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

DISCUSSION PAPER SERIES

IZA DP No. 14172

Measurement Error in Earnings Data: Replication of Meijer, Rohwedder, and Wansbeek's Mixture Model Approach to Combining Survey and Register Data

Stephen P. Jenkins
Fernando Rios-Avila

MARCH 2021

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 14172

Measurement Error in Earnings Data: Replication of Meijer, Rohwedder, and Wansbeek's Mixture Model Approach to Combining Survey and Register Data

Stephen P. Jenkins
LSE and IZA

Fernando Rios-Avila
The Levy Institute

MARCH 2021

ISSN: 2365-9793

IZA – Institute of Labor Economics

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

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

www.iza.org

ABSTRACT

Measurement Error in Earnings Data: Replication of Meijer, Rohwedder, and Wansbeek's Mixture Model Approach to Combining Survey and Register Data*

Meijer, Rohwedder, and Wansbeek (MRW, *Journal of Business & Economic Statistics*, 2012) develop methods for prediction of a single earnings figure per worker from mixture factor models fitted using earnings data from multiple linked data sources. MRW apply their method using parameter estimates of Kapteyn and Ypma's mixture factor model (KY, *Journal of Labour Economics* 2007) fitted to earnings data for Swedish workers aged 50+. First, we replicate MRW's empirical analysis using the Swedish model estimates. Second, we check the generality of their empirical finding with a new application. Using estimates of a KY model fit to a linked dataset on earnings for UK employees of all ages, we confirm that MRW's principal findings about the performance of their various predictors of true earnings also hold in this different setting.

JEL Classification: C81, C83, D31

Keywords: earnings prediction, measurement error, mixture factor model, labour earnings, Kapteyn-Ypma model

Corresponding author:
Stephen P. Jenkins
Department of Social Policy
London School of Economics and Political Science
Houghton Street
London WC2A 2AE
United Kingdom
E-mail: s.jenkins@lse.ac.uk

* We are grateful to Associate Editor Heather Anderson and three anonymous referees for their comments, to Erik Meijer for helpfully responding to our queries commenting on a first draft, and for sharing code, and to Tom Wansbeek for his encouragement. Thanks also to the FRS team at the UK Department of Work and Pensions (DWP) for facilitating this project and for helpfully responding to queries. The Linked FRS-P14 Dataset is available from the DWP through a special secure-data contract. Stata code is available from the authors.

1. INTRODUCTION

When you have observations for individuals' earnings from both survey and administrative data, which source will provide the best prediction of true earnings? The answer is straightforward if the linked administrative data are error-free. If not, combining the information from the two sources makes intuitive sense but leaves open the question of how to do this. Meijer, Rohwedder, and Wansbeek (2012, 'MRW') provide answers. They develop methods for prediction of a single earnings measure using both data sources and compare the reliability and mean squared error (MSE) of their various predictors, considering the case in which the parameters describing the generation of the earnings and measurement error processes are given. MRW provide a detailed illustration of their prediction methods using parameter estimates reported by Kapteyn and Ypma (2007, 'KY'). We provide narrow and wide replications of MRW's empirical analysis.

First, we replicate with one minor exception MRW's empirical application. Although MRW's theory of factor score prediction methods for mixture factor analysis models is general, their illustration focuses on the specific case of KY's mixture factor model and uses KY's parameter estimates derived from survey and register data on earnings for 400 Swedish workers aged 50+ in 2003. KY's paper is important because it was the first to model error in administrative data (due to mismatch), in addition to two sources of errors in survey data. They found regression to the mean in survey error to be negligible, thereby overturning a conventional wisdom.

Second, we investigate the generality of MRW's empirical findings. We apply their prediction methods using estimates from KY models fitted to data for nearly 6,000 UK workers from across the full age range. Like MRW, we find that conditionally-weighted and two-stage predictors which combine information from both sources have the best statistical properties.

2. THE KAPTEYN-YPMA (KY) MODEL AND MULTIPLE PREDICTORS OF TRUE EARNINGS

2.1 The Kapteyn-Ypma model for linked register-survey data

Suppose there are two sources of information on log earnings for each of a set of individuals: the survey measure (denoted *s*) and the register (administrative) data measure (*r*). We are interested in predicting true log earnings, ξ, which are unobserved.

KY assume that when a survey response is linked to the wrong person in the register data ('mismatch'), error-ridden measure ζ is observed. Thus, the register data are a mixture of two types of observations, cases *R*1 and *R*2:

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

For the survey data, KY assume there are three types of observation depending on whether earnings are error-free (case *S*1), have measurement error with a regression-to-the-mean component (*S*2), or measurement error plus contamination (*S*3):

$$s = \begin{cases} \xi, & \text{with probability } \pi_s & (\text{case } S1) \\ \xi + \rho(\xi - \mu_\xi) + \eta, & \text{with probability } (1-\pi_s)(1-\pi_\omega) & (\text{case } S2) \\ \xi + \rho(\xi - \mu_\xi) + \eta + \omega, & \text{with probability } (1-\pi_s)\pi_\omega & (\text{case } S3). \end{cases} \quad (2)$$

Each earnings observation pair $\boldsymbol{y} = (r, s)'$ belongs to one of six latent classes characterized by whether cases *R*1 and *R*2 are combined with *S*1, *S*2, or *S*3. The class membership probabilities are π_j, $j = 1, \ldots, 6$, and are functions of π_r, π_s, and π_ω (see MRW's Table 2). For example, the probability of being in class 4, i.e. having administrative mismatch (*R*2) and error-free survey earnings (*S*1), is $(1-\pi_r)\pi_s$.

