The Stata Journal (2023)
23, Number 1, pp. 53–85 DOI: 10.1177/1536867X231161976

Finite mixture models for linked survey and administrative data: Estimation and postestimation

Stephen P. Jenkins
Department of Social Policy
London School of Economics and Political Science
London, U.K.
s.jenkins@lse.ac.uk

Fernando Rios-Avila
The Levy Economics Institute of Bard College
Annandale-on-Hudson, NY
friosavi@levy.org

Abstract. Researchers use finite mixture models to analyze linked survey and administrative data on labor earnings, while also accounting for various types of measurement error in each data source. Different combinations of error-ridden and error-free observations characterize latent classes. Latent class probabilities depend on the probabilities of the different types of error. We introduce a suite of commands to fit finite mixture models to linked survey-administrative data: there is a general model and seven simpler variants. We also provide postestimation commands for assessment of reliability, marginal effects, data simulation, and prediction of hybrid variables that combine information from both data sources about the outcome of interest. Our commands can also be used to study measurement errors in other variables besides labor earnings.

Keywords: st0701, ky_fit, ky_estat, ky_sim, linked survey and administrative data, measurement error, finite mixture models, latent class models

1 Introduction

Linked datasets are datasets in which reports by respondents to a household survey on a variable such as labor earnings are linked to reports on the same variable in an administrative dataset (for example, income tax or Social Security Administration data) for the same respondents. Researchers have long used linked datasets to examine measurement errors in the variables of interest—to investigate whether they impart bias in the observed measures, how much spurious variation they account for, and whether errors are correlated with the "true" measure (a negative correlation means that low-earners overreport and high earners underreport). In the first generation of studies, analysts assumed that the linked administrative data provided error-free measures; all measurement errors arose in the survey reports. A selective list of examples of first-generation studies is Bound and Krueger (1991; about the United States), Bollinger

(1998; United States), Kristensen and Westergaard-Nielsen (2007; Denmark), and Angel et al. (2019; Austria). A small and more recent second generation of studies (cited later in this section) has allowed for errors in the administrative data as well. This article is a methodological contribution to second-generation studies: we provide commands to fit a wide range of models. The models can also be applied to variables other than earnings.

The statistical models underpinning virtually all second-generation studies are finite mixture models (FMMs), also known as latent class models. The key idea is that true earnings for an individual is unobserved but there are two observed earnings measures available, one from the household survey data and one from the linked administrative data. Both measures are subject to errors of various types (as explained in section 2), though not all individuals experience all types of error. We can classify individuals into a finite number of groups (latent classes) according to which types of error their earnings measures contain. Observed earnings are a combination (“mixture”) of the distributions for the latent classes. In sum, the FMMs used in second-generation studies succinctly describe both the distribution of the “true” (error-free) substantive variable of interest and the distributions of each of the latent classes and associated class membership probabilities.

These FMMs cannot be fit using readily available software such as Stata’s `fmm` suite of commands because of their specialist nature, and we are unaware of suitable community-contributed programs for Stata or other software. In this article, we provide and illustrate commands for fitting a general class of FMMs to linked data.[1] We also provide postestimation commands for assessment of reliability, marginal effects, data simulation, and prediction of hybrid variables that combine information from both data sources about the outcome of interest. The outcome of interest may be a variable other than labor earnings, as we discuss in section 5.

The FMMs we propose are generalizations of the second-generation models developed by Kapteyn and Ypma (2007; KY hereafter). KY’s model was the first to incorporate administrative data error in addition to survey measurement error. However, the characterization of administrative data error was restricted to linkage “mismatch”; that is, the situation in which an individual’s survey response is incorrectly linked to the response for some other person in the administrative data. KY’s findings, based on linked earnings data for Swedish individuals aged 50-plus, showed that even a small amount of mismatch error was consequential (their linked administrative data were less reliable than their survey data), and they found no evidence that low earners overreported and high earners underreported their earnings (a striking contrast with the findings of first-generation studies). However, KY did not consider measurement error per se

[1] More generally, FMMs can take many forms: see, for example, the semiparametric heterogeneity model of Heckman and Singer (1984) or the latent class models as discussed by Aitkin and Rubin (1985). For a textbook overview of conventional FMMs, see Cameron and Trivedi (2005, sec. 18.5).

in the administrative data, that is, error arising in its compilation (typically involving reporting by employers to tax or Social Security authorities).[2]

In our companion article (Jenkins and Rios-Avila Forthcoming), we extend KY's model to more general FMMs that include administrative measurement error in addition to linkage mismatch and survey measurement error. This is our first innovation. Our second is to allow the parameters describing the distributions in our FMMs to vary with individual characteristics. This introduces greater flexibility and hence potentially better fits to data. It also provides a succinct way to address substantive questions such as the following: Does survey earnings measurement error differ between older and younger workers? How does administrative data error differ between private- and public-sector employees? Our third contribution is to extend the methods for earnings prediction proposed by Meijer, Rohwedder, and Wansbeek (2012; MRW hereafter) to our more general models. MRW derived formulas for a number of hybrid earnings predictors that combined information from both survey and administrative data, and showed that they were more reliable than either the survey or the administrative data measure. However, MRW's illustrations focused entirely on KY's model and their estimates based on Swedish data.[3]

By comparison with Jenkins and Rios-Avila (Forthcoming), the current article focuses on the software development side of our work. As we explain in sections 2 and 3, our general approach encompasses eight model specifications, ranging from model 1 (basic) through the most general model 8. The empirical examples in this article relate to models 1–4 (model 4 is KY's most general model). Jenkins and Rios-Avila's (Forthcoming) substantive application uses U.K. linked data on employment earnings for individuals of all ages and focuses discussion on estimates from fitting models 4, 5, 7, and 8.

In section 2, we describe our FMMs and explain how to fit them using maximum likelihood. In section 3, we present our new commands for estimation and postestimation analysis. In section 4, we illustrate the commands, drawing on KY's and MRW's empirical analyses, and confirm that our commands reproduce their estimates. Section 5 contains conclusions. The appendix contains additional results that we draw on in the main text.

[2] A small number of second-generation studies allow for administrative data error in earnings: see Abowd and Stinson (2013; using data for the United States), Bingley and Martinello (2017; Denmark), Hyslop and Townsend (2020; New Zealand), and Bollinger et al. (2018; United States), who also allow for linkage mismatch. Jenkins and Rios-Avila (2020) fit KY models to linked data for the United Kingdom. Jenkins and Rios-Avila (Forthcoming) review first- and second-generation studies in more detail.

[3] Our replication of MRW's analysis using U.K. linked data (Jenkins and Rios-Avila 2021) was also restricted to KY models.

2 FMMs for linked survey and administrative data

We set out our FMMs in this section and assume that the variable of interest is the logarithm of the labor earnings of employees ("earnings"). For each of many individuals in a linked dataset, we have an observation pair referring to the worker's earnings derived from the survey data and from the administrative data.

We assume, following KY, that there is a latent variable ξ_i that represents the true variable of interest (log earnings) for each individual $i = 1, \ldots, N$. This variable is not observed directly, but there are two measures of it, each potentially error ridden: one from administrative data, r_i, and one from survey data, s_i.

2.1 Administrative data: Three types of observations

We assume the administrative data are a mixture of three types of observations. First, we distinguish between observations for whom the record linkage between administrative and survey data is correct, which occurs with probability π_r, and observations who are mismatched, with probability $1 - \pi_r$. The administrative data measure for mismatched observations is ζ_i, the earnings of some other person in the administrative data. Second, among the correctly matched observations, we assume that the administrative data earnings measure is error free with probability π_υ or contains measurement error υ_i with probability $1 - \pi_\upsilon$. (KY assumed $\pi_\upsilon = 1$.) In the case with measurement error, errors may be correlated with true earnings with the correlation denoted by ρ_r. If $\rho_r < 0$, we have mean-reverting errors: high earners underreport and low earners overreport; if $\rho_r > 0$, the reverse occurs. The three types of observations, labeled R1, R2, and R3, are summarized in (1).

$$r_i = \begin{cases} \xi_i & \text{with probability } \pi_r \pi_\upsilon & \text{(R1)} \\ \xi_i + \rho_r(\xi_i - \mu_\xi) + \upsilon_i & \text{with probability } \pi_r(1 - \pi_\upsilon) & \text{(R2)} \\ \zeta_i & \text{with probability } 1 - \pi_r & \text{(R3)} \end{cases} \tag{1}$$

2.2 Survey data: Three types of observations

We assume the survey data are a mixture of three types of observations (following KY). Type S1 respondents are those who report their true earnings: s_i equals true latent earnings ξ_i with probability π_s. The survey earnings of type S2 respondents differ from true earnings by a measurement error component representing noise (η_i), plus a mean-reversion component allowing for a correlation (ρ_s) between true earnings and error. A third type, S3, contains observations with error-ridden survey earnings (as for type S2), except that there is additional "contamination" (ω_i).[4] The probability of contamination is π_ω. Type S2 occurs with probability $(1-\pi_s)(1-\pi_\omega)$; type S3 occurs with probability $(1-\pi_s)\pi_\omega$. The three types of observation are summarized in (2).

$$s_i = \begin{cases} \xi_i & \text{with probability } \pi_s \quad \text{(S1)} \\ \xi_i + \rho_s(\xi_i - \mu_\xi) + \eta_i & \text{with probability } (1-\pi_s)(1-\pi_\omega) \quad \text{(S2)} \\ \xi_i + \rho_s(\xi_i - \mu_\xi) + \eta_i + \omega_i & \text{with probability } (1-\pi_s)\pi_\omega \quad \text{(S3)} \end{cases} \tag{2}$$

In sum, observations in the linked dataset are a mixture of nine types (latent classes $j = 1, \ldots, 9$) depending on the combination of administrative and survey observation types. The latent class probabilities are π_j, $j = 1, \ldots, 9$. For example, group 1 contains observations with the combination (R1, S1) with probability $\pi_1 = \pi_r \pi_\upsilon \pi_s$, group 2 contains observations with the combination (R1, S2) with probability $\pi_2 = \pi_r \pi_\upsilon (1 - \pi_s)(1-\pi_\omega)$, etc. The FMM specification is completed by assumptions about the latent class earnings densities, $f_j(r_i, s_i)$ for each $j = 1, \ldots, 9$.

We assume that true earnings (ξ_i), mismatched earnings (ζ_i), and errors (υ_i, η_i, ω_i) are each normally distributed with the exception that true earnings and reference period errors (ω_i) are bivariate normal. We assume normality (as other researchers do) to fit models by maximum likelihood (see below) and because it facilitates postestimation derivations.

The distributions are identically distributed and mutually independent (assumptions we relax shortly). Thus, the distributions of the factors may be written as

$$\begin{pmatrix} \xi_i \\ \omega_i \end{pmatrix} = N\left(\begin{pmatrix} \mu_\xi \\ \mu_\omega \end{pmatrix}, \begin{pmatrix} \sigma_\xi^2 & \rho_\omega \sigma_\xi \sigma_\omega \\ \rho_\omega \sigma_\xi \sigma_\omega & \sigma_\omega^2 \end{pmatrix} \right)$$
$$\zeta_i \sim N\left(\mu_\zeta, \sigma_\zeta^2\right), \quad \eta_i \sim N\left(\mu_\eta, \sigma_\eta^2\right), \quad \text{and} \quad \upsilon_i \sim N\left(\mu_\upsilon, \sigma_\upsilon^2\right)$$

where μ and σ denote mean and standard deviation (SD), respectively, and ρ_ω is the correlation between true earnings and contamination. Jenkins and Rios-Avila (Forthcoming) argue there are grounds for expecting $\rho_\omega < 0$. (KY assumed $\rho_\omega = 0$.) We do not restrict error means to equal zero, because errors may introduce systematic bias.

Table 1 summarizes the nine latent classes, their probabilities, and densities.

