Published in final edited form as:

J Bus Econ Stat. 2012 ; 30(2): 191–201. doi:10.1198/jbes.2011.08166.

Measurement error in earnings data: Using a mixture model approach to combine survey and register data

Erik Meijer,

RAND Corporation, Santa Monica, CA 90407-2138

Susann Rohwedder, and

RAND Corporation, Santa Monica, CA 90407-2138

Tom Wansbeek

University of Groningen, 9700 AV Groningen, The Netherlands

Abstract

Survey data on earnings tend to contain measurement error. Administrative data are superior in principle, but they are worthless in case of a mismatch. We develop methods for prediction in mixture factor analysis models that combine both data sources to arrive at a single earnings figure. We apply the methods to a Swedish data set. Our results show that register earnings data perform poorly if there is a (small) probability of a mismatch. Survey earnings data are more reliable, despite their measurement error. Predictors that combine both and take conditional class probabilities into account outperform all other predictors.

Keywords

Factor score; structural equation model; finite mixture; administrative data; validation study

1 Introduction

When both survey and administrative data are available, for example measuring individual earnings, the question arises how to make the best use of the available information in empirical applications. Measurement error in individual earnings data has received considerable attention in the literature, see Bound et al. (2001) for an extensive overview. Typically, the presence and extent of measurement error are studied by comparing survey answers with matched administrative data, either from company records or government agencies (e.g., Bound and Krueger, 1991; Pischke, 1995). In these studies, it is assumed that the administrative data are error-free. One of the findings in this literature has been that measurement error is nonclassical in the sense that it is negatively correlated with true earnings (mean reverting). Kapteyn and Ypma (2007) conduct a similar study, but they allow for errors in the matching of the administrative data to the survey data, and as a consequence, the administrative data may be incorrect. One of their findings is that mean reversion almost disappears after allowing for mismatch. This is an important finding as it calls into question the approach commonly adopted in empirical applications of discarding the self-reported data in favor of the supposedly more accurate administrative data. While using administrative data may solve some of the precision issues it may at the same time introduce new issues—a fact that has largely been ignored so far. In case of a mismatch, for example, the (reliably measured) earnings of one particular individual are then attributed to another individual, introducing an error that may be much worse than the imprecision of a self-reported figure. It seems intuitive that one could make better use of the available information by combining the observations from the survey data and the administrative data.

The objective of this paper is to derive several predictors of a single earnings variable and compare their statistical properties. We build on the model by Kapteyn and Ypma (2007), who introduce a mixture factor analysis model for survey and register earnings data. They estimate their model using data from Sweden. In this model, the true earnings variable is a latent variable. Our aim is to predict this latent variable taking the parameter estimates as given. In fact, throughout the paper, we abstract from model specification, identification, and estimation issues.

While both mixture models and factor score prediction have been extensively studied, the intersection of the two has received very little attention. The work by Zhu and Lee (2001) is most closely related to ours. They offer a Bayesian approach, proposing the Gibbs sampler to estimate the joint posterior distribution of the parameters and the latent variables, conditional on the observed variables. However, they do not explicitly consider prediction issues, which is the focus of our paper.

In our search for an optimal combination of survey and administrative data, we develop a theory of factor score prediction in mixture factor analysis models. We formulate the prediction problem in a generalized manner so that it covers the whole class of mixtures of linear structural equation models as well as mixtures of random coefficient linear regression models.

In the next section we introduce the Kapteyn-Ypma model. This model is a special case of a mixture factor analysis model, which we present in its full generality in section 3. To introduce notation and to provide the context for discussing the mixture case we briefly recapitulate the basics of factor score prediction in section 4. In section 5 we develop an array of predictors for the mixture model. In section 6, we adapt the Kapteyn-Ypma model to our model specification, apply the various predictors from section 5 and compare their performance in terms of mean squared error and reliability. Section 7 concludes.

2 The Kapteyn-Ypma model

The data used in Kapteyn and Ypma (2007) come from Sweden and were collected as part of a validation study to inform subsequent data collection efforts in the Survey of Health, Ageing and Retirement in Europe (SHARE). The objective was to compare self-reported information obtained in a survey with information obtained from administrative data sources (registers) to learn about measurement error, in particular in respondents' self-reports. The survey was fielded in the spring of 2003, shortly after tax preparation season, when people are most likely to have relevant financial information fresh in their minds. The sample was designed to cover individuals age 50 and older and their spouses; it was drawn from an existing database of administrative records (LINDA) which contains a five percent random subsample of the entire Swedish population. Out of 1431 sampled individuals 881 responded. Kapteyn and Ypma (2007) investigate the relation between survey data and register data for three different variables. We use their model on individual earnings without covariates, which is based on 400 observations with positive values recorded both in the register and in the survey data.

In the Kapteyn-Ypma model, there is a single latent variable ξ , representing the true value of the logarithm of earnings of an individual. (We omit the index denoting the individual.) This true value is not measured directly. However, there are two indicators available, one coming from administrative data and the other from individuals' self-reports collected in a survey. Figure 1 depicts the observations: the logarithms of the survey earnings are plotted against the logarithms of the register earnings. Most points are close to the 45° line. Thus, for most respondents, earnings as reported in the survey are close to the earnings observed in the administrative records. Still, for a noticeable number of points this is not the case.

Unlike most of the literature in this area, the Kapteyn-Ypma model does not assume that this is necessarily due to measurement error in the survey earnings. Rather, Kapteyn and Ypma assume that *both* indicators may contain error. However, the structure of the error differs between the indicators.

Let ξ and three other random variables, ζ , η , and ω , be independent of each other and normally distributed with means and variances (μ_ξ, σ_ξ^2) , $(\mu_\zeta, \sigma_\zeta^2)$, $(\mu_\eta, \sigma_\eta^2)$, and $(\mu_\omega, \sigma_\omega^2)$, respectively. Let r denote the value of log earnings from the administrative (register) data and let s be the corresponding value from the survey.

Kapteyn and Ypma (2007) assumed that any errors in the administrative data are due to a mismatch. Let π_r be the probability with which the observed administrative value, r , equals the true value of the respondent's earnings, ξ . Then $1 - \pi_r$ is the probability of a mismatch. In case of a mismatch, the administrative value r corresponds to the true value of someone else. This mismatched value will be denoted by ζ . According to the assumptions made above, ζ is not correlated with ξ . According to the same assumptions, the means and variances of ζ and ξ are allowed to differ. This reflects the background of the data collection. The survey data only cover a subset of the population restricted to individuals age 50 and older and their spouses, but a mismatch may involve an individual from the entire administrative database, which is a random sample of the entire Swedish population. To formalize this part of the model, the register data r are mixture normal with

$$r = \begin{cases} \xi & \text{with probability } \pi_r \\ \zeta & \text{with probability } 1 - \pi_r, \end{cases} \quad \begin{matrix} \text{(R1)} \\ \text{(R2)} \end{matrix}$$

where the first case reflects a match and the second case reflects a mismatch.