KY assume that true earnings and error components are each independently and identically normally distributed: $\xi \sim N(\mu_\xi,\ \sigma_\xi^2)$, $\zeta \sim N(\mu_\zeta,\ \sigma_\zeta^2)$, $\eta \sim N(\mu_\eta,\ \sigma_\eta^2)$, and $\omega \sim N(\mu_\omega,\ \sigma_\omega^2)$. They derive parameter estimates by maximizing the model log-likelihood (KY, 2007: Appendix B), having for identification set the size of the 'completely labeled' group, which comprises observations with error-free earnings (class 1: *R*1–*S*1).

2.2 MRW's predictors of true earnings

MRW provide multiple approaches to the problem of how to combine administrative and survey data on earnings to obtain the best prediction of ξ, which we now summarize. MRW

begin by deriving two predictors for the case of a single latent class. The first, $\hat{\boldsymbol{\xi}}$, is a linear function of $\boldsymbol{y}$ that minimizes the mean squared error (MSE). The second, $\hat{\boldsymbol{\xi}}_{\mathbf{U}}$, is a linear combination that minimizes the MSE conditional on unbiasedness. If class membership were manifest, these predictors could be applied in the multiple class case: for each class j, use class-specific versions of these predictors, $\hat{\boldsymbol{\xi}}_j$ and $\hat{\boldsymbol{\xi}}_{\mathbf{U},j}$ respectively.

However, since class membership is unobserved, MRW consider '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' (2012: 194).

Weighted average predictors are of two types depending on whether unconditional latent class probabilities (the π_j) or class probabilities conditional on $\boldsymbol{y}$, $p_j(\boldsymbol{y})$, are used. *Two-stage* predictors first allocate each individual to the class with the largest probability conditional on $\boldsymbol{y}$. Second, given predicted class $\hat{\jmath}$, true earnings are predicted using a within-class estimator. The idea is that if class membership is predicted accurately, two-stage estimators inherit the desirable properties of the within-class predictor. Because either $\hat{\boldsymbol{\xi}}_j$ or $\hat{\boldsymbol{\xi}}_{\mathbf{U},j}$ could be the within-predictor applied to each of the three approaches, there are six potential predictors. *System-wide* predictors minimize total MSE directly. MRW derive one predictor satisfying linearity, $\tilde{\xi}_{\mathrm{L}}$, and another without restrictions, $\tilde{\xi}$, which is identical to the conditionally-weighted predictor based on $\hat{\boldsymbol{\xi}}_j$.

For each of the seven predictors and also predicted class membership probabilities, MRW derive specific expressions for the case of the KY model. Each expression is a function of the model parameters; hence, given a set of parameter estimates, the predictors can be calculated.

3. REPLICATION OF MRW'S EMPIRICAL RESULTS

MRW use the parameter estimates reported by KY (Table C2), reproduced in the first column of Table 1, to illustrate their prediction methods. (The UK estimates in Table 1 are discussed in Section 4.) Taking the parameters as given, MRW abstract from estimation and identification issues. Although MRW report statistics to two decimal places (d.p.), their calculations are based on parameter estimates reported to three d.p., as is our replication of their work. MRW use a combination of R and Stata. We use Stata for all analysis. The

predictors were straightforward to calculate given MRW's expressions. More challenging to reproduce were their Figures 1 and 2, used to summarize results (see below), as this task requires grid searches to find the points defining region boundaries. Our Stata code makes this task easier for future researchers.

<Table 1 near here>

Figure 1 shows predicted class memberships ($\hat{j}$) based on MRW's two-step approach, for each (r, s) combination excluding the region for class 1 ($r = s$). No observation is predicted to be in class 4. We find minor differences from MRW's Figure 2, attributable to a typographical error in MRW's code which led to an incorrect error variance calculation and thence the boundaries of classes 5 and 6. As a consequence, regions marked as class 5 in MRW's Figure 2 are wider than in our Figure 1. But these differences are small and have no knock-on effects for the subsequent calculations to assess predictor performance. Mismatched observations lie further away from the mean of r (12.17) than do correctly-matched ones. Observations with survey measurement error (class 5) tend to be relatively close to the mean of s (12.20); those also with contamination (class 6) are further away from it.