[4] Kapteyn and Ypma (2007) state that contamination error "can be the result of erroneously reporting income as annual, whereas the amount is a monthly amount, or vice versa, omitting a second job or working only part of the year" (2007, 528). Jenkins and Rios-Avila (Forthcoming) relabel contamination error as reference period error because, in their U.K. application, a particularly important reason for potential differences between survey and administrative data observations is that the reference period for earnings used by the survey differs from the reference period in the administrative data.

Table 1. Latent class probabilities and distributions

Label, j	Combination	Latent class probability, π_j	Latent class distribution densities, $f_j(r_i, s_i)$
1	R1,S1	$\pi_1 = \pi_r \pi_\upsilon \pi_s$	$N\left(\begin{pmatrix}\mu_\xi \\ \mu_\xi\end{pmatrix}, \begin{pmatrix}\sigma_\xi^2 & 1 \\ 1 & \sigma_\xi^2\end{pmatrix}\right)$
2	R1,S2	$\pi_2 = \pi_r \pi_\upsilon (1-\pi_s)(1-\pi_\omega)$	$N\left(\begin{pmatrix}\mu_\xi \\ \mu_\xi + \mu_\eta\end{pmatrix}, \begin{pmatrix}\sigma_\xi^2 & (1+\rho_s)\sigma_\xi^2 \\ (1+\rho_s)\sigma_\xi^2 & (1+\rho_s)^2\sigma_\xi^2 + \sigma_\eta^2\end{pmatrix}\right)$
3	R1,S3	$\pi_3 = \pi_r \pi_\upsilon (1-\pi_s)\pi_\omega$	$N\left(\begin{pmatrix}\mu_\xi \\ \mu_\xi + \mu_\eta + \mu_\omega\end{pmatrix},\right.$ $\left.\begin{pmatrix}\sigma_\xi^2 & (1+\rho_s)\sigma_\xi^2 + \rho_\omega\sigma_\xi\sigma_\omega \\ (1+\rho_s)\sigma_\xi^2 + \rho_\omega\sigma_\xi\sigma_\omega & (1+\rho_s)^2\sigma_\xi^2 + \sigma_\eta^2 + \sigma_\omega^2 + 2\rho_\omega\sigma_\xi\sigma_\omega\end{pmatrix}\right)$
4	R2,S1	$\pi_4 = \pi_r(1-\pi_\upsilon)\pi_s$	$N\left(\begin{pmatrix}\mu_\xi + \mu_\upsilon \\ \mu_\xi\end{pmatrix}, \begin{pmatrix}(1+\rho_r)^2\sigma_\xi^2 + \sigma_\upsilon^2 & (1+\rho_r)\sigma_\xi^2 \\ (1+\rho_r)\sigma_\xi^2 & \sigma_\xi^2\end{pmatrix}\right)$
5	R2,S2	$\pi_5 = \pi_r(1-\pi_\upsilon)(1-\pi_s)$ $(1-\pi_\omega)$	$N\left(\begin{pmatrix}\mu_\xi + \mu_\upsilon \\ \mu_\xi + \mu_\eta\end{pmatrix}, \begin{pmatrix}(1+\rho_r)^2\sigma_\xi^2 + \sigma_\upsilon^2 & (1+\rho_r)(1+\rho_s)\sigma_\xi^2 \\ (1+\rho_r)(1+\rho_s)\sigma_\xi^2 & (1+\rho_s)^2\sigma_\xi^2 + \sigma_\eta^2\end{pmatrix}\right)$
6	R2,S3	$\pi_6 = \pi_r(1-\pi_\upsilon)(1-\pi_s)\pi_\omega$	$N\left(\begin{pmatrix}\mu_\xi + \mu_\upsilon \\ \mu_\xi + \mu_\eta + \mu_\omega\end{pmatrix},\right.$ $\left.\begin{pmatrix}(1+\rho_r)\sigma_\xi^2 + \sigma_\upsilon^2 & (1+\rho_r)(1+\rho_s)\sigma_\xi^2 + (1+\rho_r)\rho_\omega\sigma_\xi\sigma_\omega \\ (1+\rho_r)(1+\rho_s)\sigma_\xi^2 + (1+\rho_r)\rho_\omega\sigma_\xi\sigma_\omega & (1+\rho_s)^2\sigma_\xi^2 + \sigma_\eta^2 + \sigma_\omega^2 + 2\rho_\omega\sigma_\xi\sigma_\omega\end{pmatrix}\right)$
7	R3,S1	$\pi_7 = (1-\pi_r)\pi_s$	$N\left(\begin{pmatrix}\mu_\zeta \\ \mu_\xi\end{pmatrix}, \begin{pmatrix}\sigma_\zeta^2 & 0 \\ 0 & \sigma_\varepsilon^2\end{pmatrix}\right)$
8	R3,S2	$\pi_8 = (1-\pi_r)(1-\pi_s)(1-\pi_\omega)$	$N\left(\begin{pmatrix}\mu_\zeta \\ \mu_\xi + \mu_\eta\end{pmatrix}, \begin{pmatrix}\sigma_\zeta^2 & 0 \\ 0 & (1+\rho_s)^2\sigma_\varepsilon^2 + \sigma_\eta^2\end{pmatrix}\right)$
9	R3,S3	$\pi_9 = (1-\pi_r)(1-\pi_s)\pi_\omega$	$N\left(\begin{pmatrix}\mu_\zeta \\ \mu_\xi + \mu_\eta + \mu_\omega\end{pmatrix}, \begin{pmatrix}\sigma_\zeta^2 & 0 \\ 0 & (1+\rho_s)^2\sigma_\xi^2 + \sigma_\eta^2 + \sigma_\omega^2 + 2\rho_\omega\sigma_\xi\sigma_\omega\end{pmatrix}\right)$

We allow distributions to vary with observed characteristics by writing transformations of model parameters as linear indices of characteristics; that is,

$$G(\gamma_i) = \boldsymbol{\beta}'_\gamma \mathbf{x}_i \tag{3}$$

For each model parameter with generic label γ, where $\mathbf{x}_i$ is a vector of observed characteristics for individual i, including a constant. Transformation function $G(\cdot)$ is the identity function for means (μ), the logarithmic function for SDs (σ), the logistic function for probabilities (π), and Fisher's z transformation for correlations (ρ).[5] See the next section for further details. Some previous research has allowed the mean of true earnings (μ_ξ) to vary with characteristics but not other model parameters. Allowing measurement error distributions to differ across individuals has two advantages. The increased flexibility can improve model fit to data, and researchers can answer substantive questions by examining whether there are differences in parameters (and thence error distributions) across different groups, as stated in the *Introduction*.

The discussion so far refers to our most general model, which we label model 8. Simpler versions of our general model (models 1–7) can be fit using our estimation commands, as we explain below, including several of KY's models.

2.3 Estimation

We fit the FMM by maximum likelihood. The general expression for the log-likelihood function of our finite mixture is

$$\log\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\Pi}) = \sum_{i=1}^{N} \log \sum_{j=1}^{9} \pi_j f_j(r_i, s_i|\boldsymbol{\theta})$$

where we now write each latent class density as conditional on the set of parameters, $\boldsymbol{\theta}$, that describe the bivariate distributions and $\boldsymbol{\Pi} = \{\pi_r, \pi_s, \pi_\upsilon, \pi_\omega\}$ are the error probabilities that characterize the class probabilities π_j.

The FMM is identified by the assumptions about the relationships between the two observed measures and true earnings and the nonnormal error structure arising from the mixture of distributions: see Kapteyn and Ypma (2007, 532). See also Yakowitz and Spragins (1968), who prove that finite mixtures are identifiable if the mixture is of multivariate Gaussian distributions, which is the case here. Observe too that, although there are nine latent class probabilities, these depend on only four parameters (see table 1).

The definition of the first latent class (group 1) also plays an important role. Identification uses the assumption that the members of class 1 are "completely labeled" (as KY term it). These individuals correctly report their earnings in the survey data, are correctly matched to their administrative data records, and there is no error in their administrative earnings. Hence, both observed earnings measures equal true earnings;

[5] Reversion to the mean in the models with a heterogeneous mean earnings function refers to reversion to the mean among individuals with the same observed characteristics.

that is, $r_i = s_i = \xi_i$ if $i \in$ class 1. This assumption has two consequences for the log-likelihood function (Redner and Walker 1984).

First, because $r_i = s_i$, the class 1 distribution degenerates to a univariate normal distribution with mean μ_ξ and variance σ_ξ^2. Second, because class membership is known for observations in this group, the log-likelihood function becomes

$$\log\mathcal{L}_i(\boldsymbol{\theta}, \boldsymbol{\Pi}) = \sum_{i \in \text{class } 1} \pi_1 \log\{f_1(\xi_i|\boldsymbol{\theta})\} + \sum_{i \notin \text{class } 1} \log\left(\sum_{j=2}^{9} \pi_j f_j(r_i, s_i|\boldsymbol{\theta})\right)$$

In principle, μ_ξ and σ_ξ^2 are fully identified using the sample of class 1 observations. In practice, the sample of completely labeled observations may be too small for reliable identification of these moments. KY's strategy was to broaden the definition of equality to include observations for which survey and administrative earnings were sufficiently "close". This is an empirical judgment call.[6]

3 The ky suite of commands for estimation and postestimation

This section describes the commands for fitting our general FMM and special cases of it and commands for postestimation analysis and prediction. We assume the linked dataset is in wide format, that is, with one row per individual. There are variables corresponding to r_i and s_i and also (optionally) variables used to define explanatory variables in models with covariates.

3.1 Model estimation: ky_fit

Command `ky_fit` fits the general FMM and special cases of it. The syntax for the command is as follows,

`ky_fit` *r_var s_var* [*cl_var*] [*if*] [*in*] [*weight*] [`, delta(`*#d*`) model(`*#*`)` *options*]

where *r_var* and *s_var* are required variables. They correspond to the administrative log earnings measure r_i (*r_var*) and the survey log earnings measure s_i (*s_var*). `aweight`s, `fweight`s, `iweight`s, and `pweight`s are allowed.

[6] In their application, KY defined an observation as completely labeled if earnings in the two data sources differed by less than 1,000 Swedish Kronor (14.8% of their sample). Jenkins and Rios-Avila (2020), using U.K. data, assess the sensitivity of parameter estimates to different assumptions, varying the fraction of completely labeled observations from 0.25% to 16.93%, finding small differences for estimates the latent variable distributions, but some larger effects on estimates of the probability of correctly reporting earnings in the survey (π_s).

Optionally, you can refer to a binary variable *cl_var* that identifies observations that belong to the completely labeled class. If *cl_var* is not declared, `ky_fit` creates a binary indicator variable named `__ll__` equal to one for observations for which `abs(`*r_var* `-` *s_var*`) <= delta(`*#d*`)`. The default is `delta(0)`, but other values can be declared using `delta(`*#d*`)`.

`model(`*#*`)` specifies which version of the FMM is fit. Table 2 lists the model variants available, showing for each model the parameter restrictions imposed relative to the most general model and the combinations of types of observations present in the administrative and survey data. The default is `model(1)`, which assumes error-free administrative data plus mean-reverting errors in the survey data (but without contamination). The classical measurement error model is model 1 with $\mu_\eta = 0$ and without mean-reverting errors. The most general model, described in section 2, corresponds to model 8. KY's "full" model is model 4. Jenkins and Rios-Avila (Forthcoming) focus on models 4, 5, 7, and 8; model 5 is the best-fitting model in their application.