The second part of the model covers the survey data. Three cases are distinguished: earnings are either reported without error, with error, or with error and contamination. The probability that the observed survey value is correct is denoted by π_s . With probability $1 - \pi_s$ the survey data contain response error, part of which is mean-reverting, as expressed by the term $\rho(\xi - \mu_\xi)$, where $\rho < 0$ implies a mean-reverting response error in the sense of Bound and Krueger (1991). When the data contain response error, there is a probability of π_ω that the data are contaminated, modeled by adding an additional error term, ω . Contamination can, for example, result from erroneously reporting monthly earnings as annual or vice versa. Collecting the elements of this second part of the model, the survey data s are mixture normal with

$$s = \begin{cases} \xi & \text{with probability } \pi_s \\ \xi + \rho(\xi - \mu_\xi) + \eta & \text{with probability } (1 - \pi_s)(1 - \pi_\omega) \\ \xi + \rho(\xi - \mu_\xi) + \eta + \omega & \text{with probability } (1 - \pi_s)\pi_\omega. \end{cases} \quad \begin{matrix} \text{(S1)} \\ \text{(S2)} \\ \text{(S3)} \end{matrix}$$

Having established the distributions of r and s separately, the distribution of (r, s) is a mixture of bivariate normals, with $2 \times 3 = 6$ classes.

3 The general mixture factor analysis model

As we will show in section 6, the Kapteyn-Ypma model is a special case of a mixture factor analysis model, which has been applied in the literature before. Therefore, it is worthwhile to develop our theory in a generalized form.

Let $\mathbf{y}_n$ be a vector of observed random variables for observation n . We assume that there are J types or classes of observations, denoted by $j = 1, \dots, J$. Let π_{nj} be the (unconditional) probability that observation n is from class j . In our model, π_{nj} will typically be strictly between 0 and 1, so that class membership is not known with certainty, although we discuss known membership below as well. In each class the standard factor analysis (FA) model (e.g., Wansbeek and Meijer, 2000, chap. 7) is assumed to hold under normality. Thus, it is assumed that an observation from class j is generated from the model

$$\mathbf{y}_n = \tau_{nj} + \mathbf{B}_{nj}\xi_n + \varepsilon_n, \quad (1)$$

where $\xi_n \sim \mathcal{N}(\kappa_{nj}, \Phi_{nj})$ is a vector of unobserved (latent) random variables for observation n and $\varepsilon_n \sim \mathcal{N}(0, \Omega_{nj})$ is a vector of random errors, which are assumed to be independent of ξ_n , given the value of j . The errors may be interpreted as measurement errors or disturbances, depending on the application. Our aim is to predict ξ_n given an observed value of $\mathbf{y}_n$.

The vector of observed variables $\mathbf{y}_n$ follows a multivariate finite mixture distribution, in particular a finite mixture of multivariate normals. Because of the FA structure of (1), this is a mixture FA model (e.g., Yung, 1997). Unlike the standard FA case where observations are routinely centered and hence intercepts play no role, we explicitly incorporate intercepts since they will generally vary across classes and constitute an essential ingredient of the predictors. The parameter vectors τ_{nj} and κ_{nj} and matrices $\mathbf{B}_{nj}$, Φ_{nj} and Ω_{nj} are allowed to differ across classes.

We have not attached j subscripts to ξ_n and ε_n because our aim is to predict ξ_n and it is conceptually easier if the meaning of this is the same for all classes. For example, in the Kapteyn-Ypma model, ξ_n captures the true value of log earnings, which is defined irrespective of whether earnings are reported with measurement error or whether there is a mismatch of records, or both. However, we allow the distributions of ξ_n and ε_n to be different for different classes, so omitting the j subscript does not reduce the generality of the model. Applications with class-specific factors ξ_{nj} can be incorporated into our notation by stacking all ξ_{nj} , $j = 1, \dots, J$ into the vector ξ_n , and letting the columns of $\mathbf{B}_{nj}$ corresponding with ξ_{nk} , $k \neq j$ be zero. Our analysis below accommodates rank-deficient matrices, so our setup allows for such a specification.

In the notation used so far, we have also allowed the parameter vectors τ_{nj} and κ_{nj} , the parameter matrices $\mathbf{B}_{nj}$, Φ_{nj} and Ω_{nj} , and the class probabilities π_{nj} to differ across observations n . In a typical application of mixture FA models, the parameters will be assumed constant within the class, but allowing them to be observation-specific includes several models of interest. In particular, this includes all mixtures of linear structural equation models (e.g., Jedidi et al., 1997) and mixtures of linear random coefficient models. These models write the parameters as functions of covariates or of “deeper” parameters, or both. See the supplemental online appendix for further elaboration.

In line with the factor score prediction literature, we assume the parameters to be known and we abstract from estimation issues, and hence from issues of identification. Furthermore, we assume independence of observations. Consequently, only the observed variables $\mathbf{y}_n$ and observation-specific parameters of observation n are relevant for our purpose of predicting the latent variable ξ_n . All other observations can be ignored. Therefore, in the remainder of this paper, we drop the observation index n . Hence, our model for a generic observation from class j is now written as

$$\mathbf{y} = \boldsymbol{\tau}_j + \mathbf{B}_j \boldsymbol{\xi} + \boldsymbol{\varepsilon}, \quad (2a)$$

$$\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{\kappa}_j, \boldsymbol{\Phi}_j), \quad (2b)$$

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \boldsymbol{\Omega}_j), \quad (2c)$$

and the probability that the observation falls in class j is π_j . A direct implication of (2) is

$$\begin{pmatrix} \mathbf{y} \\ \boldsymbol{\xi} \end{pmatrix} \Big| j \sim \mathcal{N} \left[\begin{pmatrix} \boldsymbol{\mu}_j \\ \boldsymbol{\kappa}_j \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_j & \mathbf{B}_j \boldsymbol{\Phi}_j \\ \boldsymbol{\Phi}_j \mathbf{B}_j' & \boldsymbol{\Phi}_j \end{pmatrix} \right], \quad (3)$$

with $\boldsymbol{\mu}_j \equiv \boldsymbol{\tau}_j + \mathbf{B}_j \boldsymbol{\kappa}_j$ and $\boldsymbol{\Sigma}_j \equiv \mathbf{B}_j \boldsymbol{\Phi}_j \mathbf{B}_j' + \boldsymbol{\Omega}_j$. The mean $\boldsymbol{\mu}_y$ and covariance matrix $\boldsymbol{\Sigma}_y$ of $\mathbf{y}$, the mean $\boldsymbol{\mu}_{\boldsymbol{\xi}}$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}$ of $\boldsymbol{\xi}$, and the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\xi}y}$ of $\boldsymbol{\xi}$ and $\mathbf{y}$ follow by combining these expressions and the class probabilities π_j .

4 Factor score prediction with one class

For reference later on, we recapitulate some basics of factor score prediction when there is just one class. Given our earlier assumptions, all variables are normally distributed in this case, although a large part of the theory of factor score prediction still holds under nonnormality, see, e.g. Wansbeek and Meijer (2000, sec. 7.3). We use the notation as above, but without the class identifier j . Let

$$\widehat{\boldsymbol{\xi}} \equiv \boldsymbol{\kappa} + \boldsymbol{\Phi} \mathbf{B}' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \quad (4)$$

$$\mathbf{F} \equiv \boldsymbol{\Phi} - \boldsymbol{\Phi} \mathbf{B}' \boldsymbol{\Sigma}^{-1} \mathbf{B} \boldsymbol{\Phi}. \quad (5)$$

Using the expression for the conditional normal distribution we obtain $\boldsymbol{\xi} \mid \mathbf{y} \sim \mathcal{N}(\widehat{\boldsymbol{\xi}}, \mathbf{F})$. We consider the minimization of the mean-squared error of a function of $\mathbf{y}$ as a predictor of $\boldsymbol{\xi}$. Using a standard result in statistics, the MSE is minimized by the conditional expectation. So $\widehat{\boldsymbol{\xi}}$ is the minimum MSE predictor of $\boldsymbol{\xi}$.