<Figure 1 near here>

Table 2 shows estimates for the six unconditional class membership probabilities and weights for the linear predictors, replicating MRW's Table 5 entries to 2 d.p. Class 2 is the largest, with probability 69%. When there is mismatch (classes 4–6), the within-class predictors place zero weight on the administrative data but different weights are placed on the survey data depending on the error type. For class 5, the weight of 1.01 makes a small adjustment for regression to mean error.

<Table 2 near here>

The weights for unbiased and unrestricted within-class predictors for class 6 differ markedly. The unrestricted predictor substantially down-weights the survey data (relative to the unbiased predictor): it 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' (MRW: 198). The two unconditional probability-weighted predictors have the same weights, 89% on the administrative data and 11% on the survey data (a ratio of about 8:1). The register data get a high weight because the mismatch probability is only 4%. In contrast, the linear projection predictor minimizing MSE ($\tilde{\xi}_{\mathrm{L}}$) gives a relatively low weight to the register data, 22%, and a relatively high weight to the

survey data (55%), i.e. a ratio of about 0.4:1.

For the conditionally-weighted predictors and two-stage predictors, the weights vary with *r* and *s*. For $\tilde{\xi}\ (=\hat{\xi}^{*}) = \tilde{a}(\boldsymbol{y})r + \tilde{b}(\boldsymbol{y})s + \tilde{c}(\boldsymbol{y})$, where the $\tilde{a}(\boldsymbol{y}), \tilde{b}(\boldsymbol{y})$, and $\tilde{c}(\boldsymbol{y})$ terms are weighted averages of the *a*, *b*, and *c* coefficients, and similarly for $\hat{\tilde{\xi}}^{*}_{\mathrm{U}}$. MRW calculate the relative weight given to *r* compared to *s*, $\tilde{a}(\boldsymbol{y}) \,/\, \left[\tilde{a}(\boldsymbol{y}) + \tilde{b}(\boldsymbol{y})\right] = w_r(\boldsymbol{y})$, say, and plot contour lines for $w_r(\boldsymbol{y})$ for different combinations of *r* and *s*.

Figure 2 showing the results for the unrestricted system-wide estimator, $\tilde{\xi}$, replicates MRW's Figure 3 exactly. Close to the mean of true earnings (12.28), $\tilde{\xi}$ puts almost all the weight on *r* (regardless of *s*) but this weight falls very quickly as we move away from the mean. Thus, the two-stage predictors and system-wide predictor $\tilde{\xi}$ have similar relative weights in practice. The two-stage predictors give zero weight to *r* in the case of mismatch and Figure 1 shows that regions where classes 5 and 6 locate correspond to areas of Figure 2 with low $w_r(\boldsymbol{y})$.

<Figure 2 near here>

MRW compare the overall performance of the seven predictors, and of *r* and *s*, in terms of their MSE, and their reliability defined as the square of correlation between the proxy and ξ. Because closed form expressions do not exist for several predictors, MRW compute performance statistics through simulation, using 10,000 draws based on the assumed model and the parameter estimates shown in Table 1.

Table 3 corresponds to MRW's Table 6, with our calculations based on a simulation strategy replicating theirs. Corresponding table entries are the same except for the four marked with '†' which each differ by one percentage point from MRW's. We attribute the differences to simulation variability and rounding.

Table 3 shows, first, that '*r* is a clear loser [relative to *s*]' (MRW: 199). Its reliability is less than one-half (0.49) whereas *s*'s reliability is 0.69, and *r*'s MSE is more than twice that for *s*. The probability of mismatch is small but has substantial adverse consequences in terms of statistical performance. The two unconditional weighted predictors, $\hat{\xi}_U$ and $\hat{\xi}$, have remarkably low reliability and high MSE, the latter being driven by high variance rather than bias (see the rightmost two columns in Table 3).

The system-wide linear predictor $\tilde{\xi}_{\mathrm{L}}$ performs better (reliability = 0.76), but what stands out is the excellent performance of the remaining predictors, including the two two-stage predictors, which perform about as well as each other. The unrestricted system-wide predictor $\tilde{\xi}\ \left(=\hat{\xi}^{*}\right)$ has the highest reliability (0.98) but this is virtually matched by $\hat{\tilde{\xi}}^{*}_{\mathrm{U}}$ and the

two-stage predictors, each with reliability of 0.97, and almost identical MSE (near zero). The excellent performance of the two-stage predictors reflects the high probability (96%) of correctly predicting class membership, with over half of the remaining 4% due to class 3 cases being misclassified as class 2. This has no effect on the precision of the two-stage predictors because the within-class predictor is *r* for both classes.

<Table 3 near here>

4. THE PERFORMANCE OF MRW'S PREDICTORS IN A DIFFERENT CONTEXT

It is useful to know whether MRW's results about the performance of different predictors hold when the application context differs. To investigate this issue, we use estimates of KY model parameters reported by Jenkins and Rios-Avila (2020) derived from UK data.