Table 2. FMM variants and parameter restrictions

Model	Parameter restrictions	Types of observation	
#		Administrative data	Survey data
1	$\mu_\omega = 0$; $\sigma_\omega = 0$; $\pi_\omega = 0$; $\mu_\upsilon = 0$; $\sigma_\upsilon = 0$; $\pi_\upsilon = 1$; $\mu_\zeta = 0$; $\sigma_\zeta = 0$; $\pi_r = 1$; $\rho_r = 0$; $\rho_\omega = 0$	R1	S1, S2
2	$\mu_\upsilon = 0$; $\sigma_\upsilon = 0$; $\pi_\upsilon = 1$; $\mu_\zeta = 0$; $\sigma_\zeta = 0$; $\pi_r = 1$; $\rho_r = 0$; $\rho_\omega = 0$	R1	S1, S2, S3
3	$\mu_\upsilon = 0; \sigma_\upsilon = 0; \pi_\upsilon = 1; \rho_r = 0;$ $\mu_\omega = 0$; $\sigma_\omega = 0$; $\pi_\omega = 0$; $\rho_\omega = 0$	R1, R2	S1, S2
4	$\mu_\upsilon = 0$; $\sigma_\upsilon = 0$; $\pi_\upsilon = 1$; $\rho_r = 0$; $\rho_\omega = 0$	R1, R3	S1, S2, S3
5	$\rho_\omega = 0$	R1, R2, R3	S1, S2, S3
6	$\mu_\omega = 0$; $\sigma_\omega = 0$; $\pi_\omega = 0$; $\rho_\omega = 0$	R1, R2, R3	S1, S2
7	$\mu_\upsilon = 0$; $\sigma_\upsilon = 0$; $\pi_\upsilon = 1$; $\rho_r = 0$	R1, R3	S1, S2, S3
8	No restrictions	R1, R2, R3	S1, S2, S3

Optionally, you can specify the parameters of any of the models listed in table 2 as functions of covariates, as described by (3). Table 3 provides a walk-through of the estimated parameters, the parameter-specific options in `ky_fit` for declaring covariates, and the internal transformation used for maximization. If a model-specific parameter is constrained (as described by table 2), a declaration of covariates for that parameter is ignored. Because parameters (apart from means) are fit in a transformed metric, they need to be back-transformed to see them in their "natural" metric, and `margins` does this: see section 3.3.

Table 3. Options to allow parameters to be functions of covariates

Parameter	`ky_fit` option	Transformation
μ_ξ	`mu_e(`*varlist*`)`	Identity
σ_ξ	`ln_sig_e(`*varlist*`)`	$\sigma_\xi = \exp($`ln_sig_e()`$)$
μ_ω	`mu_w(`*varlist*`)`	Identity
σ_ω	`ln_sig_w(`*varlist*`)`	$\sigma_\omega = \exp($`ln_sig_w()`$)$
μ_η	`mu_n(`*varlist*`)`	Identity
σ_η	`ln_sig_n(`*varlist*`)`	$\sigma_\eta = \exp($`ln_sig_n()`$)$
μ_υ	`mu_v(`*varlist*`)`	Identity
σ_υ	`ln_sig_v(`*varlist*`)`	$\sigma_\upsilon = \exp($`ln_sig_v()`$)$
μ_ζ	`mu_t(`*varlist*`)`	Identity
σ_ζ	`ln_sig_t(`*varlist*`)`	$\sigma_\zeta = \exp($`ln_sig_t()`$)$
ρ_r	`arho_r(`*varlist*`)`	$\rho_r = \tanh($`arho_r()`$)$
ρ_s	`arho_s(`*varlist*`)`	$\rho_s = \tanh($`arho_s()`$)$
ρ_ω	`arho_w(`*varlist*`)`	$\rho_\omega = \tanh($`arho_w()`$)$
π_r	`lpi_r(`*varlist*`)`	$\pi_r = \text{logistic}($`lpi_r()`$)$
π_s	`lpi_s(`*varlist*`)`	$\pi_s = \text{logistic}($`lpi_s()`$)$
π_ω	`lpi_w(`*varlist*`)`	$\pi_\omega = \text{logistic}($`lpi_w()`$)$
π_υ	`lpi_v(`*varlist*`)`	$\pi_\upsilon = \text{logistic}($`lpi_v()`$)$

Our code fits models sequentially using `ml`: we use the parameter estimates of simpler (more restricted) models as starting values for more flexible models. Additional restrictions on model specifications can be applied using `constraint()`. If you want to use other initial values, `ml` options `search()` and `repeat()` are available. You can also provide specific initial values for model parameters using option `from()`.

We recommend that you experiment with multiple sets of initial values to check that the more complex models converge to a global maximum rather than some local maximum. This is a well-known issue for FMM models and occasionally arose in our own work (Jenkins and Rios-Avila Forthcoming) when fitting models 4–8 with many covariates. Our sequential fitting approach reduces the risk of convergence to local maximums but cannot remove it altogether (that is impossible).

`ky_fit` also allows the maximization options `technique()`, `trace`, and `difficult`.

`ky_fit` reports standard errors derived from asymptotic theory by default. Optionally, you may use `robust` and `cluster(`*cluster_var*`)`.

3.2 Postestimation tools: ky_estat

`ky_estat` is a postestimation command that provides summary statistics for a fitted model. It is integrated with Stata's built-in postestimation command `estat` and has the following syntax:

```
estat [pr_t pr_j pr_sr pr_all reliability xirel] [, sim reps(#)
    seed(#) ]
```

`pr_t` reports error probabilities π_r, π_s, π_υ, and π_ω.

`pr_j` reports latent class probabilities π_1 through π_9.

`pr_sr` reports the probabilities of each observation type S1–S3 and R1–R3.

`pr_all` reports all probabilities.

For models without covariates, `estat` reports error probabilities in their original metric (rather than the metric used for estimation). If you specify error probabilities as functions of covariates, `estat` reports average predicted probabilities.

If the error probabilities are modeled without covariates, `reliability` produces a full report of all unconditional probabilities. It also reports two reliability summary statistics for each of the survey and administrative data, based on the analytically predicted variances of the observed earnings data (r_i, s_i), and their covariances with (model-specific) estimated true latent earnings (ξ_i). The two reliability statistics are

$$R_1^r = \frac{\text{Cov}(\xi_i, r_i)}{\text{Var}(r_i)}; \qquad R_1^s = \frac{\text{Cov}(\xi_i, s_i)}{\text{Var}(s_i)}$$

$$R_2^r = \frac{\text{Cov}(\xi_i, r_i)^2}{\text{Var}(\xi_i)\text{Var}(r_i)}; \quad R_2^s = \frac{\text{Cov}(\xi_i, s_i)^2}{\text{Var}(\xi_i)\text{Var}(s_i)}$$

R_1 is analogous to the reliability statistic reported for the classical measurement error model with mean reversion and is equal to the slope coefficient from a (hypothetical) regression of true earnings on the observed earnings measure (Bound and Krueger 1991, 9). Its values may be greater than one. R_2, a more conventional psychometric measure of reliability (and used by MRW), is the squared correlation between true earnings and an observed earnings measure. We present analytical expressions for unconditional variance and covariances for model 8 in the appendix. Expressions for simpler model variants are special cases of these.

If you model error probabilities as functions of covariates, `reliability` produces simulation-based reliability estimates. Use option `reps(#)` to specify the number of replications (the default is `reps(50)`). For reproducibility, set the seed using `seed(#)`.

You can also request simulation-based reliability statistics using option `sim` even if error probabilities have not been declared as functions of covariates.

The final postestimation subcommand is `xirel`. This uses simulated data to estimate the reliability statistics, mean squared error (MSE), bias, and variance of bias of the seven latent earnings predictors proposed by MRW (see the next section). This option also produces corresponding statistics for the observed administrative and survey measures. Use `reps(#)` and `seed(#)` to set the number of replications and seed.

3.3 Postestimation predictions and marginal effects: ky_p

ky_p is a postestimation program for obtaining predictions for all relevant parameters of FMMs and is integrated with Stata's postestimation commands predict and margins. Table 4 lists the options available. The analytical formulas for the constructed moments correspond to those listed in table 1.

Table 4. ky_p options compatible with predict and margins

Option	Description
Structural parameters	
mean_e, mean_n, mean_w, mean_t	Conditional means of latent variables ξ, η, ω, and ζ, respectively
sig_e, sig_n, sig_w, sig_t	Conditional SDs of latent variables ξ, η, ω, and ζ, respectively
pi_s, pi_r, pi_w, pi_v	Error probabilities
rho_s, rho_r	Mean-reversion parameters for survey data (ρ_s) and administrative data (ρ_r)
rho_w	Conditional correlation between latent true earnings (ξ) and contamination (ω)
Constructed moments	
mean_r1, mean_r2, mean_r3	Mean of administrative earnings: R1, R2, R3, respectively
sig_r1, sig_r2, sig_r3	SD of administrative earnings: R1, R2, R3, respectively
pi_r1, pi_r2, pi_r3	Probability of belonging to type R1, R2, R3, respectively
mean_s1, mean_s2, mean_s3	Mean of survey earnings: S1, S2, S3, respectively
sig_s1, sig_s2, sig_s3	SD of survey earnings: S1, S2, S3, respectively
pi_s1, pi_s2, pi_s3	Probability of belonging to type S1, S2, S3, respectively
pj_1, ..., pj_9	Probability of belonging to latent class $j = 1, \ldots, 9$

NOTES: When models 3, 4, and 6 are chosen, mean_r2, sig_r2, and pi_r2 produce estimates for R3 because type R2 observations are absent.

Table 5 lists the options that are compatible with `predict` alone (because they are functions of the variables r_i and s_i), providing a description and definition. They cannot be used with `margins`. The options include predictions of posterior class probabilities and Bayesian classifications based on the posterior probabilities.

Table 5. `ky_p` options compatible with `predict` only

Option	Description	Definition
`pip_r1`, `pip_r2`, `pip_r3`	Posterior probability of belonging to R1, R2, or R3, respectively	$\pi_{R_j}(r_i) = \pi_{R_j} \times \frac{f_{R_j}(r_i \mid \boldsymbol{\theta})}{\sum_{k=1}^{3} f_{R_k}(r_i \mid \boldsymbol{\theta})}$
`pip_s1`, `pip_s2`, `pip_s3`	Posterior probability of belonging to S1, S2, or S3, respectively	$\pi_{S_j}(s_i) = \pi_{S_j} \times \frac{f_{S_j}(s_i \mid \boldsymbol{\theta})}{\sum_{k=1}^{3} f_{S_k}(s_i \mid \boldsymbol{\theta})}$
`pip_1`, `pip_2`, $\ldots$, `pip_9`	Posterior probability of belonging to class $j = 1, \ldots, 9$	$\pi_j(r_i, s_i) = \pi_j \times \frac{f_j(r_i, s_i \mid \boldsymbol{\theta})}{\sum_{k=2}^{9} f_k(r_i, s_i \mid \boldsymbol{\theta})}$
`bclass_r`, `bclass_s`	Bayesian classification of observation i to type R1, R2, or R3 and to type S1, S2, or S3, respectively	$bcX_i = j$ if $\pi_{X_j}(x_i) > \pi_{X_h}(x_i)$ $\forall h \neq j$ and $X \in \{R, S\}$ and $x \in \{r, s\}$
`bclass`	Bayesian classification of observation i to class $j = 1, \ldots, 9$	$bc_i = j$ if $\pi_j(r_i, s_i) > \pi_h(r_i, s_i)$ $\forall h \neq j$

The posterior or conditional probability of observation i belonging to a given class, say, class 2, is defined as the product of the unconditional probability of belonging to class 2 and the ratio of the likelihood of observation i belonging to class 2, divided by the sum of the likelihoods of observation i belonging to all classes (2 through 9). Given the posterior probabilities, the Bayesian classifier assigns each observation to the class for which the posterior probability is greatest. For all variants of our FMMs, the conditional probability of belonging to class 1 is equal to 1 if the observation belongs to the completely labeled group and 0 otherwise.

Finally, you can use `predict` to obtain seven different predictors of each individual's latent true earnings (ξ_i) using option `star`. The methods, proposed by MRW and extended by us to our general FMM, combine information from both administrative and survey data. The syntax of the option is as follows:

`predict` *prefix*`, star` [`replace surv_only`]

The new variables are named using *prefix* and consecutive integers from 1 to 7 and are created as data type `double`. To replace existing variable values, use option `replace`; `surv_only` requests the same predictors for the situation in which you have access to survey data only (as well as model estimates).