In general, for reasons of simplicity, we are particularly interested in predictors that are linear in $\mathbf{y}$. Since the minimum MSE predictor is already linear in $\mathbf{y}$, imposing the restriction that our minimum MSE predictor be linear leads to $\widehat{\boldsymbol{\xi}}_L = \widehat{\boldsymbol{\xi}}$, where the subscript L denotes restriction to the class of linear predictors. Below, we will see that this does not carry over to the mixture model and that linearity is a binding restriction there.

Now suppose that we view (2) as a linear regression model, with a fixed unknown coefficient vector $\boldsymbol{\xi}$ that we would like to estimate, instead of as a random vector that we would like to predict. Or, equivalently, consider conditioning on the true value of $\boldsymbol{\xi}$. (See Anderson and Rubin 1956, for these two views.) Then it follows that the predictor $\widehat{\boldsymbol{\xi}}$ is biased in the sense that

$$E(\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi} \mid \boldsymbol{\xi}) = -(\mathbf{I} - \boldsymbol{\Phi} \mathbf{B}' \boldsymbol{\Sigma}^{-1} \mathbf{B})(\boldsymbol{\xi} - \boldsymbol{\kappa}) \neq 0.$$

This may be considered a drawback of $\widehat{\xi}$. An unbiased predictor can be obtained as follows. The model, rewritten as $\mathbf{y} - \boldsymbol{\tau} = \mathbf{B}\boldsymbol{\xi} + \mathbf{e}$, can be interpreted as a regression model with $\boldsymbol{\xi}$ as the vector of regression coefficients, $\mathbf{B}$ the matrix of regressors, and $\mathbf{y} - \boldsymbol{\tau}$ as the dependent variable. In this model the GLS estimator is the best linear unbiased estimator (BLUE) of $\boldsymbol{\xi}$, and under normality it is also BUE as it attains the Cramér-Rao bound (e.g., Amemiya 1985, sec. 1.3.3). So, letting subscript U denote an unbiasedness restriction in the sense of

$E(\widehat{\xi} \mid \xi) = \xi$, and LU the predictor that is obtained by imposing both linearity and unbiasedness, we obtain that

$$\widehat{\xi}_U = \widehat{\xi}_{LU} = (\mathbf{B}'\Omega^{-1}\mathbf{B})^{-1}\mathbf{B}'\Omega^{-1}(\mathbf{y} - \boldsymbol{\tau}) = \kappa + (\mathbf{B}'\Omega^{-1}\mathbf{B})^{-1}\mathbf{B}'\Omega^{-1}(\mathbf{y} - \mu) \quad (6)$$

minimizes the conditional MSE, $E[(\widehat{\xi} - \xi)^2 \mid \xi]$, for every possible value of ξ , conditional on unbiasedness. This implies that it must also minimize the unconditional MSE, $E[(\widehat{\xi} - \xi)^2] = E\{E[(\widehat{\xi} - \xi)^2 \mid \xi]\}$, among unbiased predictors and thus that it is the best unbiased predictor of ξ .

5 A taxonomy of predictors

With known class membership, the problem of factor score prediction in the mixture FA model reduces to factor score prediction in the single class model discussed in the previous section. The resulting within-class predictors have the same optimality properties as in the single class model. In general, class membership is unknown, and there appear to be three natural ways to proceed: (1) compute the within-class predictors for each class and combine them in a weighted average; (2) predict class membership and then use the within-class predictor for the predicted class; and (3) derive predictors that minimize the total mean squared prediction error.

The weighted average is based on the analogy with the means of $\mathbf{y}$ and $\boldsymbol{\xi}$. In these means, the (unconditional) class probabilities π_j are the appropriate weights, but for predictors of $\boldsymbol{\xi}$, we can also consider using *conditional* class probabilities given $\mathbf{y}$. Using Bayes' rule, these conditional probabilities are

$$p_j(\mathbf{y}) \equiv \Pr(\text{class}=j \mid \mathbf{y}) = \frac{\pi_j f_j(\mathbf{y})}{\sum_k \pi_k f_k(\mathbf{y})}, \quad (7)$$

where $f_j(\mathbf{y})$ is the normal density with mean $\boldsymbol{\mu}_j$ and variance $\boldsymbol{\Sigma}_j$ evaluated in $\mathbf{y}$. We will consider both alternatives.

The two-stage predictor would share the desired statistical properties of its second-stage single-class predictor if the prediction of class membership in the first stage were error-free. We expect that if class membership can be predicted with a high probability, or if the within-class predictors for the most likely misclassifications in this first stage are very similar, then the two-stage predictor will have desirable properties, such as small MSE.

The system-wide predictors that minimize total MSE apply the methods for deriving single-class predictors to the mixture problem. If we can find solutions to these problems, these are optimal by construction, but their expressions may be complicated and it may not be possible to find solutions.

These three general approaches, the unconditional and conditional weighting for the weighted predictors, and the linearity and/or unbiasedness restrictions lead to an array of predictors for the mixture model, which can be grouped into a taxonomy. The resulting predictors are grouped in Table 1. The rows indicate the approaches and the type of weighting. The columns reflect the linearity and unbiasedness restrictions. In the next subsections we systematically develop all these predictors, although this turns out to be impossible for two of the system-wide predictors (in brackets in the table).

5.1 Within-class predictors

The first row of Table 1 presents the predictors from the classes. Within a class the distribution is normal, so the linearity restriction is redundant, as argued above, and just two instead of four predictors are given. We extend the notation of section 4 by adding the class subscript j .

Note that the first predictor, $\widehat{\xi}_{u,j}$ is placed under the heading ‘unbiased’ due to the reasoning along the lines of a single class; if class j were the only class present, this predictor would be unbiased, but this quality is lost in a multiclass context. It is of interest to elaborate this. From (6),

$$\widehat{\xi}_{u,j} = (\mathbf{B}_j' \Omega_j^{-1} \mathbf{B}_j)^{-1} \mathbf{B}_j' \Omega_j^{-1} (\mathbf{y} - \tau_j) \equiv \mathbf{B}_j^- (\mathbf{y} - \tau_j),$$

where the notation $\mathbf{B}_j^-$ is chosen since it is a generalized inverse of $\mathbf{B}_j$. Consequently,

$$E(\widehat{\xi}_{u,j} \mid \xi) = \mathbf{B}_j^- (E(\mathbf{y} \mid \xi) - \tau_j).$$

Let

$$q_j(\xi) \equiv \Pr(\text{class}=j \mid \xi) = \frac{\pi_j g_j(\xi)}{\sum_k \pi_k g_k(\xi)}, \quad (8)$$

where $g_j(\xi)$ is the normal density with mean $\boldsymbol{\tau}_j$ and variance $\boldsymbol{\Phi}_j$ evaluated in ξ . Then

$$E(\mathbf{y} \mid \xi) = \sum_j q_j(\xi) E(\mathbf{y} \mid \xi, j) = \sum_j q_j(\xi) (\tau_j + \mathbf{B}_j \xi) = \bar{\tau}(\xi) + \bar{\mathbf{B}}(\xi) \xi, \quad (9)$$

with $\bar{\tau}(\xi) \equiv \sum_j q_j(\xi) \tau_j$, $\bar{\mathbf{B}}(\xi) \equiv \sum_j q_j(\xi) \mathbf{B}_j$, and $E(\widehat{\xi}_{u,j} \mid \xi) = \mathbf{B}_j^- [\bar{\tau}(\xi) - \tau_j + \bar{\mathbf{B}}(\xi) \xi]$. For unbiasedness, the latter should be equal to ξ for all ξ . This will only hold in special cases. The only case we know of is when $\boldsymbol{\tau}_j$ and $\mathbf{B}_j$ do not vary across classes. The heterogeneity is then restricted to the case where the intercepts and factor loadings are homogeneous, and only the distributions of ξ and $\boldsymbol{\epsilon}$ have a mixture character.