The survey data are from the Family Resources Survey (FRS), the UK's main income survey, for financial year 2011/12. Information about gross employment earnings is derived by asking what the amount received is (for up to three jobs), followed by a question about the period to which that amount refers (a week, month, or year, etc.). Responses are converted to annual amounts pro rata. Our measure of *s* is the logarithm of total gross earnings (the sum across all jobs reported). The administrative data are for the FRS respondents in employment who gave their consent for their survey responses to be linked to records held by the tax authorities (Her Majesty's Revenue and Customs, HMRC). These P14 data are compiled from employers' returns on P14 forms to HMRC about wages and salaries paid to employees and taxes and National Insurance contributions withheld. Our measure of *r* is the logarithm of total gross earnings per year (the sum across all earnings spells reported for 2011/12).

The linked dataset contains 5,971 men and women after excluding 420 observations with imputed or otherwise edited FRS earnings values. A scatterplot of *r* and *s* for the FRS-P14 data (Appendix Figure A1) is similar to MRW's Figure 1: that the majority of the observations lie on or close to the 45° ray from the origin with relatively few observations above and below the line. However, we would expect differences in model estimates derived the two data sets, for example, differences in earnings and error variances because our data refer to UK employees of all ages whereas KY's sample were all aged 50+. The greater fraction of part-time (low earning) workers in the Swedish data likely accounts for the larger dispersion of both *r* and *s* than in the UK (Appendix Table A1). Also, a larger fraction of the Swedish sample have $r > s$ than $s < r$, whereas the fractions are approximately equal in the

UK sample (Appendix Figure A1).

Table 1 displays KY model estimates from the UK data derived assuming the fraction with error-free earnings is 13.9% (i.e. close to KY's choice). We define completely labeled observations (class 1) as those for whom $|r–s| < 0.02$. The Appendix reports UK estimates derived using a completely-labeled fraction of 3.4% (those for whom $|r–s| < 0.005$). The Appendix also shows that the two sets of estimates lead to similar conclusions about predictor performance and so are not discussed further here.[1]

For the UK and Swedish estimates, the means of measurement error and of contamination error are negative and regression to the mean in survey error is near zero. The standard deviation of earnings among mismatched observations is larger than the variance of true earnings in both estimate sets, but the differential is greater in KY's Swedish estimates (σ_ζ/σ_ξ is 2.5 by contrast with 1.7 for the UK data), which may reflect the differences in survey sample age range. In both estimate sets, contamination error is more widely dispersed than measurement error is but, again, the differential is greater for the Swedish estimates than UK ones: $\sigma_\omega/\sigma_\eta$ is greater than 12 and less than 6 respectively. The UK probability of mismatch is 6%, greater than 4% for the Swedish, consistent with Sweden's greater experience with record linkage.

Figure 3 shows the predicted class membership $\hat{\jmath}$ associated with each combination of r and s for models (a) and (b). This is similar to Figure 1. In both figures, class 2 observations are close to the 45° line. Class 3 observations are close to the mean of r but associated with a wide range of s values. Mismatched observations (classes 5 and 6) have r values distant from the mean.

There are interesting differences as well. In Figure 3, the lines demarcating the boundaries between the regions where class 3 is located on the one hand and classes 5 and 6 on the other hand have a distinctly smaller positive slope than do the corresponding lines in Figure 1. In the UK data, the probability of mismatch is larger than in the Swedish case and so, for observations around the mean of r (12.17), a relatively low – or a relatively high – value of s is more likely to imply membership of classes 5 or 6 rather than membership of class 3 by comparison with the Swedish case.

<Figure 3 near here>

Table 4 shows the UK unconditional class membership probabilities (π_j) and the

[1] How model parameter estimates differ for different values of the completely-labeled fraction is the focus of Jenkins and Rios-Avila's (2020) note.

weights for the linear predictors of true earnings. The probabilities are broadly similar to the Swedish ones with the main differences being the lower probability of class 2 membership (0.61 versus 0.69) and a higher probability of class 3 membership (0.19 versus 0.13). There is a slightly higher probability of class 5 membership as well. Given the similarities in estimates cited so far, it is unsurprising the weights for the linear predictors are similar to those for the Swedish case (cf. Table 3).

The largest cross-national difference is for the unrestricted linear predictor ($\tilde{\xi}_{\mathrm{L}}$), for which the UK case gives a larger relative weight to the register data than does the Swedish one. The UK weight on *r* is 0.32 by contrast with 0.22, whereas the weights on *s* are almost the same (0.54 and 0.55 respectively). Also, the UK within-class predictor weight on survey data for class 6 (mismatch with survey error and contamination) is larger relative to the constant (*c*). In both contexts, there is substantial shrinkage to the population mean because the contamination error variance is large but, in the UK case, the variance is smaller relative to measurement error variance and so less shrinkage.

