frac{\sigma_u^2}{2\mu_v^2}+\frac{z}{\mu_v}} \left[2 - 2\Phi\left(\frac{z}{\sigma_u} + \frac{\sigma_u}{\mu_v}\right)\right]$$

$$f(z) = \frac{1}{\mu_v} e^{\frac{\sigma_u^2}{2\mu_v^2}+\frac{z}{\mu_v}} \Phi\left[-\left(\frac{z}{\sigma_u} + \frac{\sigma_u}{\mu_v}\right)\right] \qquad (A2.8)$$

Now derive $f(\epsilon)$ based on $\epsilon = z + \omega$. Since $u$ and $v$ are statistically independent to $\omega$, the sum of $u + v = z$ is also independent of $\omega$. Thus

$$f(z,\omega) = f(z)f(\omega) = f_z(\epsilon - \omega)f_\omega(\omega) \qquad (A2.9)$$

Integrating out $\omega$ from (A2.9) yields the density if $\epsilon$, i.e.

$$f(\epsilon) = \int_0^\infty f_z(\epsilon - \omega)f_\omega(\omega)d\omega$$

$$f(\epsilon) = \frac{1}{\mu_\omega \mu_v} e^{\frac{\sigma_u^2}{2\mu_v^2}} \int_0^\infty e^{\frac{(\epsilon-\omega)}{\mu_v}} e^{-\frac{\omega}{\mu_\omega}} \Phi\left[-\left\{\frac{(\epsilon - \omega)}{\sigma_u} + \frac{\sigma_u}{\mu_v}\right\}\right] d\omega$$

$$f(\epsilon) = \frac{1}{\mu_\omega \mu_v} e^{\frac{\sigma_u^2}{2\mu_v^2}} e^{\frac{\epsilon}{\mu_v}} \int_0^\infty e^{-\omega\left(\frac{1}{\mu_v}+\frac{1}{\mu_\omega}\right)} \Phi\left[-\left\{\frac{(\epsilon - \omega)}{\sigma_u} + \frac{\sigma_u}{\mu_v}\right\}\right] d\omega$$

$$f(\epsilon) = \frac{1}{\mu_\omega \mu_v} e^{\frac{\sigma_u^2}{2\mu_v^2}} e^{\frac{\epsilon}{\mu_v}} \int_0^\infty e^{-\omega\left(\frac{1}{\mu_v}+\frac{1}{\mu_\omega}\right)} \Phi\left[\frac{\omega}{\sigma_u} - \left(\frac{\epsilon}{\sigma_u} + \frac{\sigma_u}{\mu_v}\right)\right] d\omega \qquad (A2.10)$$

Let $G = \frac{1}{\mu_\omega \mu_v} e^{\frac{\sigma_u^2}{2\mu_v^2}} e^{\frac{\epsilon}{\mu_v}}$; $a = \left(\frac{1}{\mu_\omega} + \frac{1}{\mu_v}\right)$; $b = \frac{1}{\sigma_u}$; $c = \left(\frac{\epsilon}{\sigma_u} + \frac{\sigma_u}{\mu_v}\right)$. Substituting these in (A2.10) yields the integrand as

$$G \int_0^\infty e^{-a\omega} \Phi(b\omega - c)d\omega \qquad (A2.11)$$

One way to solve this is by integration by parts. For that, let $x = \Phi(b\omega - c)$ and $dy = e^{-a\omega}d\omega$. These relationships imply that $dx = b\phi(b\omega - c)$ and $y = -\frac{1}{a}e^{-a\omega}$. Thus the integrand

$$\int_0^\infty x dy = xy|_0^\infty - \int_0^\infty y dx \qquad (A2.12)$$

$$xy|_0^\infty = -\frac{1}{a} e^{-a\omega} \Phi(b\omega - c)|_0^\infty = \frac{1}{a}\Phi(-c) \qquad (A2.13)$$

$$\int_0^\infty y dx = \int_0^\infty -\frac{b}{a} e^{-a\omega} \phi(b\omega - c) d\omega = -\frac{b}{a} \int_0^\infty e^{-a\omega} \phi(b\omega - c) d\omega$$