5.2 Weighted predictors

Given the predictors for a single class, a natural way to proceed to obtain a predictor in the mixture model is to combine them into a weighted average of the predictors per class. Potential weights for these weighted averages are the unconditional class membership probabilities π_j , by analogy with the mean of ξ , and the conditional class membership probabilities $p_j(\mathbf{y})$. The second row of Table 1 lists the predictors using the unconditional probabilities,

$$\widehat{\xi}_U = \sum_j \pi_j \widehat{\xi}_{U,j} \quad \text{and} \quad \widehat{\xi} = \sum_j \pi_j \widehat{\xi}_j.$$

Using the results from section 5.1, we can write

$$E(\widehat{\xi}_U \mid \xi) = \sum_j \pi_j \mathbf{B}_j^- \left[\bar{\tau}(\xi) - \tau_j + \bar{\mathbf{B}}(\xi) \xi \right] = \mathbf{B}^\dagger \left[\bar{\tau}(\xi) + \bar{\mathbf{B}}(\xi) \xi \right] - \tau^\dagger,$$

where $\mathbf{B}^\dagger = \sum_j \pi_j \mathbf{B}_j^-$ and $\tau^\dagger = \sum_j \pi_j \mathbf{B}_j^- \tau_j$. Again, $\widehat{\xi}_U$ will be unbiased only in special cases.

The results for weighting with the conditional class membership probabilities are displayed in the next row of the table. An asterisk distinguishes these predictors from the previous ones. Then

$$\widehat{\xi}_U^* = \sum_j p_j(\mathbf{y}) \widehat{\xi}_{U,j} \quad \text{and} \quad \widehat{\xi}^* = \sum_j p_j(\mathbf{y}) \widehat{\xi}_j. \quad (10)$$

Notice that the weights depend nonlinearly on $\mathbf{y}$ and hence the linearity property is lost when weighting. Also, the predictor $\widehat{\xi}_U^*$ will generally not be unbiased.

5.3 Two-stage predictors

An intuitively appealing alternative to computing a weighted predictor is to predict class membership first and then proceed as if class membership was known and use the class-specific predictor of the predicted class. Hence, this leads to two-stage predictors, analogous to two-step estimators. The obvious class predictor is the class with the highest conditional probability, i.e., $\hat{J} \equiv \arg \max_j p_j(\mathbf{y})$. Then the two-stage predictors, listed in the fourth row of Table 1, are

$$\widehat{\xi}_U^{(2)} = \widehat{\xi}_{U,\hat{J}} \quad \text{and} \quad \widehat{\xi}^{(2)} = \widehat{\xi}_{\hat{J}}.$$

Again, unbiasedness and linearity are lost.

5.4 System-wide predictors

The bottom row of Table 1 presents the system-wide predictors that are obtained by minimizing the MSE with or without the restrictions. Since in the mixture model normality does not hold any longer, linearity is not obtained as a byproduct, and the full range of four different predictors has to be considered. We use a tilde to denote system-wide predictors.

Prediction under linearity and unbiasedness—The starting point for finding an unbiased predictor that is linear in $\mathbf{y}$ is to find a matrix $\mathbf{L}$ and a vector $\mathbf{b}$ such that

$\tilde{\xi}_{LU} \equiv \mathbf{L}'\mathbf{y} + \mathbf{b}$ satisfies $E(\tilde{\xi}_{LU} \mid \xi) = \xi$ for all ξ . Using (9) we obtain

$$E(\tilde{\xi}_{LU} \mid \xi) = \mathbf{L}' E(\mathbf{y} \mid \xi) + \mathbf{b} = \mathbf{L}' \left[\bar{\tau}(\xi) + \bar{\mathbf{B}}(\xi) \xi \right] + \mathbf{b}.$$

This should be equal to ξ for any ξ , so also for $\xi = \mathbf{0}$. Consequently, $\mathbf{b} = -\mathbf{L}' \bar{\tau}(0)$. Hence

$$\mathbf{L}' \left[\bar{\tau}(\xi) - \bar{\tau}(0) + \bar{\mathbf{B}}(\xi)\xi \right] = \xi$$

should hold for all ξ . Because of the (nonlinear) dependence of $\bar{\tau}(\xi)$ and $\bar{\mathbf{B}}(\xi)$ on ξ , a matrix $\mathbf{L}$ with this property only exists under special circumstances. We conjecture that it only exists when τ_j and $\mathbf{B}_j$ do not vary across classes. Hence, a linear unbiased predictor will generally not exist. If τ_j and $\mathbf{B}_j$ do not vary across classes, the model can be reinterpreted as a single-class FA model, in which the factors and errors follow nonnormal distributions, in particular mixture normal distributions. Then the minimum MSE linear unbiased predictor follows from the expressions in section 4, which do not depend on normality.

Prediction under unbiasedness—Imposing linearity may preclude a solution. We relax this requirement, but maintain unbiasedness. We then consider a predictor $\tilde{\xi}_U = \mathbf{h}(\mathbf{y})$, where $\mathbf{h}(\mathbf{y})$ is a function of $\mathbf{y}$ such that $E(\mathbf{h}(\mathbf{y}) \mid \xi = \xi) = \xi$ for all ξ . Using (8), this requirement for $\mathbf{h}(\mathbf{y})$ can be rewritten as

$$\sum_j \pi_j g_j(\xi) \{E[\mathbf{h}(\mathbf{y}) \mid \xi, j] - \xi\} = 0$$

for all ξ . Attempts to derive the predictor led to expressions without tractable solutions, except for the special case in which τ_j and $\mathbf{B}_j$ do not vary across classes, which was discussed above.

Prediction under linearity—This case is conceptually the simplest one. It is based on the idea of linear projection. According to a well-known result (e.g., Angrist and Pischke, 2009, Theorem 3.1.5 for the univariate case in a slightly different form), the MSE is minimized over all linear functions by

$$\tilde{\xi}_L = \mu_\xi + \Sigma_{\xi y} \Sigma_y^{-1} (\mathbf{y} - \mu_y). \quad (11)$$

Note that the structure of the model is only used in computing the moments μ_ξ , μ_y , $\Sigma_{\xi y}$, and Σ_y . Derivation and optimality (minimum MSE within the class of linear predictors) of the predictor follow from general principles and do not use the model structure explicitly.