<Table 4 near here>

Figure 4 shows contour lines for the register data relative weights, $w_r(\boldsymbol{y})$, for the case of the unrestricted system-wide estimator, $\tilde{\xi}$. The UK chart is similar to Figure 2 for the Swedish case: relative weights fall very quickly as we move away from the mean. What differs across settings is the slope of the contour lines, being less steep for the UK case. The explanation is the higher prevalence of mismatch in the UK.

<Figure 4 near here>

Table 5 summarizes the performance of MRW's seven predictors for the UK case. The patterns seen for the Swedish case (Table 3) are seen here as well. For example, first, the survey data are more reliable and have lower MSE than the register data.[2] There is an interesting difference, however. UK reliabilities for both *r* and *s* are higher than their Swedish counterparts and, moreover, the shortfall in reliability of *r* relative to *s* is smaller in the UK case. Put differently, *r* is no longer such a 'clear loser' (MRW: 199) in the UK context by comparison with the Swedish one (reliability of 0.70 compared with 0.48). In current research using the same UK dataset, we find error variances are higher for older than younger

[2] Unfortunately, the reliabilities for *r* and *s* reported by Jenkins and Rios-Avila (2020: Table 3) are incorrect. Corrected estimates show that reliability for *r* is 0.69 over the full range of completely labeled fractions we considered, from 0.25% to 16.93%. Over the same range, reliability for *s* varies only from 0.84 to 0.80, respectively. (These estimates are derived using analytical formulae; the estimates shown in Table 6 are derived via simulation.) In sum, our conclusions about the relative reliabilities of *r* and *s*, and how they vary with the completely labelled fraction, are unaffected.

workers, suggesting that the UK-Sweden differences in reliability are related to the different age compositions of the samples.

<Table 5 near here>

Second, the weighted unconditional estimators also perform better in the UK context than the Swedish one, with a reliability of 0.75 rather than 0.54. So too does the system-wide linear predictor, with a reliability of 0.87 rather than 0.76.

Third, the remaining predictors have excellent statistical performance, with reliabilities of 0.95 or 0.96, which are only very slightly smaller than their Swedish data model counterparts (0.97 or 0.98). The excellent performance of the two-stage predictors is again due to the high accuracy with which class membership is predicted by the models, 87%, though this rate is lower than the Swedish one (96%). Contaminated observations prove harder to identify in the UK context. However, in both settings, at least one half of the observations misclassified are class 3 individuals wrongly placed in class 2.

5. SUMMARY AND CONCLUSIONS

MRW analyse approaches to predicting individuals' true earnings when you have earnings observations from linked survey and administrative data. They provide general methods for factor score prediction in mixture factor analysis models and illustrate them using KY's model of the relationships between survey, register, and true earnings, and taking KY's parameter estimates referring to Swedish data for older workers.

We have replicated MRW's empirical findings about the statistical performance of the various candidate earning predictors based on estimates derived from models fitted to data for Swedish workers aged 50+. We have also shown that their findings about predictor performance also hold in a different setting, UK employees of all ages, using Jenkins and Rios-Avila's (2020) KY model parameter estimates. The findings in common are that there are two-stage predictors and conditionally-weighted predictors of true earnings that have high reliability and low MSE. These perform much better in MSE and reliability terms than do the register or survey data used separately.

These findings about earnings predictors are conditional on the KY model used to describe the data generating process. The KY model does not allow for measurement error per se in the register data, ignores heterogeneity across subgroups of workers, and assumes each factor is normally distributed. In current work we are fitting models that relax the first two assumptions; deriving and assessing the performance of earnings predictors for these and

other models raises interesting challenges for future research.

REFERENCES

Jenkins, S. P. and Rios-Avila, F. (2020). Measurement errors in survey and administrative data on employment earnings: sensitivity to the fraction assumed to have error-free earnings, Economics Letters, online first at https://doi.org/10.1016/j.econlet.2020.109253

Kapteyn, A. and Ypma, J. Y. (2007). Measurement error and misclassification: a comparison of survey and administrative data. Journal of Labor Economics, 25(3), 513–551.

Meijer, E., Rohwedder, S., and Wansbeek, T. (2012). Measurement error in earnings data: using a mixture model approach to combine survey and register data. Journal of Business and Economic Statistics, 30(2), 191–201.

Table 1. Estimates of KY model derived from Swedish and UK data

Parameter	KY (2007) (Sweden, workers aged 50+)	Jenkins & Rios-Avila (2020) (UK, workers of all ages)
μ_ξ	12.283	9.8121
σ_ξ	0.717	0.7522
μ_ζ	9.187	8.5303
σ_ζ	1.807	1.2596
μ_ω	–0.304	–0.1499
σ_ω	1.239	0.7309
μ_η	–0.048	–0.0129
σ_η	0.099	0.1413
π_r	0.959	0.9393
π_s	0.152	0.1469
π_ω	0.156	0.2373
ρ	–0.013	–0.0279

Note See main text for sources.