$$= -\frac{b}{a\sqrt{2\pi}} \int_0^\infty e^{-a\omega} e^{-\frac{(b\omega - c)^2}{2}} d\omega$$

$$= -\frac{b}{a\sqrt{2\pi}} \times \left(-\frac{1}{b}\right) \times \sqrt{\frac{\pi}{2}} e^{\frac{a(a-2bc)}{2b^2}} \times \text{erf}\left(\frac{b(c - b\omega) - a}{b\sqrt{2}}\right)\Bigg|_0^\infty$$

$$= \frac{1}{2a} e^{\frac{a(a-2bc)}{2b^2}} \left[-1 - \text{erf}\left(\frac{bc - a}{b\sqrt{2}}\right)\right] = -\frac{1}{2a} e^{\frac{a(a-2bc)}{2b^2}} \left[1 + \text{erf}\left(\frac{bc - a}{b\sqrt{2}}\right)\right]$$

$$= -\frac{1}{2a} e^{\frac{a(a-2bc)}{2b^2}} \left[1 + 2\Phi\left(\frac{bc - a}{b}\right) - 1\right] = -\frac{1}{a} e^{\frac{a(a-2bc)}{2b^2}} \Phi\left(\frac{bc - a}{b}\right) \qquad (A2.14)$$

Combining (A2.13) and (A2.14) yields the marginal density of $\epsilon$ as

$$f(\epsilon) = \frac{G}{a}\left[\Phi(-c) + e^{\frac{a(a-2bc)}{2b^2}} \Phi\left(\frac{bc - a}{b}\right)\right]$$

where $G, a, b, c$ are defined above.

## Appendix III: Estimating $\widehat{\omega}^*$ and $\widehat{v}$

The first step to obtain joiners and leavers is to calculate $\widehat{\omega}^*$ and $\hat{v}$. Unfortunately, none of the three error components can be computed directly. The only available estimates are the estimates of the composite residual i.e. $\hat{\epsilon}$, $\hat{\mu}_{\omega^*}$, $\hat{\mu}_v$ and $\hat{\sigma}_u$. However one can still compute estimates of $\widehat{\omega}^*$ and $\hat{v}$ as their expected values conditioned on the value of $\hat{\epsilon}$. Mathematically these can be represented as

$$\widehat{\omega}^* = E(\omega^*|\hat{\epsilon}) = \int_0^\infty \omega^* f(\omega^*|\hat{\epsilon}) d\omega^* \qquad (A3.1a)$$

$$\hat{v} = E(v|\hat{\epsilon}) = \int_0^\infty v\, f(v|\hat{\epsilon}) dv \qquad (A3.1b)$$

As suggested in Johndrow et. al. (1982) these conditional means serves as individual specific estimates of the two one-sided errors. Statistically they are best linear unbiased predictor of $\widehat{\omega}^*$ and $\hat{v}$.

Solving the integral requires expression for the $f(\omega^*|\hat{\epsilon})$ and $f(v|\hat{\epsilon})$ which are unknown. However, using Bayes' theorem one can write these conditional densities as

$$f(\omega^*|\hat{\epsilon}) = \frac{f(\omega^*, \hat{\epsilon})}{f(\hat{\epsilon})} \qquad (A3.2a)$$

$$f(v|\hat{\epsilon}) = \frac{f(v, \hat{\epsilon})}{f(\hat{\epsilon})} \qquad (A3.2b)$$

Substituting these into A2.1a-1b, we obtain

$$E(\omega^*|\hat{\epsilon}) = \frac{1}{f(\hat{\epsilon})} \int_0^\infty \omega^* f(\omega^*, \hat{\epsilon}) d\omega^* \qquad (A3.3a)$$

$$E(v|\hat{\epsilon}) = \frac{1}{f(\hat{\epsilon})} \int_0^\infty v\, f(v, \hat{\epsilon}) dv \qquad (A3.3b)$$

Appendix II derives the formula for $f(\epsilon)$ which allows one to obtain $f(\hat{\epsilon})$ by substituting $\epsilon$ by $\hat{\epsilon}$. The joint densities $f(\omega^*, \hat{\epsilon})$ and $f(v, \hat{\epsilon})$ however are not directly available but can be derived from the statistical independence assumption and the density of each of the error components.