We describe the predictors ("hybrid" earnings variables) in table 6, with derivations of the formulas presented in the appendix. Predictors 1 to 6 use two within-class predictions for ξ. The first set, $\widehat{\xi}_i^j$, used for predictors 1, 3, and 5, minimizes the MSE, $E\{(\xi_i - \widehat{\xi}_i^j)^2 | \xi_i, i \in J\}$. The second set of predictors, $\widehat{\xi}_{Ui}^j$, used for cases 2, 4, and 6, minimizes the MSE conditional on $E(\xi_i - \widehat{\xi}_{Ui}^j | i \in J) = 0$. Predictors 1 and 2 provide weighted predictors using the unconditional within-class probabilities π_j. Predictors 3 and 4 provide weighted predictors using conditional or posterior within-class probabilities $\pi_j(r_i, s_i)$. Finally, predictors 5 and 6 use the two-step Bayesian classification. The seventh predictor ($\widehat{\xi_{7i}}$) is a system-wide predictor that minimizes MSE under the assumption of linearity and imposing the condition of unbiasedness.

Table 6. Seven predictors of latent true earnings

Variable name	Predictor description	Definition
prefix`1`	Weighted unconditional	$\widehat{\xi}_{1i} = \sum_{j=1}^{9} \pi_j \widehat{\xi}_i^j$
prefix`2`	Weighted unconditional and unbiased	$\widehat{\xi}_{2i} = \sum_{j=1}^{9} \pi_j \widehat{\xi}_{Ui}^j$
prefix`3`	Weighted conditional	$\widehat{\xi}_{3i} = \sum_{j=1}^{9} \pi_j(r_i, s_i) \widehat{\xi}_i^j$
prefix`4`	Weighted conditional and unbiased	$\widehat{\xi}_{4i} = \sum_{j=1}^{9} \pi_j(r_i, s_i) \widehat{\xi}_{Ui}^j$
prefix`5`	Two-step	$\widehat{\xi}_{5i} = \sum_{j=1}^{9} (bc_i = j) \widehat{\xi}_i^j$
prefix`6`	Two-step unbiased	$\widehat{\xi}_{6i} = \sum_{j=1}^{9} (bc_i = j) \widehat{\xi}_{Ui}^j$
prefix`7`	System-wide, linear	$\widehat{\xi}_{7i} = \widehat{\mu}_\xi + \mathbf{\Sigma}_{\xi y} \mathbf{\Sigma}_y^{-1} (\mathbf{y}_i - \widehat{\boldsymbol{\mu}}_{y\|x})$, $\mathbf{y}_i = (r_i, s_i)$

NOTE: $\widehat{\xi}_i^j$ is the within-class predictor that minimizes $E\{(\xi_i - \widehat{\xi}_i^j)^2 | \xi_i, i \in J\}$. $\widehat{\xi}_{Ui}^j$ is the within-class predictor that minimizes MSE under the condition $E(\xi_i - \widehat{\xi}_{Ui}^j | i \in J) = 0$. $\mathbf{\Sigma}_{\xi y}$ is the covariance matrix between ξ_i and (r_i, s_i). $\mathbf{\Sigma}_y^{-1}$ corresponds to the inverse of the variance–covariance matrix of (r_i, s_i). $\widehat{\boldsymbol{\mu}}_{y|x}$ is the system-wide expected value for (r_i, s_i). See MRW and the appendix for further details.

3.4 Data simulation: ky_sim

ky_sim is a utility command for simulating data based on the data-generating process characterized by the fitted FMM, as described in section 2 and table 2. The new dataset contains simulated values of s_i and r_i for each individual.

ky_sim simulates the joint distribution of administrative and survey log earnings in two ways. The first way allows you to simulate data by selecting the FMM that characterizes the data-generating function, setting the number of observations to be contained in the simulated dataset, and providing values for each of the parameters that characterize the given model variant. Model parameters are constant across observations—this corresponds to the specification of models without covariates. The syntax for this option is as follows:

ky_sim, model(#) nobs(#) [*options*]

model(#) specifies the model that characterizes the data-generating function, where # identifies one of the 8 models listed in table 2. model() is required.

nobs(#) sets the number of observations in the dataset to be created. nobs() is required.

seed(#) sets the random-number seed to be used for the simulation of the data.

If there is an unsaved dataset in memory, ky_sim will not generate the new simulated data unless option clear is specified.

You must specify values for the following parameters, with the specification depending on model selected:

Means: mean_e(#) mean_n(#) mean_t(#) mean_w(#) mean_v(#)

SDs: sig_e(#) sig_n(#) sig_t(#) sig_w(#) sig_v(#)

Correlations: rho_r(#) rho_s(#) rho_w(#)

Error probabilities: pi_s(#) pi_w(#) pi_r(#) pi_v(#)

If you specify a parameter value that is not required for the model selected, it is ignored. For example, a value for rho_w(#) is ignored if data are simulated using any model other than models 7 or 8.

When the program is used in this way, it also stores information in e(), so you can use the other postestimation commands described earlier.

The second way to use ky_sim is as a postestimation command. In this case, ky_sim generates simulated data using parameter estimates from a previously fitted model as well as the data currently in memory. Command syntax in this case is

ky_sim [, *options*]

If `ky_sim` is specified without any options directly after fitting a model with `ky_fit`, simulated data are created using the parameters from this previously fitted model.

Alternatively, you can use parameters from a previously fitted model that have been stored in memory using `estimates store` or saved to disk using `estimates save`, using the option `est_sto()` or `est_sav()`. If you retrieve the stored or saved estimates to use with `ky_sim` and a model with covariates had been fit, all the relevant covariates must be available in the dataset currently in memory.

The option `prefix(`*prefix*`)` allows specification of the prefix for the names of the newly created variables. By default, the program uses the variable name prefix "_". Option `replace` enables the program to overwrite variables if they already exist in the dataset, and option `seed(`*#*`)` allows you to set the seed for replication purposes.

Depending on the model chosen, `ky_sim` creates the variables shown in table 7.

Table 7. Variables created using `ky_sim`

Variable name	Description
prefix`e_var`	Latent true log(earnings)
prefix`n_var`	Factor η_i (survey data measurement error)
prefix`w_var`	Factor ω_i (survey data contamination)
prefix`v_var`	Factor υ_i (administrative data measurement error)
prefix`t_var`	Mismatched log earnings ζ_i
prefix`pi_ri`	1 if data are linked correctly
prefix`pi_vi`	1 if administrative data have no mean-reverting error
prefix`pi_si`	1 if survey data are reported correctly
prefix`pi_wi`	1 if survey data contain contamination
prefix`r_var`	Administrative log(earnings)
prefix`s_var`	Survey log(earnings)
prefix`l_var`	1 if r_i and s_i are error free

NOTES: *prefix* is empty if `ky_sim` is used as a postestimation command. By default, *prefix* = _ when using the second way to simulate data.

4 Illustrations: Estimation and postestimation

This section shows how to use the commands described in the previous section by revisiting the pioneering second-generation study by KY and MRW's companion article and showing how to reproduce their estimates. We do not have access to KY's confidential linked dataset, so we simulate their data using the parameter estimates they report and then analyze the data using the commands described earlier.

We start by setting the parameter estimates for KY's "full" (most general) model, reported in KY's table C2, based on a sample of size 400. We use globals; you could also use locals or scalars.

```
global mean_e  12.283
global mean_t   9.187
global mean_w (-0.304)
global mean_n (-0.048)
global sig_e    0.717
global sig_t    1.807
global sig_w    1.239
global sig_n    0.099
global pi_r     0.959
global pi_s     0.152
global pi_w     0.156
global rho_s  (-0.013)
```

KY's full model corresponds to model 4 of our FMM variants (see table 2). We use option `model(4)` and set the sample size with `nobs(400)`. Because `ky_sim` stores all the information in `e()`, we can also store that information in memory with `estimates store` and use it as a benchmark later.

```
ky_sim, seed(101) nobs(400) model(4)                      ///
       mean_e($mean_e) mean_t($mean_t) mean_w($mean_w)    ///
       mean_n($mean_n) sig_e($sig_e) sig_t($sig_t)        ///
       sig_w($sig_w) sig_n($sig_n)                        ///
       pi_r($pi_r) pi_s($pi_s) pi_w($pi_w) rho_s($rho_s) clear
estimates store model0
```

Using the simulated dataset, we can fit all the (simpler) models that are reported in KY's table C2 in addition to their full model (our model 4). KY's "basic" model corresponds to our model 1 with the additional restriction that $\mu_\eta = 0$. Their "no-mismatch" and "no-contamination" models correspond to our models 2 and 3.

```
constraint 1 [mu_n]_cons = 0
ky_fit r_var s_var l_var, model(1) constraint(1)
estimates store model1
ky_fit r_var s_var l_var, model(2)
estimates store model2
ky_fit r_var s_var l_var, model(3)
estimates store model3
ky_fit r_var s_var l_var, model(4)
estimates store model4
estimates table model0 model4 model3 model2 model1
```

Table 8 shows that parameter estimates derived from the simulated data are close to those reported by KY; so, too, are standard errors and log-likelihood values. The transformation of the mean-reversion correlation (`arho_s`) is large and statistically significant in the basic model but is much smaller for other models. The largest difference across models is in the estimate of `ln_sig_w`. We attribute this to the random nature of the simulated dataset.

Table 8. Estimates of KY models based on simulated data

	KY full	Simulated data							
	model	Full model		No contamination		No mismatch		Basic model	
mu_e	12.283	12.349	(0.034)	12.306	(0.038)	12.240	(0.048)	12.246	(0.037)
mu_n	−0.048	−0.061	(0.006)	−0.062	(0.006)	−0.059	(0.006)	0.000	(.)
mu_w	−0.304	−0.344	(0.148)			0.479	(0.284)		
mu_t	9.187	8.586	(0.678)	11.622	(0.256)				
ln_sig_e	−0.333	−0.406	(0.036)	−0.285	(0.036)	−0.047	(0.035)	−0.047	(0.035)
ln_sig_n	−2.313	−2.295	(0.048)	−2.270	(0.047)	−2.268	(0.046)	−0.449	(0.038)
ln_sig_w	0.592	−0.026	(0.112)			0.731	(0.100)		
ln_sig_t	0.214	0.501	(0.315)	0.622	(0.098)				
arho_s	−0.013	−0.022	(0.010)	−0.015	(0.010)	−0.026	(0.010)	−0.680	(0.054)
lpi_r	3.152	3.520	(0.335)	1.838	(0.159)				
lpi_s	−1.719	−1.844	(0.148)	−1.708	(0.150)	−1.879	(0.147)	−1.879	(0.147)
lpi_w	−1.688	−1.784	(0.189)			−1.683	(0.161)		
$\log\mathcal{L}$		−543.0		−595.5		−695.5		−1041.8	

NOTES: Standard errors in parentheses. Sample size = 400.