Prediction without restrictions—The approach here rests on the fact that the MSE is minimized by the mean of the conditional distribution. As with the linear projection result above, this result is generic and does not depend on the structure of the model. Of course, for the implementation, we need an expression of this conditional mean, which is derived from the model structure. Adapting results from section 4, we have $\xi \mid \mathbf{y}, j \sim \mathcal{N}(\tilde{\xi}_j, \mathbf{F}_j)$, with $\tilde{\xi}_j \equiv \kappa_j + \Phi_j \mathbf{B}_j' \Sigma_j^{-1} (\mathbf{y} - \mu_j)$ and $\mathbf{F}_j \equiv \Phi_j - \Phi_j \mathbf{B}_j' \Sigma_j^{-1} \mathbf{B}_j \Phi_j$. So $\xi \mid \mathbf{y}$ is mixture normal with class membership probabilities $\Pr(\text{class} = j \mid \mathbf{y}) = p_j(\mathbf{y})$ with $p_j(\mathbf{y})$ given by (7). Thus, the system-wide predictor without restrictions is given by

$$\tilde{\xi} = E(\xi \mid \mathbf{y}) = \sum_j \Pr(\text{class} = j \mid \mathbf{y}) E(\xi \mid \mathbf{y}, j) = \sum_j p_j(\mathbf{y}) \tilde{\xi}_j,$$

which equals $\widehat{\xi}^*$, the predictor obtained by weighting of the unrestricted predictors per class using the conditional class probabilities. The conditional variance of this predictor is

$$\text{Var}(\xi \mid \mathbf{y}) = \sum_j p_j(\mathbf{y}) \mathbf{F}_j + \sum_j p_j(\mathbf{y}) \widehat{\xi}_j \widehat{\xi}_j' - \widehat{\xi} \widehat{\xi}'.$$

The first term on the right-hand side is the weighted average of the variances within the classes, and the other terms represent the between-class variance.

5.5 Extensions

The derivations up till now assumed that all covariance matrices are nonsingular and that no extraneous information is available about class membership. However, in some cases, including the Kapteyn-Ypma model, these assumptions are not met. In this section, we generalize our methods to deal with this.

In the case of a *labeled class*, it is known a priori to which class an observation belongs, or this follows deterministically from the covariates. Thus, we have observed (measured) rather than unobserved (unmeasured) heterogeneity. More generally, we may know that an observation belongs to a certain subset of the classes. This can be viewed as a limiting case of observation-specific unconditional class probabilities π_{nj} , and thus is accommodated by our general setup.

Degeneracy denotes singular covariance matrices or factor loadings matrices that do not have full column rank. Several types of degeneracies are conceivable in a single-class factor analysis model. For substantive reasons, the errors ϵ of some of the indicators may be assumed to be identically zero, leading to a singular covariance matrix $\mathbf{\Omega}$. This may also be the result of the estimation procedure without a priori restrictions, i.e., a *Heywood case*. Similarly, the covariance matrix $\mathbf{\Phi}$ of the factors may be singular. Finally, the factor loadings matrix $\mathbf{B}$ may not be of full column rank. The latter does not necessarily lead to a degenerate distribution, but it does imply that some of the formulas presented earlier cannot be applied. In a single-class factor analysis setting, singularity of $\mathbf{\Phi}$ or deficient column rank of $\mathbf{B}$ are not particularly interesting or relevant, because the model is then equivalent to a model with fewer factors and thus can be reduced to a more parsimonious one. However, in a mixture setting, such forms of degeneracy may occur quite naturally. If $\mathbf{\Omega}$ is nonsingular, then the covariance matrix $\mathbf{\Sigma}$ of the observed indicators is also nonsingular. Thus, a necessary condition for singularity of $\mathbf{\Sigma}$ is singularity of $\mathbf{\Omega}$.

The formulas for the single-class predictors cannot be directly applied in the case of degeneracies, because the required inverses may not exist. Furthermore, when there are classes with degenerate distributions of the indicators $\mathbf{y}$, (7) does not hold as $\mathbf{y}$ does not have a proper density. The supplemental online appendix gives the expressions for the within-class predictors and conditional probabilities in the case of degeneracies.

With these adapted expressions, prediction of the factor scores is relatively straightforward. For the non-system-wide predictors, we can apply the standard expressions, using the adapted formulas for the within-class predictors and the conditional probabilities. For the system-wide linear predictor, (3) is still valid, and thus the predictor $\tilde{\xi}_L$ is still given by (11), with Σ_y^{-1} replaced by the Moore-Penrose generalized inverse Σ_y^+ if it is singular. The system-wide unrestricted predictor $\tilde{\xi}$ is again equivalent to the conditionally weighted unrestricted predictor $\widehat{\xi}^*$. We will apply these methods in section 6.

6 Application to the Kapteyn-Ypma model

We can now express the parameters of the general mixture FA model in terms of the parameters of the Kapteyn-Ypma model. We have $\mathbf{y} = (r, s)'$ and $\kappa_j = \mu_{\xi}$ and $\Phi_j = \sigma_{\xi}^2$ for all classes. The correspondence between the remaining parameters and the general setup is shown in Table 2. A notable feature is that Ω_j is singular for classes 1 through 4. So there is a degree of degeneracy in the model, cf. the discussion in section 5.5. Moreover Σ_1 is singular, because in class 1, $\mathbf{y}$ takes on values in the subspace $y_1 = y_2$, i.e., $r = s$, whereas Σ_j is nonsingular for $j = 2, \dots, 6$.

6.1 Predictors

Given this structure, we now turn to the issue of prediction. We discuss a number of predictors, following the taxonomy presented in Table 1.

Within-class predictors—Formulas for the predictors per class are given in Table 3.

Notice that for classes 5 and 6, the expressions for $\widehat{\xi}_{u,j}$ can be obtained from the corresponding expressions for $\widehat{\xi}_j$ by setting $\sigma_{\eta}^2 = \sigma_{\omega}^2 = 0$, i.e., by ignoring the measurement error variances in the computation.

Weighted predictors—Moving down the rows of Table 1, we obtain overall predictors by weighting the $\widehat{\xi}_{u,j}$ and $\widehat{\xi}_j$ with the probabilities π_j to obtain $\widehat{\xi}_u = ar + bs + c$ and $\widehat{\xi} = a'r + b's + c'$, where a, b, c, a', b' , and c' depend on the various parameters. The detailed expressions can be found in the supplemental online appendix.

Conditionally weighted predictors—Consider the event $r = s$. If the observation is drawn from class 1, the conditional probability of observing this event is 1, whereas this conditional probability is 0 if the observation is drawn from any of the other five classes. Hence, the conditional probability that $j = 1$ given this event is 1: $p_1((r, s)) = 1$ for all r and $p_j((r, s)) = 0$ for $j = 2, \dots, 6$. Conversely, using the same logic, $p_j((r, s)) = 0$ for all $s \neq r$. In the latter case,

$$p_j(\mathbf{y}) = \frac{\pi_j f_j(\mathbf{y})}{\sum_{k=2}^6 \pi_k f_k(\mathbf{y})}, \quad j=2, \dots, 6,$$

where $f_j(\mathbf{y}) = (2\pi)^{-1} (\det \Sigma_j)^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{y} - \mu_j)' \Sigma_j^{-1} (\mathbf{y} - \mu_j) \right\}$. It follows that, if $r = s$, $\widehat{\xi}_u^* = \widehat{\xi}^* = (r+s)/2 = r = s$ and, if $r \neq s$,

$$\widehat{\xi}_u^* = \sum_{j=2}^6 p_j(\mathbf{y}) \widehat{\xi}_{u,j} \quad \text{and} \quad \widehat{\xi}^* = \sum_{j=2}^6 p_j(\mathbf{y}) \widehat{\xi}_j. \quad (12)$$

Two-stage predictors—In the first stage, class membership is predicted according to the maximum of the $p_j(\mathbf{y})$. Thus, for $r = s$, class 1 is predicted and for $r \neq s$, class 1 is dropped from consideration. Figure 2 shows which class is predicted for a given (r, s) combination with $r \neq s$. This class predictor was also computed for the observed data points by Kapteyn and Ypma (2007, Fig. C1). From this figure, we see that close to the 45° line, where most of the observations lie in Figure 1, membership of class 2 is predicted, and farther away from

this line but still in the middle of the distribution of r , membership of class 3 is predicted. Perhaps most interesting is the area in the middle-left where membership of class 5 is predicted. There are about a dozen points in this area in Figure 1, and this is where it is predicted that r is a mismatch. The resulting factor score predictor is the within-class predictor corresponding with the predicted class, given in Table 3.