*Derivation of* $E(\omega^*|\hat{\epsilon})$:

For notational convenience we first derive $E(\omega^*|\epsilon)$ and then substitute $\epsilon$ by $\hat{\epsilon}$ to obtain $E(\omega^*|\hat{\epsilon})$.

$$f(\omega^*) = \frac{1}{\mu_{\omega^*}} e^{-\frac{\omega^*}{\mu_{\omega^*}}} \qquad (A3.4a)$$

$$f(v) = \frac{1}{\mu_v} e^{-\frac{v}{\mu_v}} \qquad (A3.4b)$$

$$f(u) = \frac{1}{\sigma_u\sqrt{2\pi}} e^{-\frac{u^2}{2\sigma_u^2}} \qquad (A3.4c)$$

All of the random variables above are statistically independent. This means that if we define $z = u - v$, the $z$ and $\omega$ will be statistically independent. Using this property one can therefore write

$$f(\omega^*, z) = f(\omega^*) f(z)$$

The density $f(\omega)$ is already known and $f(z)$ is already derived in Appendix II. Thus one can write

$$f(\omega^*, z) = \frac{1}{\mu_{\omega^*}} e^{-\frac{\omega^*}{\mu_{\omega^*}}} \frac{1}{\mu_v} e^{\frac{\sigma_u^2}{2\mu_v^2}+\frac{z}{\mu_v}} \Phi\left[-\left(\frac{z}{\sigma_u} + \frac{\sigma_u}{\mu_v}\right)\right]$$

Making the transformation $\epsilon = \omega^* + z$, the joint density of $\omega^*$ and $\epsilon$ is

$$f(\omega^*, \epsilon) = \frac{1}{\mu_{\omega^*}} e^{-\frac{\omega^*}{\mu_{\omega^*}}} \frac{1}{\mu_v} e^{\frac{\sigma_u^2}{2\mu_v^2}+\frac{(\epsilon-\omega)}{\mu_v}} \Phi\left[-\left(\frac{\epsilon - \omega^*}{\sigma_u} + \frac{\sigma_u}{\mu_v}\right)\right]$$

Let $a' = \left(\frac{1}{\mu_v} + \frac{1}{\mu_{\omega^*}}\right)$; $b' = \sigma_u$; $c' = \left(\frac{\epsilon}{\sigma_u} + \frac{\sigma_u}{\mu_v}\right)$; $\theta_1' = \frac{1}{\mu_v \mu_{\omega^*}} e^{\frac{\epsilon}{\mu_v}+\frac{\sigma_u^2}{2\mu_v^2}}$. Utilizing these re-parameterizations

$$f(\omega^*, \epsilon) = \theta_1' \, e^{-a'\omega^*} \Phi\left(\frac{\omega^*}{b'} - c'\right) \qquad (A3.5)$$

Thus the expression for $E(\omega^*|\epsilon)$ can be written as

$$E(\omega^*|\epsilon) = \int_0^\infty \omega^* \frac{f(\omega^*, \epsilon)}{f(\epsilon)} d\omega^* == \frac{1}{f(\epsilon)} \int_0^\infty \omega^* f(\omega^*, \epsilon) d\omega = \frac{\theta_1'}{f(\epsilon)} \int_0^\infty \omega^* e^{-a'\omega^*} \Phi\left(\frac{\omega^*}{b'} - c'\right) d\omega^* \quad (A3.6)$$

The right hand side of (A3.6) can be integrated by using integration by parts formula

Suppose $x' = \Phi\left(\frac{\omega^*}{b'} - c'\right)$; $dy' = \omega^* e^{-a'\omega^*} d\omega^*$. This means $dx' = \frac{1}{b'} \phi\left(\frac{\omega^*}{b'} - c'\right) d\omega$ and $y' = -\frac{(1+a'\omega^*)}{a'^2} e^{-a'\omega^*}$

Therefore the integral $\int_0^\infty \omega^* e^{-a'\omega^*} \Phi(\frac{\omega^*}{b'} - c') \, d\omega^*$ can be written as

$$\int_0^\infty x' dy' = x'y'|_0^\infty - \int_0^\infty y' dx' \qquad (A3.7)$$

As already defined

$$x'y' = = -\frac{(1+a'\omega^*)}{a'^2} e^{-a'\omega^*} \Phi\left(\frac{\omega^*}{b'} - c'\right)$$