Table 8 reports estimated parameters (other than means) in a transformed metric. We use `margins` to obtain estimates of the parameters in their natural metric. To illustrate this, we focus on the estimates from the full model derived from simulated data.

```
. margins, predict(mean_e) predict(sig_e)
>                   predict(mean_t) predict(sig_t)
>                   predict(mean_w) predict(sig_w)
>                   predict(mean_n) predict(sig_n)
>                   predict(pi_r) predict(pi_s) predict(pi_w)
>                   predict(rho_s)
  (output omitted)
```

```
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
    _predict |
           1 |   12.34936   .0335341   368.26   0.000     12.28364    12.41509
           2 |   .6659948    .023718    28.08   0.000     .6195083    .7124813
           3 |   8.586231   .6782989    12.66   0.000     7.256789    9.915672
           4 |   1.650615   .5192749     3.18   0.001     .6328545    2.668375
           5 |  -.3435237   .1479331    -2.32   0.020    -.6334672   -.0535802
           6 |    .974735   .1089581     8.95   0.000      .761181    1.188289
           7 |  -.0608566   .0063531    -9.58   0.000    -.0733084   -.0484048
           8 |   .1007999   .0048806    20.65   0.000      .091234    .1103657
           9 |   .9712426   .0093542   103.83   0.000     .9529088    .9895765
          10 |   .1365808   .0174403     7.83   0.000     .1023985    .1707632
          11 |   .1437948   .0233102     6.17   0.000     .0981077    .1894819
          12 |  -.0220813   .0097204    -2.27   0.023     -.041133   -.0030297
------------------------------------------------------------------------------
```

If you specify a model in which parameters depend on explanatory variables, margins can also be used to obtain average predictive margins (APMs) of those parameters and to test contrasts. For example, suppose your `ky_fit` command specifies that the log of the survey measurement error SD depends on a binary indicator variable for the respondent's sex using the option `ln_sig_v(i.female)` and that women are coded with `female` = `1` and men with `female` = `0`. The following **`margins`** commands provide APM estimates of σ_ν, first for the sample as a whole and then separately by sex. The third command provides a test of the difference between the APMs for sex.

```
margins, predict(sig_v)
margins female, predict(sig_v)
margins female, predict(sig_v) pwcompare(effect)
```

The first command derives the value of σ_ν for every observation from the fitted model, with values of explanatory variables (`female` in this case) set at their sample values, and then reports the average over the sample of the derived σ_ν values, as well as the associated standard error. The second command provides separate estimates for men and women. It calculates the APM of σ_ν for `female` = `0` by first setting all sample values of `female` to 0 and then averaging over the whole sample. (If other explanatory variables had been included in the equation—not the case here—they would have been left at their sample values.) The command calculates the APM of σ_ν for `female` = `1` analogously.[7] The third command provides the test of the binary contrast in APMs. You can also use other pairwise and contrast options (**`help margins`**).

Let us now return to KY's full-model estimates and consider the reliability of the survey and administrative data. MRW showed how to investigate reliability by using a simulation-based method as well as by using analytical solutions (implied by the fitted model). MRW illustrated their methods using KY's estimates, showing that their survey data were more reliable than their administrative data, attributing this to the small but consequential prevalence of linkage mismatch.

The reliability statistics reported in MRW's table 6 can be obtained using our postestimation commands and the estimates reported by KY. For this illustration, we compare simulation-based and analytical reliability statistics using **`estat reliability`** and **`estat reliability, sim`**. We also use Jann's (2007) **`esttab`** utility, part of his **`estout`** package, for reporting results. We first show the code. Table 9 summarizes the results.

[7] **`margins, predict(sig_v) over(female)`** provides an alternative calculation. This derives estimates in the same way as the first command, except that the averaging is done separately for men and for women. In our experience, the estimates derived using this approach are very similar to those derived using the second command's approach.

```
ky_sim, seed(101) nobs(400) model(4)                   ///
       mean_e($mean_e) mean_t($mean_t) mean_w($mean_w) ///
       mean_n($mean_n) sig_e($sig_e) sig_t($sig_t)     ///
       sig_w($sig_w) sig_n($sig_n)                     ///
       pi_r($pi_r) pi_s($pi_s) pi_w($pi_w) rho_s($rho_s) clear

quietly: estat reliability
matrix rel_analytical = r(rel)
quietly: estat reliability, sim reps(100) seed(10)
matrix rel_simulation = r(rel)
esttab matrix(rel_analytical, fmt(4)) using table9,    ///
       mtitle("Analytical statistics") rtf replace b(4)
esttab matrix(rel_simulation, fmt(4)) using table9,    ///
       mtitle("Simulation statistics") rtf append b(4)
```

Table 9. Reliability statistics: Replication of MRW's table 6

Derivation method	Var	Cov	Rel1	Rel2
Analytical				
Administrative data	1.0038	0.4930	0.4912	0.4710
Survey data	0.7257	0.5084	0.7006	0.6929
Simulation				
Administrative data	0.9947	0.4866	0.4892	0.4662
Survey data	0.7169	0.5055	0.7051	0.6981

Table 9 shows that corresponding analytical and simulation-based statistics are similar. According to both derivation methods, we conclude that the survey data are more reliable than the administrative data, even though the mismatch probability is only 4.1%. The "analytical" statistics are the same as those reported in MRW's table 6.

MRW's main contribution was derivation of expressions for multiple predictors of latent true log earnings that combine information from survey and administrative measures with FMM estimates. To obtain observation-specific values for MRW's seven predictors, we use the `star` option to `predict`. To evaluate the statistical performance of the various predictors (assuming the data-generating process represented by model estimates is correct), we use postestimation command `estat xirel`. Internally, this calls on `ky_sim` to simulate data and `predict, star` to obtain the predictions.

```
. estat xirel, seed(10) reps(100)
Rel Statistics for 'e' predictions
            Rel1       Rel2        MSE    E(Bias)  Var(Bias)
r_var     0.5005     0.4786     0.5480    -0.1267     0.5321
s_var     0.7097     0.7021     0.2227    -0.0783     0.2165
  e_1     0.5605     0.5353     0.4344    -0.1189     0.4204
  e_2     0.5600     0.5375     0.4342    -0.1178     0.4204
  e_3     1.0020     0.9795     0.0105     0.0009     0.0105
  e_4     0.9879     0.9738     0.0137     0.0013     0.0136
  e_5     0.9892     0.9758     0.0125    -0.0002     0.0125
  e_6     0.9805     0.9714     0.0150    -0.0003     0.0150
  e_7     1.0068     0.7627     0.1216     0.0018     0.1217
```

The outputs for `e_1` to `e_7` correspond closely to what is shown in MRW's table 6. Observe the extremely good statistical performance of these predictors, especially `e_3` through `e_6` (see our table 6 for details of their definitions).

5 Conclusions

This article introduced a new set of commands for estimation and postestimation analysis of measurement error models for linked survey and administrative data. Our FMM specifications are those proposed by Jenkins and Rios-Avila (Forthcoming), which extend those proposed by KY. In particular, we allow for measurement error in the administrative data, as well as linkage mismatch and measurement error in the survey data. We also provide a suite of postestimation commands for simulation, assessing reliability, and deriving highly reliable hybrid earnings predictors of latent true earnings, building on the work of MRW. As Abowd and Stinson have pointed out, such predictors 'could be used by statistical agencies to produce a measure of "true earnings" [...], a valuable measure for researchers that would allow agencies to release information from administrative data while limiting confidentiality concerns' (2013, 1467).

Although our discussion has referred to labor earnings, our programs could also be used to examine measurement errors in other income variables. For example, Kapteyn and Ypma (2007) fitted their models to linked data on pensions and tax payments as well as employment earnings. Our approach could potentially be applied to other continuous variables such as height and body weight. (For example, a researcher may have, for each of many study participants, a self-reported measure of height or weight and a measure taken by a specialist interviewer: compare Cawley [2004].) A researcher has to decide before using our commands whether it is appropriate to assume that the unobserved true distribution of the concept of interest is normally distributed.

We hope that our commands will help researchers compare measurement error processes over time and across countries using a common approach that is based on a relatively general model. Linked datasets are becoming more commonly available. One limitation of our models is that they refer to cross-sectional data. We do not exploit the additional information provided by longitudinal linked datasets, as done in different ways by, for example, Abowd and Stinson (2013), Bollinger et al. (2018), and Hyslop and Townsend (2020). Adding longitudinal features to our FMM models is a task for future research.

6 Acknowledgments

Thanks to the FRS team at the U.K. Department of Work and Pensions for facilitating this project as part of their "secure data pilot" initiative. We also thank the anonymous referee and Nicholas J. Cox (handling editor) for helpful comments.

7 Programs and supplemental materials

Our software suite works with Stata 14 or later. To install a snapshot of the corresponding software files as they existed at the time of publication of this article, type

```
. net sj 23-1
. net install st0701      (to install program files, if available)
. net get st0701          (to install ancillary files, if available)
```

8 References

Abowd, J. M., and M. H. Stinson. 2013. Estimating measurement error in annual job earnings: A comparison of survey and administrative data. *Review of Economics and Statistics* 95: 1451–1467. https://doi.org/10.1162/REST_a_00352.

Aitkin, M., and D. B. Rubin. 1985. Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society, Series B* 47: 67–75. https://doi.org/10.1111/j.2517-6161.1985.tb01331.x.

Angel, S., F. Disslbacher, S. Humer, and M. Schnetzer. 2019. What did you really earn last year? Explaining measurement error in survey income data. *Journal of the Royal Statistical Society, Series A* 182: 1411–1437. https://doi.org/10.1111/rssa.12463.

Bingley, P., and A. Martinello. 2017. Measurement error in income and schooling and the bias of linear estimators. *Journal of Labor Economics* 35: 1117–1148. https://doi.org/10.1086/692539.

Bollinger, C. R. 1998. Measurement error in the current population survey: A nonparametric look. *Journal of Labor Economics* 16: 576–594. https://doi.org/10.1086/209899.

Bollinger, C. R., B. T. Hirsch, C. M. Hokayem, and J. P. Ziliak. 2018. The good, the bad and the ugly: Measurement error, non-response and administrative mismatch in the CPS. Working Paper, Gatton College of Business, University of Kentucky. http://christopherbollinger.com/wp-content/uploads/2019/09/GoodBadUglyFull.pdf.

Bound, J., and A. B. Krueger. 1991. The extent of measurement error in longitudinal earnings data: Do two wrongs make a right? *Journal of Labor Economics* 9: 1–24. https://doi.org/10.1086/298256.

Cameron, A. C., and P. K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.

Cawley, J. 2004. The impact of obesity on wages. *Journal of Human Resources* 39: 451–474. https://doi.org/10.2307/3559022.

Heckman, J., and B. Singer. 1984. A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* 52: 271–320. https://doi.org/10.2307/1911491.

Hyslop, D. R., and W. Townsend. 2020. Earnings dynamics and measurement error in matched survey and administrative data. *Journal of Business and Economic Statistics* 38: 457–469. https://doi.org/10.1080/07350015.2018.1514308.

Jann, B. 2007. Making regression tables simplified. *Stata Journal* 7: 227–244. https://doi.org/10.1177/1536867X0700700207.

Jenkins, S. P., and F. Rios-Avila. 2020. Modelling errors in survey and administrative data on employment earnings: Sensitivity to the fraction assumed to have error-free earnings. *Economics Letters* 192: 109253. https://doi.org/10.1016/j.econlet.2020.109253.

———. 2021. Measurement error in earnings data: Replication of Meijer, Rohwedder, and Wansbeek's mixture model approach to combining survey and register data. *Journal of Applied Econometrics* 36: 474–483. https://doi.org/10.1002/jae.2811.

———. Forthcoming. Reconciling reports: Modelling employment earnings and measurement errors using linked survey and administrative data. *Journal of the Royal Statistical Society*, Series A.

Kapteyn, A., and J. Y. Ypma. 2007. Measurement error and misclassification: A comparison of survey and administrative data. *Journal of Labor Economics* 25: 513–551. https://doi.org/10.1086/513298.

Kristensen, N., and N. Westergaard-Nielsen. 2007. A large-scale validation study of measurement errors in longitudinal survey data. *Journal of Economic and Social Measurement* 32: 65–92. https://doi.org/10.3233/JEM-2007-0283.

Meijer, E., S. Rohwedder, and T. Wansbeek. 2012. Measurement error in earnings data: Using a mixture model approach to combine survey and register data. *Journal of Business and Economic Statistics* 30: 191–201. https://doi.org/10.1198/jbes.2011.08166.

Redner, R. A., and H. F. Walker. 1984. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review* 26: 195–239. https://doi.org/10.1137/1026034.

Yakowitz, S. J., and J. D. Spragins. 1968. On the identifiability of finite mixtures. *Annals of Mathematical Statistics* 39: 209–214. https://doi.org/10.1214/aoms/1177698520.