System-wide predictors—The last row of Table 1 presents four predictors. As indicated in section 5.4, expressions for the two unbiased predictors $\tilde{\xi}_{LU}$ and $\tilde{\xi}_U$ could not be found. This leaves us with the projection-based predictor $\tilde{\xi}_L$ and the conditional expectation-based predictor $\tilde{\xi}$. The expression for the former is given in (11). We need to elaborate the mean and covariance matrices of $\mathbf{y}$ and ξ . By weighting the rows of Table 2 with the π 's we obtain, after some straightforward algebra, expressions for the elements of $\boldsymbol{\mu}_y$, $\boldsymbol{\Sigma}_y$ and $\boldsymbol{\Sigma}_{\xi y}$. Substitution in (11) yields again a linear expression, $\tilde{\xi}_L = a''r + b''s + c''$. The full expression for the coefficients is given in the supplemental online appendix. Finally, $\tilde{\xi} = \widehat{\xi}^*$, as was already stated in section 5.4. The formula for the latter has been given by (12).

6.2 Empirical results

We now turn to the empirical outcomes of the prediction process using the Kapteyn-Ypma estimates of the model parameters. Our objective is to predict earnings by combining the information from the survey and administrative data. We do so on the basis of a simple model, and of the estimation results from that model. Kapteyn and Ypma (2007) provide an extensive discussion of their results. Here, we only reproduce the relevant parameter estimates that we will use in implementing the various predictors. Estimated means and variances of the normal distributions are given in Table 4. The estimates of the other parameters are $\pi_r = .96$, $\pi_s = .15$, $\pi_\omega = .16$, and $\rho = -.01$.

Table 5 presents results per class and the resulting unconditionally weighted predictors. The fourth column gives the class probabilities. The largest class is the second one, representing the case of measurement error, but no mismatch. Columns 5–10 present the numerical results corresponding with Table 3. The predictors in Table 3 have been written as linear functions $ar + bs + c$ of r and s , the three weights a , b , and c being given for $\widehat{\xi}_{U,j}$ first and next for $\widehat{\xi}_j$.

As was already apparent from Table 3, the rows for classes 5 and 6 are the most interesting ones. These represent the cases of a mismatch, hence the value from the register is discarded and only the value from the survey is used; and the latter suffers from measurement error. If there is no (further) contamination, the unbiased predictor is obtained from the survey figure by slightly overweighting it to compensate for the mean-reverting measurement error. The unrestricted predictor is obtained by slightly downweighting it, due to shrinkage to the overall mean (cf. Wansbeek and Meijer 2000, p. 165). If there is contamination, the results diverge strongly. The unrestricted predictor is predominantly a constant, plus some effect from the survey, reflecting a large amount of shrinkage due to the large variance of the contamination term, so that more weight is given to the population mean. This phenomenon is sometimes called *borrowing strength* in the random coefficients literature (e.g., Raudenbush et al., 2004, p. 9). In contrast, the unbiased predictor is very similar to the predictor obtained when there is no contamination.

The last row of the table gives the results for the predictors $\widehat{\xi}_U$ and $\widehat{\xi}$ that are obtained by weighting the preceding rows by the unconditional class membership probabilities given in the second column. Apart from the constant c , the results from the two predictors only differ

from the third digit onwards, and imply an 89% weight to the register data and an 11% weight to the survey data. The linear projection-based predictor $\widehat{\xi}_L$ can also be expressed as a linear combination of r and s :

$$\widehat{\xi}_L = .22r + .55s + 2.85.$$

This is a striking result, with very low weight given to the register data, vastly unlike the case with $\widehat{\xi}_U$ and $\widehat{\xi}$.

The predictors discussed so far were all based on uniform weights across the (r, s) space. The conditionally weighted predictors $\widehat{\xi}_U^*$ and $\widehat{\xi}$ and the two-stage predictors $\widehat{\xi}_U^{(2)}$ and $\widehat{\xi}^{(2)}$ cannot be expressed as a simple function of r and s since the weights now depend on $\mathbf{y} = (r, s)'$ through the densities $p_j(\mathbf{y})$, cf. (10). For the two-stage predictors, Figure 2 indicates which class-specific predictor is used for a given value of (r, s) , and thus from Table 5 the corresponding weights can be found. Similarly, we can express the conditionally weighted predictors as “linear” combinations of r and s with the weights depending on the values of r and s : $\widehat{\xi} = \bar{a}(\mathbf{y})r + \bar{b}(\mathbf{y})s + \bar{c}(\mathbf{y})$, where $\bar{a}(\mathbf{y})$, $\bar{b}(\mathbf{y})$, and $\bar{c}(\mathbf{y})$ are weighted averages of the a , b , and c coefficients in the last three columns of Table 5, with weights $p_j(\mathbf{y})$, and similarly for $\widehat{\xi}_U^*$. We have computed the resulting relative weight given to r , as compared to s , computed as $\bar{a}(\mathbf{y}) / [\bar{a}(\mathbf{y}) + \bar{b}(\mathbf{y})]$, over the (r, s) space. The result for $\widehat{\xi}$ is given in Figure 3. The figure for $\widehat{\xi}_U^*$ is very similar and is omitted. We see that in the middle of the figure the predictor is (almost) fully dominated by r , but further from the mean of ξ ($\mu_\xi = 12.28$), more weight is attached to s . Since the variance of ξ is large, the relative weight largely depends on the value of r only and is rather insensitive to the value of s .

It is also interesting to compare Figure 3 with Figure 2. This shows that in the areas of the (r, s) space where classes 2 or 3 are predicted for the two-stage predictors, which corresponds to these two-stage predictors being equal to r and thus giving a 100% weight to r , the relative weight for r in $\widehat{\xi}$ and $\widehat{\xi}_U^*$ is also very high, more than 90%. Conversely, in the areas where classes 5 or 6 are predicted for the two-stage predictors, r is predicted to be a mismatch and its weight in computing the two-stage predictors is zero. In these areas, the relative weight of r in $\widehat{\xi}$ and $\widehat{\xi}_U^*$ is very low, less than 10%. The weight for r in $\widehat{\xi}$ and $\widehat{\xi}_U^*$ changes continuously, whereas it changes discretely for the two-stage predictors, but Figure 3 shows that “continuously” is still quite fast, so that the two-stage predictors and conditionally weighted predictors are very similar.

In order to obtain an overall comparison of the performance of the seven predictors we have computed their reliabilities. The reliability of a proxy is a measure for the precision with which it measures the latent variable of interest. When ξ is one-dimensional, the reliability of a proxy x is the squared correlation between ξ and x . The reliability of a linear combination $x = ar + bs + c$ can be straightforwardly obtained in closed form. For the conditionally weighted and two-stage predictors, reliabilities cannot be computed analytically since they are nonlinear combinations of r and s . We have computed the reliabilities of these predictors by simulating 100,000 draws from the model and computing the squared sample correlation between factor and predictor from these. (We have also computed the reliabilities of the linear proxies using these draws, which corresponded closely to the results obtained from the closed-form expressions for these proxies.)