Table 2. Unconditional class membership probabilities (π_j) for Swedish earnings data model and predictors of linear expressions *ar* + *bs* + *c*: replication of MRW's Table 5

				$\hat{\xi}_{U,j}$			$\hat{\xi}_j$		
Class (*j*)	*r*	*s*	π_j	*a*	*b*	*c*	*a*	*b*	*c*
1	R1	S1	0.15	0.5	0.5	0	0.5	0.5	0
2	R1	S2	0.69	1	0	0	1	0	0
3	R1	S3	0.13	1	0	0	1	0	0
4	R2	S1	0.01	0	1	0	0	1	0
5	R2	S2	0.03	0	1.01	–0.11	0	0.99	0.12
6	R2	S3	0.01	0	1.01	0.19	0	0.25	9.32
			$\hat{\xi}_U$:	0.89	0.11	–0.00	$\hat{\xi}$: 0.89	0.11	0.05
			$\tilde{\xi}_L$:	0.22	0.55	2.85			

Note The π_j do not sum to 1 because of rounding.

Table 3. Precision of the predictors: replication of MRW's Table 6

	Predictor	Reliability	MSE	Bias	Variance
Administrative data	r	0.48†	0.54	–0.12†	0.52
Survey	s	0.69	0.23	–0.08	0.23†
Weighted (unconditional), unbiased	$\hat{\xi}_{\mathrm{U}}$	0.54	0.43	–0.12	0.41
Weighted (unconditional)	$\hat{\xi}$	0.54	0.43	–0.12	0.41
Weighted (conditional), unbiased	$\hat{\xi}^{*}_{\mathrm{U}}$	0.97	0.01	0.00	0.01
Two-stage, unbiased	$\hat{\xi}^{(2)}_{\mathrm{U}}$	0.97	0.02	–0.00	0.02
Two-stage	$\hat{\xi}^{(2)}$	0.97	0.01	–0.00	0.01
System-wide linear	$\tilde{\xi}_{\mathrm{L}}$	0.76	0.13†	–0.00	0.13†
System-wide	$\tilde{\xi} = \hat{\xi}^{*}$	0.98	0.01	0.00	0.01

Note Cells marked with '†' differ from the corresponding MRW estimates by one percentage point.

Table 4. Unconditional class membership probabilities (π_j) for UK earnings data model and predictors of linear expressions *ar* + *bs* + *c*

				$\hat{\xi}_{\mathrm{U},j}$			$\hat{\xi}_j$		
Class (*j*)	*r*	*s*	π_j	*a*	*b*	*c*	*a*	*b*	*c*
1	R1	S1	0.14	0.5	0.5	0	0.5	0.5	0
2	R1	S2	0.61	1	0	0	1	0	0
3	R1	S3	0.19	1	0	0	1	0	0
4	R2	S1	0.01	0	1	0	0	1	0
5	R2	S2	0.04	0	1.03	–0.27	0	0.99	0.09
6	R2	S3	0.01	0	1.03	–0.11	0	0.51	4.94
			$\hat{\xi}_{\mathrm{U}}$:	0.87	0.13	–0.01	$\hat{\xi}$: 0.87	0.12	0.08
			$\tilde{\xi}_{\mathrm{L}}$:	0.32	0.54	1.04			

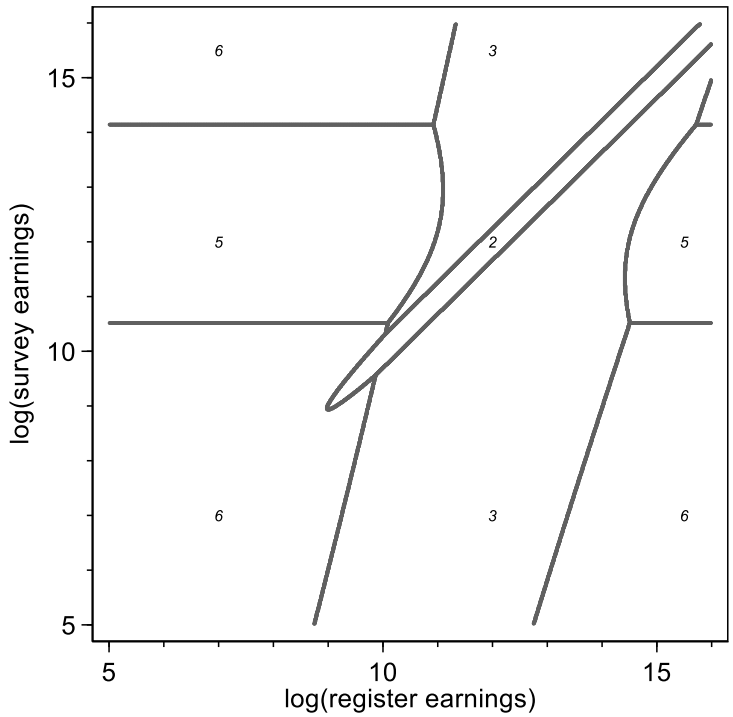
Notes Authors' calculations from KY model estimates reported by Jenkins and Rios-Avila (2020) for completely labelled fraction = 13.9%. The π_j do not sum to 1 because of rounding.