About the authors

Stephen P. Jenkins is Professor of Economic and Social Policy at the London School of Economics and Political Science. He is also a Research Fellow at IZA, Bonn. Much of his substantive research is about inequality and poverty and related topics. He also has interests in applied microeconometrics, including survival analysis and statistical graphics, and the use of survey and administrative record data. He has contributed many commands to Statistical Software Components and written articles for the *Stata Journal*. He is currently an Editor of the *Stata Journal*.

Fernando Rios-Avila is a research scholar at the Levy Economics Institute of Bard College under the Distribution of Income and Wealth program. His research interests include applied econometrics, labor economics, and poverty and inequality.

A Appendix

This appendix contains three sections. Appendix A.1 discusses the relationship between conditional and unconditional correlations for a pair of random variables. Appendix A.2 provides expressions for expected values (means), variances, and covariances for the components in our general FMM. Appendix A.3 provides expressions for hybrid earnings predictors of latent true earnings for our general model, building on MRW's work.

A.1 Unconditional and conditional covariances between variables

Consider two random variables e_i and u_i defined as follows

$$e_i = \mu_{e|\mathbf{x}} + \varepsilon_{i,e}; \quad u_i = \mu_{u|\mathbf{x}} + \varepsilon_{i,u}$$
$$\begin{pmatrix} \varepsilon_{i,e} \\ \varepsilon_{i,u} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_e^2 & \rho\sigma_e\sigma_u \\ \rho\sigma_e\sigma_u & \sigma_u^2 \end{bmatrix} \right)$$

where $\mu_{k|\mathbf{x}} = E(k_i|\mathbf{x})$ for $k_i \in \{e_i, u_i\}$ and $\mathbf{x}$ is a vector of observed characteristics for individual $i = 1, \ldots, N$. Based on the law of total variance, and assuming $(\varepsilon_{i,e}, \varepsilon_{i,u})$ are independently distributed from $\mathbf{x}$, we have

$$\begin{aligned} \text{Var}(k_i) &= E\{\text{Var}(k_i|\mathbf{x})\} + \text{Var}\{E(k_i|\mathbf{x})\} \\ \text{Var}(k_i) &= \sigma_k^2 + \text{Var}(\mu_{k|\mathbf{x}}) \quad \text{for} \quad k_i \in \{e_i, u_i\} \end{aligned}$$

Similarly, using the law of total covariance, we have

$$\begin{aligned} \text{Cov}(e_i, u_i) &= E\{\text{Cov}(e_i, u_i|\mathbf{x})\} + \text{Cov}\{E(e_i|\mathbf{x}), E(u_i|\mathbf{x})\} \\ \text{Cov}(e_i, u_i) &= \rho\sigma_e\sigma_u + \text{Cov}(\mu_{e|\mathbf{x}}, \mu_{u|\mathbf{x}}) \end{aligned}$$

Thus, even if e_i and u_i are conditionally uncorrelated, their unconditional covariance may be nonzero.

A.2 Expected values, variances, and covariances for the general FMM

We provide expressions for the moments of the administrative data and the survey data, in turn.

A.2.1 Administrative data

The data structure for administrative data is

$$r_i = \left\{ \begin{array}{ll} r_{1,i} = \xi_i & \text{with probability } \pi_{r_1} = \pi_r\pi_\nu \\ r_{2,i} = \xi_i + \rho_r(\xi_i - \mu_{\xi|\mathbf{x}}) + \nu_i & \text{with probability } \pi_{r_2} = \pi_r(1-\pi_\nu) \\ r_{3,i} = \zeta_i & \text{with probability } \pi_{r_3} = 1 - \pi_r \end{array} \right\}$$

The data-generating process for the latent variables is

$$\begin{pmatrix} \xi_i \\ \nu_i \\ \zeta_i \end{pmatrix} = N\left(\begin{bmatrix} \mu_{\xi|\mathbf{x}} \\ \mu_{\nu|\mathbf{x}} \\ \mu_{\zeta|\mathbf{x}} \end{bmatrix}, \begin{bmatrix} \sigma_\xi^2 & 0 & 0 \\ 0 & \sigma_\nu^2 & 0 \\ 0 & 0 & \sigma_\zeta^2 \end{bmatrix} \right)$$

where $\mu_{\gamma|\mathbf{x}}$ can be expressed as a linear function of $\mathbf{x}$ for each $\gamma \in \{\xi, \nu, \zeta\}$.

Unconditional moments by data type (class)

Class 1: $r_{1,i} = \xi_i$

Expected value:

$$E(r_{1,i}) = \mu_\xi$$

Variance:

$$\text{Var}(r_{1,i}) = \text{Var}(\xi_i) = \sigma_\xi^2 + \text{Var}(\mu_\xi|\mathbf{x})$$

Covariance with ξ_i:

$$\text{Cov}(\xi_i, r_{1,i}) = \text{Var}(\xi_i) = \sigma_\xi^2 + \text{Var}(\mu_{\xi|\mathbf{x}})$$

Class 2: $r_{2,i} = \xi_i + \rho_r(\xi_i - \mu_{\xi|\mathbf{x}}) + \nu_i$

Expected value:

$$\begin{aligned} E(r_{2,i}) &= E(\xi_i + \rho_r(\xi_i - \mu_{\xi|\mathbf{x}}) + \nu_i) \\ &= \mu_\xi + \mu_\nu \end{aligned}$$

Variance:

$$\begin{aligned} \text{Var}(r_{2,i}) &= \text{Var}\{\xi_i + \rho_r(\xi_i - \mu_{\xi|\mathbf{x}}) + \nu_i\} \\ &= \text{Var}(\mu_{\xi|\mathbf{x}} + (1+\rho_r)(\xi_i - \mu_{\xi|\mathbf{x}}) + \nu_i) \\ &= \sigma^2_{\mu_{\xi|\mathbf{x}}} + (1+\rho_r)^2\sigma_\xi^2 + \text{Var}(\nu_i) + 2\text{Cov}(\mu_{\xi|\mathbf{x}}, \mu_{\nu|\mathbf{x}}) \end{aligned}$$

Covariance with ξ_i:

$$\begin{aligned} \text{Cov}(\xi_i, r_{2,i}) &= \text{Cov}\{\xi_i, \xi_i + \rho_r(\xi_i - \mu_{\xi|\mathbf{x}}) + \nu_i\} \\ &= \text{Var}(\xi_i) + \rho_r\sigma_\xi^2 + \text{Cov}(\mu_{\xi|\mathbf{x}}, \mu_{\nu|\mathbf{x}}) \\ &= \text{Var}(\mu_{\xi|\mathbf{x}}) + (1+\rho_r)\sigma_\xi^2 + \text{Cov}(\mu_{\xi|\mathbf{x}}, \mu_{\nu|\mathbf{x}}) \end{aligned}$$

Class 3: $r_{3,i} = \zeta_i$

Expected value:

$$E(r_{3,i}) = E(\zeta_i) = \mu_\zeta$$

Variance:

$$\text{Var}(r_{3,i}) = \text{Var}(\zeta_i) = \text{Var}(\mu_{\zeta|\mathbf{x}}) + \sigma_\zeta^2$$

Covariance with ξ_i:

$$\text{Cov}(\xi_i, r_{3,i}) = \text{Cov}(\xi_i, \zeta_i) = \text{Cov}(\mu_{\xi|\mathbf{x}}, \mu_{\zeta|\mathbf{x}})$$

Moments for administrative data, overall

Expected value:

$$\begin{aligned} E(r_i) &= \pi_{r_1} E(r_{1,i}) + \pi_{r_2} E(r_{2,i}) + \pi_{r_3} E(r_{3,i}) \\ &= \pi_{r_1}\mu_\xi + \pi_{r_2}(\mu_\xi + \mu_\nu) + \pi_{r_3}\mu_\zeta \\ &= (\pi_{r_1} + \pi_{r_2})\mu_\xi + \pi_{r_2}\mu_\nu + \pi_{r_3}\mu_\zeta \end{aligned}$$

Variance:

$$\mathrm{Var}(r_i) = \sum_{j=1}^{3} \pi_{r_j} \mathrm{Var}(r_{j,i}) + \mathrm{Var}\{E(r_{j,i})\}$$

where

$$\mathrm{Var}\{E(r_{j,i})\} = \sum_{j=1}^{3} \pi_{r_j} \{E(r_{j,i}) - E(r_i)\}^2$$

Covariance with ξ_i:

$$\mathrm{Cov}(\xi_i, r_i) = \sum_{j}^{3} \pi_{r_j} \mathrm{Cov}(\xi_i, r_{j,i})$$

A.2.2 Survey data

The data structure for survey data is

$$s_i = \left\{ \begin{array}{ll} s_{1,i} = \xi_i & \text{with probability } \pi_{s1} = \pi_s \\ s_{2,i} = \xi_i + \rho_s(\xi_i - \mu_{\xi|\mathbf{x}}) + \eta_i & \text{with probability } \pi_{s2} = (1-\pi_s)(1-\pi_\omega) \\ s_{3,i} = \xi_i + \rho_s(\xi_i - \mu_{\xi|\mathbf{x}}) + \eta_i + \omega_i & \text{with probability } \pi_{s3} = (1-\pi_s)\pi_\omega \end{array} \right\}$$

The data-generating process for the latent variables is

$$\begin{pmatrix} \xi_i \\ \eta_i \\ \omega_i \end{pmatrix} = N\left(\begin{bmatrix} \mu_{\xi|\mathbf{x}} \\ \mu_{\eta|\mathbf{x}} \\ \mu_{\omega|\mathbf{x}} \end{bmatrix}, \begin{bmatrix} \sigma_\xi^2 & 0 & \rho_\omega\sigma_\xi\sigma_\omega \\ 0 & \sigma_\eta^2 & 0 \\ \rho_\omega\sigma_\xi\sigma_\omega & 0 & \sigma_\omega^2 \end{bmatrix} \right)$$

where $\mu_{\gamma|\mathbf{x}}$ can be expressed as a linear function of $\mathbf{x}$ for each $\gamma = \{\xi, \eta, \omega\}$.

Unconditional moments by data class

Class 1: $s_{1,i} = \xi_i$

Expected value:

$$E(s_{1,i}) = \mu_\xi$$

Variance:

$$\mathrm{Var}(s_{1,i}) = \mathrm{Var}(\xi_i) = \sigma_\xi^2 + \mathrm{Var}(\mu_{\xi|\mathbf{x}})$$

Covariance with ξ_i:

$$\mathrm{Cov}(\xi_i, s_{1,i}) = \mathrm{Var}(\xi_i) = \sigma_\xi^2 + \mathrm{Var}(\mu_{\xi|\mathbf{x}})$$

Class 2: $s_{2,i} = \xi_i + \rho_s(\xi_i - \mu_{\xi|\mathbf{x}}) + \eta_i$

Expected value:

$$\begin{aligned} E(s_{2,i}) &= E\{\xi_i + \rho_s(\xi_i - \mu_{\xi|\mathbf{x}}) + \eta_i\} \\ &= \mu_\xi + \mu_\eta \end{aligned}$$

Variance:

$$\begin{aligned} \mathrm{Var}(s_{2,i}) &= \mathrm{Var}\{\xi_i + \rho_s(\xi_i - \mu_{\xi|\mathbf{x}}) + \eta_i\} \\ &= \mathrm{Var}\{\mu_{\xi|\mathbf{x}} + (1+\rho_s)(\xi_i - \mu_{\xi|\mathbf{x}}) + \eta_i\} \\ &= \sigma^2_{\mu_{\xi|\mathbf{x}}} + (1+\rho_s)^2\sigma^2_\xi + \mathrm{Var}(\eta_i) + 2\mathrm{Cov}(\mu_{\xi|\mathbf{x}}, \mu_{\eta|\mathbf{x}}) \end{aligned}$$