Evaluating at $(0 \; and \; \infty)$ yields

$$x'y' = \frac{1}{a'^2} \Phi(-c') \qquad (A3.8)$$

The second part of (A3.7) is

$$\int_0^\infty y'dx' = \frac{1}{b'a'^2} \int_0^\infty -(1+a'\omega^*) e^{-a'\omega^*} \phi\left(\frac{\omega^*}{b'} - c'\right) d\omega^*$$

Expressing $\phi(.)$ in explicit form yields

$$= \frac{1}{b'a'^2\sqrt{2\pi}} \int_0^\infty -(1+a'\omega^*) e^{-a'\omega^*} e^{-\frac{1}{2}\left(\frac{\omega^{*2}}{b'^2} + c'^2 - 2\frac{\omega^* c'}{b'}\right)^2} d\omega^*$$

$$= -\frac{e^{-\frac{c'^2}{2}}}{b'a'^2\sqrt{2\pi}} \int_0^\infty (1+\omega^* a') e^{-a'\omega^* + \frac{\omega^* c'}{b'}} e^{-\frac{\omega^{*2}}{2b'^2}} d\omega^*$$

The definite integral of the above expression is

$$= \frac{e^{-\frac{c'^2}{2}}}{b'a'^2\sqrt{2\pi}} \left[ \sqrt{\frac{\pi}{2}} b'\left(a'^2 b'^2 - a'b'c' - 1\right) e^{\frac{(c'-a'b')^2}{2}} - \sqrt{\frac{\pi}{2}} b'\left(a'^2 b'^2 - a'b'c' - 1\right) e^{\frac{(c'-a'b')^2}{2}} \operatorname{erf}\left(\frac{a'b'-c'}{\sqrt{2}}\right) - a'b'^2 \right] \quad (A3.9)$$

Simplifying *(A3.9)* we obtain

$$\int_0^\infty y'dx' = \frac{e^{-\frac{c'^2}{2}}}{a'^2} \left[ \left(a'^2 b'^2 - a'b'c' - 1\right) e^{\frac{(c'-a'b')^2}{2}} \Phi(c' - a'b') - \frac{ab}{\sqrt{2\pi}} \right] \qquad (A3.10)$$

Combining (A3.8) and (A3.10) we obtain

$$x'y'|_0^\infty - \int_0^\infty y'dx' = \frac{\Phi(-c')}{a'^2} - \frac{e^{-\frac{c'^2}{2}}}{a'^2} \left[ \left(a'^2 b'^2 - a'b'c' - 1\right) e^{\frac{(c'-a'b')^2}{2}} \Phi(c' - a'b') - \frac{ab}{\sqrt{2\pi}} \right]$$

By substituting $\epsilon$ by $\hat{\epsilon}$, $\mu_v$ by $\hat{\mu}_v$, $\mu_{\omega*}$ by $\hat{\mu}_{\omega^*}$, $\sigma_u$ by $\hat{\sigma}_u$ in the following expression and in $a', b', c'$, now the conditional expectation can be written as

$$E(\omega^*|\hat{\epsilon}) = \frac{1}{f(\hat{\epsilon})\hat{\mu}_v\hat{\mu}_{\omega^*}} e^{\frac{\hat{\epsilon}}{\hat{\mu}_v}+\frac{\sigma_u^2}{2\hat{\mu}_v^2}} \left\{ \frac{\Phi(-c')}{a'^2} - \frac{e^{-\frac{c'^2}{2}}}{a'^2}\left[\left(a'^2b'^2 - a'b'c' - 1\right)e^{\frac{(c'-a'b')^2}{2}}\Phi(c'-a'b') - \frac{ab}{\sqrt{2\pi}}\right]\right\}$$

*Derivation of* $E(v|\hat{\epsilon})$:

For notational convenience we first derive $E(v|\epsilon)$ and then substitute $\epsilon$ by $\hat{\epsilon}$ to obtain $E(v|\hat{\epsilon})$.