Table 5. Precision of the predictors: UK earnings data model

Label	Predictor	Reliability	MSE	Bias	Variance
Administrative data	r	0.70	0.23	–0.08	0.22
Survey	s	0.81	0.13	–0.04	0.13
Weighted (unconditional), unbiased	$\hat{\xi}_{\mathrm{U}}$	0.75	0.17	–0.07	0.17
Weighted (unconditional)	$\hat{\xi}$	0.75	0.18	–0.07	0.17
Weighted (conditional), unbiased	$\hat{\xi}^{*}_{\mathrm{U}}$	0.96	0.02	0.00	0.02
Two-stage, unbiased	$\hat{\xi}^{(2)}_{\mathrm{U}}$	0.95	0.03	0.00	0.03
Two-stage	$\hat{\xi}^{(2)}$	0.95	0.03	0.00	0.03
System-wide linear	$\tilde{\xi}_{\mathrm{L}}$	0.87	0.08	0.00	0.08
System-wide	$\tilde{\xi} = \hat{\xi}^{*}$	0.96	0.02	0.00	0.02

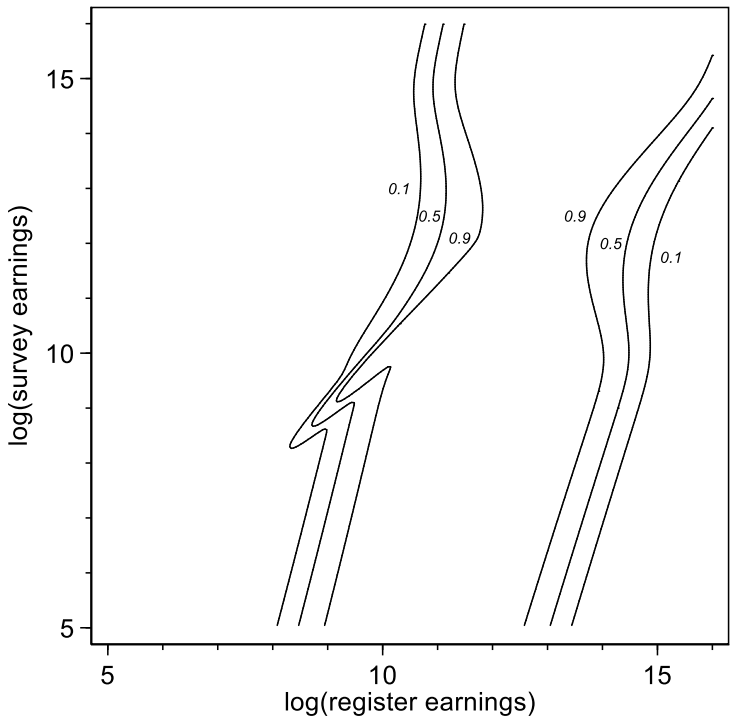
Notes Authors' calculations from KY model estimates reported by Jenkins and Rios-Avila (2020), for completely labelled fraction = 13.9%.

Figure 1. Predicted class membership $\hat{\jmath}$ for each combination of register and survey data (excluding observations with $r = s$): replication of MRW's Figure 2



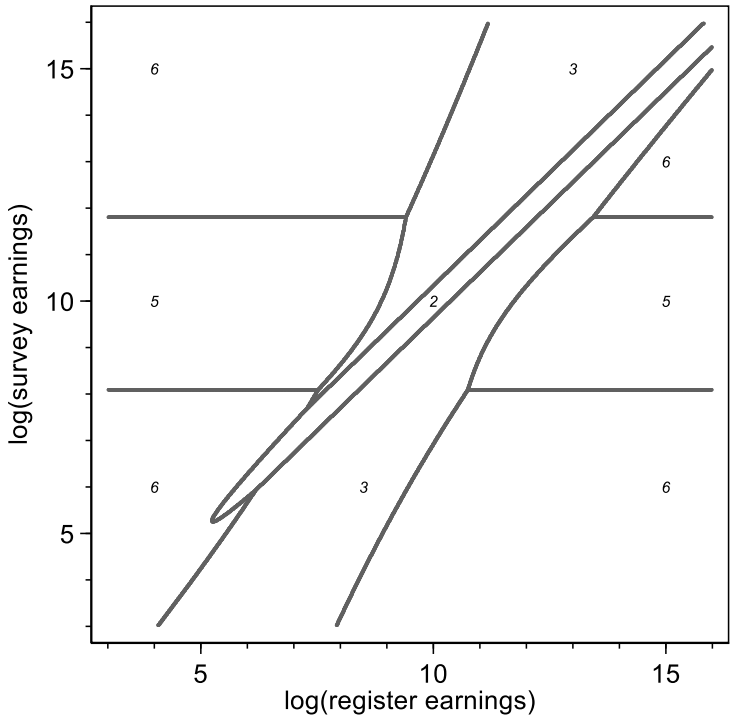
Note This figure differs slightly from MRW's: see main text.