Covariance with ξ_i:

$$\begin{aligned} \mathrm{Cov}(\xi_i, s_{2,i}) &= \mathrm{Cov}\{\xi_i, \xi_i + \rho_s(\xi_i - \mu_{\xi|\mathbf{x}}) + \eta_i\} \\ &= \mathrm{Var}(\xi_i) + \rho_s\sigma^2_\xi + \mathrm{Cov}(\mu_{\xi|\mathbf{x}}, \mu_{\eta|\mathbf{x}}) \\ &= \mathrm{Var}(\mu_{\xi|\mathbf{x}}) + (1+\rho_s)\sigma^2_\xi + \mathrm{Cov}(\mu_{\xi|\mathbf{x}}, \mu_{\eta|\mathbf{x}}) \end{aligned}$$

Class 3: $s_{3,i} = \xi_i + \rho_s(\xi_i - \mu_{\xi|\mathbf{x}}) + \eta_i + \omega_i$

Expected value:

$$\begin{aligned} E(s_{3,i}) &= E\{\xi_i + \rho_s(\xi_i - \mu_{\xi|\mathbf{x}}) + \eta_i + \omega_i\} \\ &= \mu_\xi + \mu_\eta + \mu_\omega \end{aligned}$$

Variance:

$$\begin{aligned} \mathrm{Var}(s_{3,i}) &= \mathrm{Var}\{\xi_i + \rho_s(\xi_i - \mu_{\xi|\mathbf{x}}) + \eta_i + \omega_i\} \\ &= \mathrm{Var}\{\mu_{\xi|\mathbf{x}} + (1+\rho_s)(\xi_i - \mu_{\xi|\mathbf{x}}) + \eta_i + \omega_i\} \\ &= \sigma^2_{\mu_{\xi|\mathbf{x}}} + (1+\rho_s)^2\sigma^2_\xi + \mathrm{Var}(\eta_i) + \mathrm{Var}(\omega_i) + 2\mathrm{Cov}(\mu_{\xi|\mathbf{x}}, \mu_{\eta|\mathbf{x}}) \\ &\quad + 2\mathrm{Cov}(\mu_{\xi|\mathbf{x}}, \mu_{\omega|\mathbf{x}}) + 2(1+\rho_s)\rho_\omega\sigma_\xi\sigma_\omega + 2\mathrm{Cov}(\mu_{\omega|\mathbf{x}}, \mu_{\eta|\mathbf{x}}) \end{aligned}$$

Covariance with ξ_i:

$$\begin{aligned} \mathrm{Cov}(\xi_i, s_{3,i}) &= \mathrm{Cov}\{\xi_i, \xi_i + \rho_s(\xi_i - \mu_{\xi|\mathbf{x}}) + \eta_i + \omega_i\} \\ &= \mathrm{Var}(\xi_i) + \rho_s\sigma^2_\xi + \mathrm{Cov}(\mu_{\xi|\mathbf{x}}, \mu_{\eta|\mathbf{x}}) + \mathrm{Cov}(\mu_{\xi|\mathbf{x}}, \mu_{\omega|\mathbf{x}}) + \rho_\omega\sigma_\xi\sigma_\omega \\ &= \mathrm{Var}(\mu_{\xi|\mathbf{x}}) + (1+\rho_s)\sigma^2_\xi + \mathrm{Cov}(\mu_{\xi|\mathbf{x}}, \mu_{\eta|\mathbf{x}}) + \mathrm{Cov}(\mu_{\xi|\mathbf{x}}, \mu_{\omega|\mathbf{x}}) \\ &\quad + \rho_\omega\sigma_\xi\sigma_\omega \end{aligned}$$

Moments for survey data, overall

Expected value:

$$\begin{aligned} E(s_i) &= \pi_{s_1}E(s_{1,i}) + \pi_{s_2}E(s_{2,i}) + \pi_{s_3}E(s_{3,i}) \\ &= \pi_{s_1}\mu_\xi + \pi_{s_2}(\mu_\xi + \mu_\eta) + \pi_{s_3}(\mu_\xi + \mu_\eta + \mu_\omega) \\ &= \mu_\xi + (\pi_{s_2} + \pi_{s_3})\mu_\eta + \pi_{s_3}\mu_\omega \end{aligned}$$

Variance:

$$\text{Var}(s_i) = \sum_{j=1}^{3} \pi_{s_j}\text{Var}(s_{j,i}) + \text{Var}\{E(s_{j,i})\}$$

where

$$\text{Var}\{E(s_{j,i})\} = \sum_{j=1}^{3} \pi_{s_j}\{E(s_{j,i}) - E(s_i)\}^2$$

Covariance with ξ_i

$$\text{Cov}(\xi_i, s_i) = \sum_{j}^{3} \pi_{s_j}\text{Cov}(\xi_i, s_{j,i})$$

A.2.3 Conditional moments by data class

Table A1. Mean and variance of r_i and s_i, conditional on $\mathbf{x}$, by class

Data type	$E(\cdot\|\mathbf{x})$ or $\mu_{\cdot\|\mathbf{x}}$	$\text{Var}(\cdot\|\mathbf{x})$	$\text{Cov}(\xi_{i,\cdot}\|\mathbf{x})$
$r_{1,i}$	$\mu_{\xi\|\mathbf{x}}$	σ^2_ξ	σ^2_ξ
$r_{2,i}$	$\mu_{\xi\|\mathbf{x}} + \mu_{\upsilon\|\mathbf{x}}$	$(1+\rho_r)^2\sigma^2_\xi + \sigma^2_\upsilon$	$(1+\rho_r)\sigma^2_\xi$
$r_{3,i}$	$\mu_{\xi\|\mathbf{x}}$	σ^2_ζ	0
$s_{1,i}$	$\mu_{\xi\|\mathbf{x}}$	σ^2_ξ	σ^2_ξ
$s_{2,i}$	$\mu_{\xi\|\mathbf{x}} + \mu_{\eta\|\mathbf{x}}$	$(1+\rho_s)^2\sigma^2_\xi + \sigma^2_\eta$	$(1+\rho_s)\sigma^2_\xi$
$s_{3,i}$	$\mu_{\xi\|\mathbf{x}} + \mu_{\eta\|\mathbf{x}} + \mu_{\omega\|\mathbf{x}}$	$(1+\rho_s)^2\sigma^2_\xi + \sigma^2_\eta + \sigma^2_\omega$ $+2(1+\rho_s)\rho_\omega\sigma_\xi\sigma_\omega$	$(1+\rho_s)\sigma^2_\xi + \rho_\omega\sigma_\xi\sigma_\omega$

Table A2. Covariance between r_i and s_i, conditional on $\mathbf{x}$, by class

$\mathrm{Cov}(\cdot\vert\mathbf{x})$	$s_{1,i}$	$s_{2,i}$	$s_{3,i}$
$r_{1,i}$	σ_ξ^2	$(1+\rho_s)\sigma_\xi^2$	$(1+\rho_s)\sigma_\xi^2+\rho_\omega\sigma_\xi\sigma_\omega$
$r_{2,i}$	$(1+\rho_r)\sigma_\xi^2$	$(1+\rho_r)(1+\rho_s)\sigma_\xi^2$	$(1+\rho_r)(1+\rho_s)\sigma_\varepsilon^2+(1+\rho_r)\rho_\omega\sigma_\xi\sigma_\omega$
$r_{3,i}$	0	0	0

Overall covariance conditional on x

$$
\begin{aligned}
\mathrm{Cov}(r_i,s_i|\mathbf{x}) &= \sum_{h=1}^{3}\sum_{k=1}^{3}\pi_{r_h}\pi_{s_k}\mathrm{Cov}(r_{h,i},s_{k,i}|\mathbf{x})\\
&= \pi_{r_1}[\pi_{s_1}\sigma_\xi^2+\pi_{s_2}(1+\rho_s)\sigma_\xi^2+\pi_{s_3}\{(1+\rho_s)\sigma_\xi^2+\rho_\omega\sigma_\xi\sigma_\omega\}]\\
&\quad\; +\pi_{r_2}[\pi_{s_1}(1+\rho_r)\sigma_\xi^2+\pi_{s_2}(1+\rho_r)(1+\rho_s)\sigma_\xi^2\\
&\quad\; +\pi_{s_3}\{(1+\rho_r)(1+\rho_s)\sigma_\varepsilon^2+(1+\rho_r)\rho_\omega\sigma_\xi\sigma_\omega\}]+\pi_{r_3}(0)\\
&= \pi_{r_1}[\{1+(\pi_{s_2}+\pi_{s_3})\rho_s\}\sigma_\xi^2+\pi_{s_3}\rho_\omega\sigma_\xi\sigma_\omega]\\
&\quad\; +\pi_{r_2}[\{1+(\pi_{s_2}+\pi_{s_3})\rho_s\}(1+\rho_r)\sigma_\xi^2+\pi_{s_3}(1+\rho_r)\rho_\omega\sigma_\xi\sigma_\omega]
\end{aligned}
$$

Overall unconditional covariance

$$
\mathrm{Cov}(r_i,s_i)=\mathrm{Cov}(r_i,s_i|\mathbf{x})+\mathrm{Cov}(\mu_{r|\mathbf{x}},\mu_{s|\mathbf{x}})
$$

where

$$
\begin{aligned}
\mu_{r|\mathbf{x}} &= E(r_i|\mathbf{x})=(\pi_{r_1}+\pi_{r_2})\mu_{\xi|\mathbf{x}}+\pi_{r_2}\mu_{\nu|\mathbf{x}}+\pi_{r_3}\mu_{\zeta|\mathbf{x}}\\
\mu_{s|\mathbf{x}} &= \mu_{\xi|\mathbf{x}}+(\pi_{s_2}+\pi_{s_3})\mu_{\eta|\mathbf{x}}+\pi_{s_3}\mu_{\omega|\mathbf{x}}
\end{aligned}
$$

A.3 Predictors of latent true earnings

Following MRW, we differentiate between within-class predictors and a system-wide predictor. For the second case, we consider the simplest scenario of prediction under linearity.

A.3.1 System-wide predictor under linearity

Consider two measures r_i and s_i, which are manifest measures of latent true earnings, ξ_i, but are measured with error. Without loss of generality, assume that $\mu_k = \mu_{k|\mathbf{x}} = 0$. A predictor for the latent variable, $\widehat{\xi_i}$, can be derived as a linear combination as follows:

$$\widehat{\xi_i} = \theta_1 r_i + \theta_2 s_i \tag{A1}$$

The system-wide predictor will be characterized given a set of weights θ_1 and θ_2 that minimize the MSE between the predictor and the true latent variable ξ_i.

$$\min_{\theta_1,\theta_2} \text{MSE} = E\left\{\left(\xi_i - \widehat{\xi_i}\right)^2\right\} = E\left\{\xi_i - (\theta_1 r_i + \theta_2 s_i)\right\}^2) \tag{A2}$$

The first-order conditions are

$$\begin{aligned}
\frac{\partial \text{MSE}}{\partial \theta_1} &= E\left\{(\xi_i - \theta_1 r_i - \theta_2 s_i)\, r_i\right\} \\
&= E\left(\xi_i r_i - \theta_1 r_i^2 - \theta_2 r_i s_i\right) \\
&= \text{Cov}(\xi_i, r_i) - \theta_1 \text{Var}(r_i) - \theta_2 \text{Cov}(r_i, s_i) = 0 && \text{(A3)} \\
\frac{\partial \text{MSE}}{\partial \theta_2} &= E\left\{(\xi_i - \theta_1 r_i - \theta_2 s_i)\, s_i\right\} \\
&= E(\xi_i s_i - \theta_1 r_i s_i - \theta_2 s_i^2) \\
&= \text{Cov}(\xi_i, s_i) - \theta_1 \text{Cov}(r_i, s_i) - \theta_2 \text{Var}(s_i) = 0 && \text{(A4)}
\end{aligned}$$

Solving the system of equations given by (A3) and (A4), we have

$$\begin{bmatrix} \text{Cov}(\xi_i, r_i) \\ \text{Cov}(\xi_i, s_i) \end{bmatrix} = \begin{bmatrix} \text{Var}(r_i) & \text{Cov}(r_i, s_i) \\ \text{Cov}(r_i, s_i) & \text{Var}(s_i) \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$$

$$\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} \text{Var}(r_i) & \text{Cov}(r_i, s_i) \\ \text{Cov}(r_i, s_i) & \text{Var}(s_i) \end{bmatrix}^{-1} \begin{bmatrix} \text{Cov}(\xi_i, r_i) \\ \text{Cov}(\xi_i, s_i) \end{bmatrix}$$

Given solutions for θ_1 and θ_2, we can substitute them into (A1), which provides the system-wide predictor for $\widehat{\xi_i}$.

$$\begin{aligned}
\widehat{\xi_i} &= \begin{bmatrix} \theta_1 & \theta_2 \end{bmatrix} \begin{bmatrix} r_i \\ s_i \end{bmatrix} \\
\widehat{\xi_i} &= \begin{bmatrix} (\text{Cov}(\xi_i, r_i) & \text{Cov}(\xi_i, s_i) \end{bmatrix} \begin{bmatrix} \text{Var}(r_i) & \text{Cov}(r_i, s_i) \\ \text{Cov}(r_i, s_i) & \text{Var}(s_i) \end{bmatrix}^{-1} \begin{bmatrix} r_i \\ s_i \end{bmatrix}
\end{aligned}$$

This is the same predictor as given by MRW's (11), page 96. We label this predictor 7 in the main text.