The results are given in Table 6, for the seven predictors considered throughout, plus r and s as a reference. We also report the (unconditional) mean squared errors of the predictors, $MSE = E(\text{predictor} - \xi)^2$, which are, of course, strongly negatively related to the reliabilities, although they are not perfect substitutes. For example, unlike the reliability, the MSE of a linear combination $x = ar + bs + c$ does depend on the constant c . In addition, we present the (unconditional) bias $E(\text{predictor} - \xi)$ and (unconditional) variance $\text{Var}(\text{predictor} - \xi)$ of the prediction errors.

The table shows some striking results. The register data r have a squared correlation with the true log earnings ξ of less than .5. Informally stated, r is a clear loser, which is surprising. Apparently a high probability of being exactly correct is not sufficient and the small probability of being an independent draw from a different distribution has dramatic consequences for the statistical properties of the indicator. The survey data s perform considerably better than r . The predictors obtained by unconditional weighting per class perform poorly, but the predictor based on linear projection performs much better. However, all predictors that use the conditional class membership probabilities are nearly perfect (with $\tilde{\xi}$ performing marginally better than the others), again, of course, against the background of the postulated model.

In the last two columns of Table 6, we see that the biases of the first four predictors are sizeable, but the squared biases are negligible compared to the variances, so that the variances dominate the MSEs. It is also noteworthy that both r and s have negative bias. The bias of r is due to mismatch, whereas the bias of s is due to the negative means of the measurement error η and contamination ω . The biases of r and s are identified on the basis of the $r = s$ cases, which are higher up on the earnings distribution (on average) than the other cases.

The excellent performance in terms of reliability and MSE of the two-stage predictor, both in an absolute sense and compared to the minimum MSE predictor $\tilde{\xi}$, is related to the high probability with which class membership is correctly predicted. Overall, the probability of correctly classifying an observation is 96%. Among the misclassifications, a classification of an observation from class 3 as an observation from class 2 takes more than 2 of the remaining 4 percent. But because the within-class predictor is r in both class 2 and class 3, this has no consequence for the precision of the two-stage predictors. We expect that in situations where the class probabilities are less concentrated and the within-class predictors vary more widely across classes, the two-stage predictors will be noticeably less precise.

7 Discussion

This paper has dealt with a very simple question. We are given two numerical values for the same quantity, which are often not equal. What is the best guess as to the truth? As always, getting to a sensible answer involves some model building, necessarily based on some assumptions. The model in case was of the mixture factor analysis type, and answering our question amounted to factor score prediction in that model. Since that topic appeared to be hardly researched, we first derived the methods needed. Taking the standard factor score prediction literature as our point of departure, we systematically explored the options for constructing predictors for the mixture case. This produced seven predictors for further consideration.

We then explored the consequences of two extensions of the model. The first is the presence of labeled classes, which means that class membership is either known a priori or is a deterministic function of the covariates. The second extension is the incorporation of degenerate distributions and rank-deficient coefficient matrices in the model.

We applied the tools we developed to the case of Swedish earnings data, where measurements were available for each individual both from an administrative database and from a survey. We took the Kapteyn-Ypma (Kapteyn and Ypma, 2000) model as given, including their parameter estimates, and studied prediction of true (log) earnings from the two observed measurements. The major conclusion is that conditionally weighted predictors and the two-stage predictors perform extremely well, much better than either the register data or the survey data by themselves, predictors based on unconditional weights, or the linear-projection based estimator.

A topic for further research is to study the generality of our empirical findings. In particular, researchers increasingly use administrative data matched to surveys and simply treat the administrative data as if they contained the correct values. Our empirical results show that, if there is a potential for mismatch, this approach may be counterproductive, because we showed that in the Kapteyn-Ypma model, the survey earnings have a higher reliability and lower MSE than the register earnings.

In the U.S., administrative data from the Social Security Administration (SSA) are a primary source of such matched data sources. They have been matched to nationally representative surveys, such as the Health and Retirement Study (HRS), Survey of Income and Program Participation (SIPP), and the Current Population Survey (CPS). The algorithm that the SSA uses for matching administrative records to survey records tends to be conservative in declaring a correct match, but it is certainly not guaranteed to be error-free. Given the findings in this paper, it seems worthwhile to study the prevalence of mismatches using a model like the Kapteyn-Ypma one, and their potential consequences using an analysis like the one in the current paper.

There remains the puzzle why the register data perform so poorly. It is beyond the scope of this paper to resolve this puzzle, but a speculative explanation may be the following. In the Kapteyn-Ypma model, the register data are assumed to deviate from the true earnings values only in case of a mismatch. However, the register data might contain some measurement error themselves, possibly in the case of unreported earnings from a second or third job. Failure to report these earnings could, for example, be due to the earnings being minor or because they pertain to unreported employment in the underground economy (see Kapteyn and Ypma 2007, footnote 5, for a similar example). If this is the explanation, “mismatch” should be reinterpreted as “misreport”. Under such a reinterpretation, the model specification would likely need to be changed slightly, because then register earnings under misreport would not be independent from the true earnings. It is not immediately clear, though, whether they would be positively or negatively correlated with true earnings, and thus whether the reliability of register earnings as a measure of true earnings would be higher or even lower than estimated in the current paper. In any case, it seems that even under the most favorable circumstances, the quality of the register data is lower than is typically assumed in validation studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank Jelmer Ypma and Arie Kapteyn for sharing their code with us and for discussions about their model, and Jos ten Berge and conference participants at Aston Business School, Birmingham, in particular Battista Severgnini, for stimulating discussions and comments on an earlier version of this paper. We would also like to thank the editors (Arthur Lewbel and Keisuke Hirano), an associate editor, and a referee for helpful comments. The National Institute on Aging supported the collection of the survey data and subsequent match to

administrative records (grant R03 AG21780). Rohwedder acknowledges financial support from NIA grant P01 AG08291.

References

- Amemiya, T. Advanced econometrics. Harvard University Press; Cambridge, MA: 1985.
- Anderson, TW.; Rubin, H. Statistical inference in factor analysis. In: Neyman, J., editor. Proceedings of the third Berkeley symposium on mathematical statistics and probability V. University of California Press; Berkeley: 1956. p. 111-150.[Available from <http://projecteuclid.org/euclid.bsm/1200511860>]
- Angrist, JD.; Pischke, J-S. Mostly harmless econometrics: An empiricist's companion. Princeton University Press; Princeton, NJ: 2009.
- Bound, J.; Brown, C.; Mathiowetz, N. Measurement error in survey data. In: Heckman, JJ.; Leamer, E., editors. Handbook of econometrics. Vol. 5. North-Holland; Amsterdam: 2001. p. 3705-3843.
- Bound J, Krueger AB. The extent of measurement error in longitudinal earnings data: Do two wrongs make a right? *Journal of Labor Economics*. 1991; 9:1–24.
- Jedidi K, Jagpal HS, DeSarbo WS. Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Marketing Science*. 1997; 16:39–59.
- Kapteyn A, Ypma JY. Measurement error and misclassification: A comparison of survey and administrative data. *Journal of Labor Economics*. 2007; 25:513–551.
- Pischke J-S. Measurement error and earnings dynamics: Some estimates from the PSID validation study. *Journal of Business & Economic Statistics*. 1995; 13:305–314.
- Raudenbush, SW.; Bryk, AS.; Cheong, YF.; Congdon, R. HLM 6: Hierarchical linear and nonlinear modeling. Scientific Software International; Chicago: 2004.
- Wansbeek, TJ.; Meijer, E. Measurement error and latent variables in econometrics. North-Holland; Amsterdam: 2000.
- Yung Y-F. Finite mixtures in confirmatory factor-analysis models. *Psychometrika*. 1997; 62:297–330.
- Zhu H-T, Lee S-Y. A Bayesian analysis of finite mixtures in the LISREL model. *Psychometrika*. 2001; 66:133–152.