The first step to derive $E(v|\hat{\epsilon})$ is to find the conditional density $f(v,\hat{\epsilon})$. Since $\omega^*, v$ and $u$ are statistically independent

$$f(\omega^*, v, u) = f(\omega^*)f(v)f(u) \qquad (A3.11)$$

Substituting $u = \epsilon + v - \omega^*$ in (A3.11) yields

$$f(\omega^*, v, \epsilon) = \frac{1}{\mu_v\mu_{\omega^*}\sigma_u\sqrt{2\pi}} e^{-\frac{v}{\mu_v}-\frac{(\epsilon+v)^2}{2\sigma_u^2}} e^{-\frac{\omega^*}{\mu_{\omega^*}}-\frac{[\omega^{*2}-2\omega^*(\epsilon+v)]}{2\sigma_u^2}}$$

$$f(v,\epsilon) = \frac{1}{\mu_v\mu_{\omega^*}\sigma_u\sqrt{2\pi}} e^{-\frac{v}{\mu_v}-\frac{(\epsilon+v)^2}{2\sigma_u^2}} \int_0^\infty e^{-\frac{\omega^*}{\mu_{\omega^*}}-\frac{[\omega^{*2}-2\omega^*(\epsilon+v)]}{2\sigma_u^2}} d\omega^*$$

Let $k = \frac{1}{\mu_{\omega^*}} - \frac{(\epsilon+v)}{\sigma_u^2}, l = \frac{1}{2\sigma_u^2}$

$$f(v,\epsilon) = \frac{1}{\mu_v\mu_{\omega^*}\sigma_u\sqrt{2\pi}} e^{-\frac{v}{\mu_v}-\frac{(\epsilon+v)^2}{2\sigma_u^2}} \int_0^\infty e^{-k\omega^*} e^{-\frac{\omega^{*2}}{b}} d\omega^*$$

$$= \frac{1}{\mu_v\mu_{\omega^*}\sigma_u\sqrt{2\pi}} e^{-\frac{v}{\mu_v}-\frac{(\epsilon+v)^2}{2\sigma_u^2}} \frac{1}{2}\sqrt{\pi}\sqrt{l}e^{\frac{k^2 l}{4}}\text{erf}\left(\frac{kl+2\omega^*}{2\sqrt{l}}\right)|_0^\infty$$

$$= \frac{1}{\mu_v\mu_{\omega^*}\sigma_u\sqrt{2\pi}} e^{-\frac{v}{\mu_v}-\frac{(\epsilon+v)^2}{2\sigma_u^2}} \frac{1}{2}\sqrt{\pi\sqrt{l}}e^{\frac{k^2 l}{4}}\left(1-\text{erf}\left(\frac{k\sqrt{l}}{2}\right)\right) \qquad (A3.12)$$

Simplifying (A3.12) yields

$$f(v,\epsilon) == \frac{1}{\mu_v\mu_{\omega^*}} e^{-v(\frac{1}{\mu_v}+\frac{1}{\mu_{\omega^*}})} e^{\frac{\sigma_u^2}{2\mu_{\omega^*}^{2}}-\frac{\epsilon}{\mu_{\omega^*}}} \Phi\left(\frac{v}{\sigma_u}+\left(\frac{\epsilon}{\sigma_u}-\frac{\sigma_u}{\mu_{\omega^*}}\right)\right)$$

Now one can define the conditional expectation of $v$ as

$$E(v|\epsilon) = \frac{1}{f(\epsilon)} \int_0^{\infty} v\, f(v,\epsilon) dv$$

$$\rightarrow E(v|\epsilon) = \frac{1}{f(\epsilon)} \int_0^{\infty} v \frac{1}{\mu_v \mu_{\omega^*}} e^{-v(\frac{1}{\mu_v}+\frac{1}{\mu_{\omega^*}})} e^{\frac{\sigma_u^2}{2\mu_{\omega^*}{}^2}-\frac{\epsilon}{\mu_{\omega^*}}} \Phi\left(\frac{v}{\sigma_u} + \left(\frac{\epsilon}{\sigma_u} - \frac{\sigma_u}{\mu_{\omega^*}}\right)\right) dv$$

$$\rightarrow E(v|\epsilon) = \frac{1}{f(\epsilon)} \frac{1}{\mu_v \mu_{\omega^*}} e^{\frac{\sigma_u^2}{2\mu_{\omega^*}{}^2}-\frac{\epsilon}{\mu_{\omega^*}}} \int_0^{\infty} v\, e^{-v\left(\frac{1}{\mu_v}+\frac{1}{\mu_{\omega^*}}\right)} \Phi\left(\frac{v}{\sigma_u} + \left(\frac{\epsilon}{\sigma_u} - \frac{\sigma_u}{\mu_{\omega^*}}\right)\right) dv \qquad (A3.13)$$