A.3.2 Within-class predictors

For the estimates that rely on within-class predictors (predictors 1–6 in the main text), MRW discuss two estimators: linear estimators that minimize the within-class MSE $\widehat{\xi}_i^j$ and the estimator that minimizes the MSE conditional on the estimator being unbiased $\widehat{\xi}_{Ui}^j$.

The general form for the within-class predictor $\widehat{\xi}_i^j$ follows the same structure as (A2), but for each subclass 2–9, and so is not discussed further here. However, the unbiased estimator depends on the specific class.

The solutions for classes 1, 2, 3, 4, and 7 are straightforward to derive because they assume that either r_i or s_i is an error-free measure of ξ_i. Thus, we concentrate on the predictors corresponding to classes 5, 6, 8, and 9.

Classes 8 and 9

These two classes assume that only s_i contains information that can be used to construct the predictor for ξ. We refer here to the predictor for class 9 as the more general case. Without loss of generality, we assume that the unconditional and conditional (on $\mathbf{x}$) means of all variables in the model are equal to zero.

Under these assumptions, the predictor $\widehat{\xi}$ for class 9 is a linear transformation of s_i given by

$$\widehat{\xi}_{Ui}^9 = \theta s_{3,i}$$

where θ is selected so it minimizes the within-class MSE, conditional on the predictor being unbiased estimate for ξ. We start with the second condition:

$$\begin{aligned}
E(\xi_i - \theta s_{3,i}|\xi_i) &= 0\\
&= E\left\{\xi_i - \theta(\xi_i + \rho_s\xi_i + \eta_i + \omega_i)|\xi_i\right\}\\
&= E(\xi_i|\xi_i) - \theta(1+p_s)E(\xi_i|\xi_i) - \theta E(\eta_i|\xi_i)\\
&= \xi_i - \theta(1+p_s)\xi_i - 0 - \theta\rho_\omega\frac{\sigma_\omega}{\sigma_\xi}\xi_i\\
&\Rightarrow 1 - \theta(1+p_s) - \theta\rho_\omega\frac{\sigma_\omega}{\sigma_\xi} = 0\\
&\Rightarrow \theta = \frac{1}{1+p_s+\rho_\omega\frac{\sigma_\omega}{\sigma_\xi}}
\end{aligned}$$

Thus, the ξ unbiased predictor for class 9 is

$$\widehat{\xi}_{Ui}^9 = \theta s_{3,i} = \frac{s_{3,i}}{1+p_s+\rho_\omega\frac{\sigma_\omega}{\sigma_\xi}} \tag{A5}$$

and the unbiased predictor for class 8 is

$$\widehat{\xi}_{Ui}^8 = \theta s_{2,i} = \frac{s_{2,i}}{1+p_s} \tag{A6}$$

Equations (A5) and (A6) imply that the unbiased predictors for classes 8 and 9 are defined uniquely by imposing the unbiasedness assumption.

Classes 5 and 6

For classes 5 and 6, two measures can be used as proxies for ξ, each with its own sources of errors. We refer here to the solution for class 6 as the more general case.

Consider first the unbiased predictors that could be derived using data from r_{2i} or s_{3i}, which follow the same structure as (A3) and (A4):

$$\begin{aligned} \widehat{\xi}_{Ui}^{6r2} &= \frac{r_{2,i}}{1+p_r} = \theta_{r2} r_{2,i} \\ \widehat{\xi}_{Ui}^{6s3} &= \frac{s_{3,i}}{1+p_s+\rho_\omega \frac{\sigma_\omega}{\sigma_\xi}} = \theta_{s3} s_{3,i} \end{aligned} \tag{A7}$$

An unbiased ξ predictor for class 6 that combines the information from both sources can be obtained using a weighted average between both predictors:

$$\begin{aligned} \widehat{\xi}_{Ui}^{6} &= \delta \widehat{\xi}_{Ui}^{6r} + (1-\delta)\widehat{\xi}_{Ui}^{6s} \\ \widehat{\xi}_{Ui}^{6} &= \delta \theta_{r2} r_{2,i} + (1-\delta)\theta_{s3} s_{3,i} \end{aligned}$$

To determine the optimal weight, we need to find the value δ that minimizes the MSE, which is given by

$$\min_\delta E\left[\left\{\xi_i - \delta\theta_{r2} r_{2,i} - (1-\delta)\theta_{s3} s_{3,i}\right\}^2\right]$$

The first-order condition is

$$\begin{aligned} \frac{\partial \text{MSE}}{\partial \delta} &= E\left[\left\{\xi_i - \delta\theta_{r2} r_{2,i} - (1-\delta)\theta_{s3} s_{3,i}\right\}(\theta_{r2} r_{2,i} - \theta_{s3} s_{3,i})\right] = 0 \\ \theta_{r2}\text{Cov}(\xi_i, r_{2,i}) - \theta_{s3}\text{Cov}(\xi_i, s_{3,i}) - \delta\theta_{r2}^2\text{Var}(r_{2,i}) & \\ + \delta\theta_{r2}\theta_{s3}\text{Cov}(r_{2,i}, s_{3,i}) - (1-\delta)\theta_{r2}\theta_{s3}\text{Cov}(r_{2,i}, s_{3,i}) & \\ + (1-\delta)\theta_{s3}^2\text{Var}(s_{3,i}) = 0 & \end{aligned}$$

Finally, solving for δ, we have

$$\delta = \frac{\theta_{r2}\text{Cov}(\xi_i, r_{2,i}) - \theta_{s3}\text{Cov}(\xi_i, s_{3,i}) - \theta_{r2}\theta_{s3}\text{Cov}(r_{2,i}, s_{3,i}) + \theta_{s3}^2\text{Var}(s_{3,i})}{\theta_{r2}^2\text{Var}(r_{2,i}) - 2\theta_{r2}\theta_{s3}\text{Cov}(r_{2,i}, s_{3,i}) + \theta_{s3}^2\text{Var}(s_{3,i})} \tag{A8}$$

Substituting (A8) into (A7) provides the unbiased predictor for class 6.

To summarize, table A3 presents the expressions for the within-class predictions for all 9 classes, assuming that our general model (model 8) describes the data-generating process. The expressions for the other models are simplified versions of the expressions in the table.

Table A3. Expressions for the within-class predictors as functions of the parameters (general FMM)

Class (j)	r	s	$\widehat{\xi}^j$	$\widehat{\xi}^j_U$
1	$r_{1,i}$	$s_{1,i}$	$\frac{1}{2}(r+s)$	$\frac{1}{2}(r+s)$
2	$r_{1,i}$	$s_{2,i}$	r	r
3	$r_{1,i}$	$s_{3,i}$	r	r
4	$r_{2,i}$	$s_{1,i}$	s	s
5	$r_{2,i}$	$s_{2,i}$	$\mu_{\xi\|\mathbf{x}} + \mathbf{\Sigma}'_{\xi,5}\mathbf{\Sigma}_5^{-1}\begin{bmatrix} r_i - \mu_{r_2\|\mathbf{x}} \\ s_i - \mu_{s_2\|\mathbf{x}} \end{bmatrix}$	$\mu_{\xi\|\mathbf{x}} + \begin{bmatrix} \delta_{r_2,s_2}\theta_{r2} \\ (1-\delta_{r_2,s_2})\theta_{s2}) \end{bmatrix}' \begin{bmatrix} r_i - \mu_{r_2\|\mathbf{x}} \\ s_i - \mu_{s_2\|\mathbf{x}} \end{bmatrix}$
6	$r_{2,i}$	$s_{3,i}$	$\mu_{\xi\|\mathbf{x}} + \mathbf{\Sigma}'_{\xi,6}\mathbf{\Sigma}_6^{-1}\begin{bmatrix} r_i - \mu_{r_2\|\mathbf{x}} \\ s_i - \mu_{s_3\|\mathbf{x}} \end{bmatrix}$	$\mu_{\xi\|\mathbf{x}} + \begin{bmatrix} \delta_{r_2,s_3}\theta_{r2} \\ (1-\delta_{r_2,s_3})\theta_{s3} \end{bmatrix}' \begin{bmatrix} r_i - \mu_{r_2\|\mathbf{x}} \\ s_i - \mu_{s_3\|\mathbf{x}} \end{bmatrix}$
7	$r_{3,i}$	$s_{1,i}$	s	s
8	$r_{3,i}$	$s_{2,i}$	$\mu_{\xi\|\mathbf{x}} + \frac{\mathrm{Cov}(\xi_i, s_{2,i}\|\mathbf{x})}{\mathrm{Var}(s_{2,i}\|\mathbf{x})}(s_i - \mu_{s2\|\mathbf{x}})$	$\mu_{\xi\|\mathbf{x}} + \frac{1}{\theta_{s2}}(s_i - \mu_{s2\|\mathbf{x}})$
9	$r_{3,i}$	$s_{3,i}$	$\mu_{\xi\|\mathbf{x}} + \frac{\mathrm{Cov}(\xi_i, s_{3,i}\|\mathbf{x})}{\mathrm{Var}(s_{3,i}\|\mathbf{x})}(s_i - \mu_{s3\|\mathbf{x}})$	$\mu_{\xi\|\mathbf{x}} + \frac{1}{\theta_{s3}}(s_i - \mu_{s3\|\mathbf{x}})$

NOTES: $\mathbf{\Sigma}_{\xi,j}$ represents the covariances between ξ_i and (r_i, s_i), conditional on characteristics $\mathbf{x}$ and class j. $\mathbf{\Sigma}_j$ represents the variance–covariance matrix between r_i and s_i, conditional on characteristics $\mathbf{x}$ and class j. Also, $\delta_{r_j,s_k} = \{\theta_{r_j}\mathrm{Cov}(\xi_i, r_{j,i}) - \theta_{s_k}\mathrm{Cov}(\xi_i, s_{k,i}) - \theta_{r_j}\theta_{s_k}\mathrm{Cov}(r_{j,i}, s_{k,i}) + \theta^2_{s_k}\mathrm{Var}(s_{k,i})\}/\{\theta^2_{r_j}\mathrm{Var}(r_{j,i}) - 2\theta_{r_j}\theta_{s_k}\mathrm{Cov}(r_{j,i}, s_{k,i}) + \theta^2_{s_k}\mathrm{Var}(s_{k,i})\}$; $\theta_{r_2} = 1/(1+\rho_r)$; $\theta_{s_2} = 1/(1+\rho_s)$; and $\theta_{s_3} = 1/\{1+\rho_s+\rho_\omega(\sigma_\omega/\sigma_\xi)\}$.