Figure 1.
Survey data versus register data.

Figure 2. Predicted class membership $\hat{\mathcal{J}}$ for each combination of register and survey data (excluding $r = s$).

Figure 3.

Relative weight of $\log(\text{register earnings})$ r in the conditionally weighted predictor $\tilde{\xi}$. The lines connect points with the same relative weight given to r , computed as $(\mathbf{y})/[(\mathbf{y}) + \tilde{h}(\mathbf{y})]$.

Table 1

A taxonomy of predictors in a mixture structural equation model.

	Restrictions			
	Linear unbiased	Unbiased	Linear	None
Within class (j)	$\hat{\xi}_{U,j}$		$\hat{\xi}_j$	
Weighted (unconditional)	$\hat{\xi}_U$		$\hat{\xi}$	
Weighted (conditional)	$\hat{\xi}_U^*$		$\hat{\xi}^*$	
Two-stage	$\hat{\xi}_U^{(2)}$		$\hat{\xi}^{(2)}$	
System-wide	$\tilde{\xi}_{LU}$	$\tilde{\xi}_U$	$\tilde{\xi}_L$	$\tilde{\xi} (= \hat{\xi}^*)$

Table 2

Parameterization of the Swedish earnings data model in terms of a mixture structural equation model.

Class (<i>j</i>)	<i>r</i>	<i>s</i>	π_j	τ_j	B_j	ω_j
1	R1	S1	$\pi_r \pi_s$	$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$
2	R1	S2	$\pi_r(1 - \pi_s)(1 - \pi_\omega)$	$\begin{pmatrix} 0 \\ \mu_\eta - \rho\mu_\xi \end{pmatrix}$	$\begin{pmatrix} 1 \\ 1 + \rho \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & \sigma_\eta^2 \end{pmatrix}$
3	R1	S3	$\pi_r(1 - \pi_s)\pi_\omega$	$\begin{pmatrix} 0 \\ \mu_\eta + \mu_\omega - \rho\mu_\xi \end{pmatrix}$	$\begin{pmatrix} 1 \\ 1 + \rho \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & \sigma_\eta^2 + \sigma_\omega^2 \end{pmatrix}$
4	R2	S1	$(1 - \pi_r)\pi_s$	$\begin{pmatrix} \mu_\zeta \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 1 \end{pmatrix}$	$\begin{pmatrix} \sigma_\zeta^2 & 0 \\ 0 & 0 \end{pmatrix}$
5	R2	S2	$(1 - \pi_r)(1 - \pi_s)(1 - \pi_\omega)$	$\begin{pmatrix} \mu_\zeta \\ \mu_\eta - \rho\mu_\xi \end{pmatrix}$	$\begin{pmatrix} 0 \\ 1 + \rho \end{pmatrix}$	$\begin{pmatrix} \sigma_\zeta^2 & 0 \\ 0 & \sigma_\eta^2 \end{pmatrix}$
6	R2	S3	$(1 - \pi_r)(1 - \pi_s)\pi_\omega$	$\begin{pmatrix} \mu_\zeta \\ \mu_\eta + \mu_\omega - \rho\mu_\xi \end{pmatrix}$	$\begin{pmatrix} 0 \\ 1 + \rho \end{pmatrix}$	$\begin{pmatrix} \sigma_\zeta^2 & 0 \\ 0 & \sigma_\eta^2 + \sigma_\omega^2 \end{pmatrix}$

Note: R1 = correct match; R2 = mismatch; S1 = no measurement error; S2 = measurement error; S3 = measurement error and contamination.

Table 3

Expressions for the within-class predictors for the Swedish earnings data model as functions of the parameters.

Class (<i>j</i>)	<i>r</i>	<i>s</i>	$\xi_{U,j}$	ξ_j
1	R1	S1	$\frac{1}{2}(r+s)$	$\frac{1}{2}(r+s)$
2	R1	S2	r	r
3	R1	S3	r	r
4	R2	S1	s	s
5	R2	S2	$\mu_{\xi} + \frac{1}{1+\rho} s'$	$\mu_{\xi} + \frac{(1+\rho)\sigma_{\xi}^2}{(1+\rho)^2\sigma_{\xi}^2 + \sigma_{\eta}^2} s'$
6	R2	S3	$\mu_{\xi} + \frac{1}{1+\rho} (s' - \mu_{\omega})$	$\mu_{\xi} + \frac{(1+\rho)\sigma_{\xi}^2}{(1+\rho)^2\sigma_{\xi}^2 + \sigma_{\eta}^2 + \sigma_{\omega}^2} (s' - \mu_{\omega})$

Notes: R1 = correct match; R2 = mismatch; S1 = no measurement error; S2 = measurement error; S3 = measurement error and contamination; $s' \equiv s - \mu_{\xi} - \mu_{\eta}$

Table 4

Parameter estimates for the Swedish earnings data model.

	ξ	ζ	η	ω
μ	12.28	9.20	-0.05	-0.31
σ^2	0.51	3.27	0.01	1.52

Source: Kapteyn and Ypma (2007).

Unconditional class membership probabilities for the Swedish earnings data model and predictors as linear expressions $ar + bs + c$.

Table 5

Class (j)	r	s	π_j	$\xi_{U,j}$			ξ_j		
				a	b	c	a	b	c
1	R1	S1	.15	0.5	0.5	0	0.5	0.5	0
2	R1	S2	.69	1	0	0	1	0	0
3	R1	S3	.13	1	0	0	1	0	0
4	R2	S1	.01	0	1	0	0	1	0
5	R2	S2	.03	0	1.01	-0.11	0	0.99	0.12
6	R2	S3	.01	0	1.01	0.19	0	0.25	9.32
				$\xi_U : 0.89$			$\xi : 0.89$		
				-0.00			0.11		
							0.05		

Notes: R1 = correct match; R2 = mismatch; S1 = no measurement error; S2 = measurement error; S3 = measurement error and contamination; probabilities do not sum to 1 due to rounding.

Table 6

Precision of the predictors.

	Predictor	Reliability	MSE	Bias	Variance
Register	r	.47	.54	-.13	.52
Survey	s	.69	.23	-.08	.22
Weighted (unconditional), unbiased	$\hat{\xi}_U$	.53	.43	-.12	.41
Weighted (unconditional)	$\hat{\xi}$	.53	.43	-.12	.41
Weighted (conditional), unbiased	$\hat{\xi}_U^*$	.97	.01	.00	.01
Two-stage, unbiased	$\hat{\xi}_U^{(2)}$	.97	.02	.00	.02
Two-stage	$\hat{\xi}^{(2)}$	.97	.01	.00	.01
System-wide linear	$\tilde{\xi}_L$	.76	.12	.00	.12
System-wide	$\tilde{\xi} = \hat{\xi}_S^*$	.98	.01	.00	.01