Let $a' = \left(\frac{1}{\mu_v} + \frac{1}{\mu_{\omega^*}}\right)$; $b' = \sigma_u$; $d' = \left(\frac{\epsilon}{\sigma_u} - \frac{\sigma_u}{\mu_{\omega^*}}\right)$; $\theta_2' = \frac{1}{f(\epsilon)} \frac{1}{\mu_v \mu_{\omega^*}} e^{\frac{\sigma_u^2}{2\mu_{\omega^*}{}^2}-\frac{\epsilon}{\mu_{\omega^*}}}$ Utilizing these re-parameterizations in (A3.13) yields

$$E(v|\epsilon) = \theta_2' \int_0^{\infty} v\, e^{-a'v}\, \Phi\left(\frac{v}{b'} + d'\right) dv \qquad (A3.14)$$

The above integral can be solved by integration by parts

Let $x'' = \Phi\left(\frac{v}{b'} + d'\right) \Rightarrow dx'' = \frac{1}{b'} \phi\left(\frac{v}{b'} + d'\right) dv$; $dy'' = v e^{-a'v} dv \Rightarrow y'' = -\frac{(1+a'v)}{a'^2} e^{-a'v}$

$$\int_0^{\infty} x'' dy'' = x''y''|_0^{\infty} - \int_0^{\infty} y'' dx''$$

Evaluating the first term on the R.H.S.

$$x''y''|_0^{\infty} = \frac{\Phi(d')}{2a'^2}$$

The second term

$$\int_0^{\infty} y'' dx'' = \int_0^{\infty} -\frac{(1+a'v)}{a'^2} e^{-a'v} \frac{1}{b'} \phi\left(\frac{v}{b'} + d'\right) dv$$

$$= \frac{e^{-\frac{d'^2}{2}}}{a'^2 b'} \int_0^{\infty} -(1+a'v) e^{-v(a'+\frac{d'}{b'})} e^{2\frac{v^2}{2b'^2}} dv$$

Integrating yields

$$\int_0^{\infty} y'' dx'' = \frac{e^{-\frac{d'^2}{2}}}{a'^2} \left[ (a'^2 b'^2 + a'b'd' - 1) e^{\frac{(a'b'+d')^2}{2}} \Phi(-a'b' - d') \right] - \frac{a'b'}{\sqrt{2\pi}}$$

By substituting $\epsilon$ by $\hat{\epsilon}$, $\mu_v$ by $\hat{\mu}_v$, $\mu_{\omega*}$ by $\hat{\mu}_{\omega^*}$, $\sigma_u$ by $\hat{\sigma}_u$ in the following expression and in $a', b', d'$, the conditional expectation can be written as

$$E(v|\hat{\epsilon}) = \theta_2' \left[ x''y'' - \int_0^\infty y''dx'' \right]$$

$$\rightarrow E(v|\hat{\epsilon}) = \frac{1}{f(\hat{\epsilon})} \frac{1}{\hat{\mu}_v \mu_{\omega^*}} e^{\frac{\sigma_u^2}{2\hat{\mu}_{\omega^*}{}^2} - \frac{\hat{\epsilon}}{\hat{\mu}_{\omega^*}}} \left[ \frac{\Phi(d')}{a'^2} - \frac{e^{-\frac{d'^2}{2}}}{a'^2} \left\{ \left(a'^2 b'^2 + a'b'd' - 1\right) e^{\frac{\left(a'b'+d'\right)^2}{2}} \Phi(-a'b' - d') - \frac{a'b'}{\sqrt{2\pi}} \right\} \right]$$