Figure 2. Relative weight of *r* in the conditionally weighted predictor $\tilde{\xi}$: replication of MRW's Figure 3



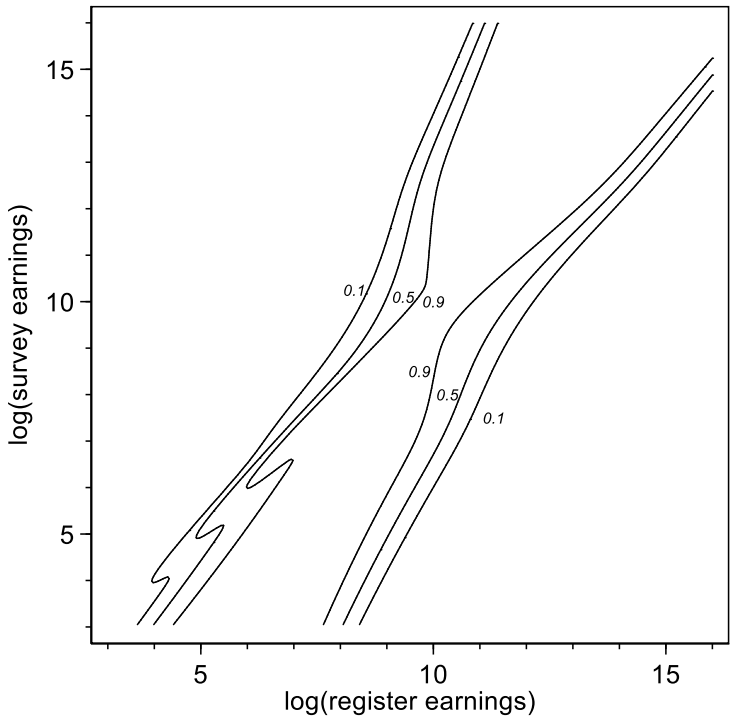
Notes The lines connect points with the same relative weight given to *r*, $w(\boldsymbol{y}) = \tilde{a}(\boldsymbol{y}) \,/\, \left[\tilde{a}(\boldsymbol{y}) + \tilde{b}(\boldsymbol{y})\right]$.

Figure 3. Predicted class membership $\hat{j}$ for each combination of register and survey data (excluding observations with $r = s$): KY model applied to UK earnings data



Note The completely labeled fraction = 13.9%.

Figure 4. Relative weight of *r* in the conditionally weighted predictor $\tilde{\xi}$: KY model applied to UK earnings data



Note The lines connect points with the same relative weight given to *r*, $w(\boldsymbol{y}) = \tilde{a}(\boldsymbol{y}) \,/\, \left[\tilde{a}(\boldsymbol{y}) + \tilde{b}(\boldsymbol{y})\right]$.

APPENDIX

(online Supplementary Material)

for

Measurement error in earnings data: replication of Meijer, Rohwedder, and Wansbeek's mixture model approach to combining survey and register data

Stephen P. Jenkins
(LSE)

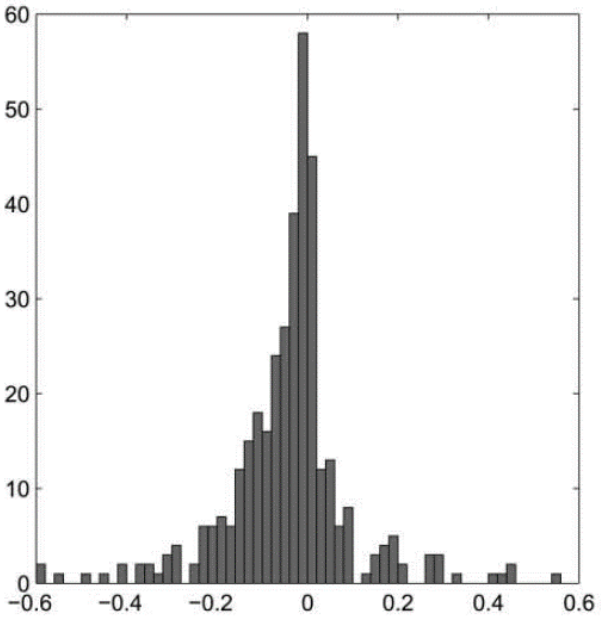
Fernando Rios-Avila
(The Levy Institute)

Table A1. Summary statistics, register (*r*) and survey (*s*) measures of log earnings, by dataset

		Kapteyn & Ypma (2007) (Sweden, workers aged 50+)	Jenkins & Rios-Avila (2020) (UK, workers of all ages)
N		400	5,971
s	Mean	12.196	9.771
	(std. dev.)	(0.821)	(0.813)
r	Mean	12.172	9.754
	(std. dev.)	(0.961)	(0.842)
s–*r*	Mean	0.024	0.016
	(std. dev.)	(0.656)	(0.496)

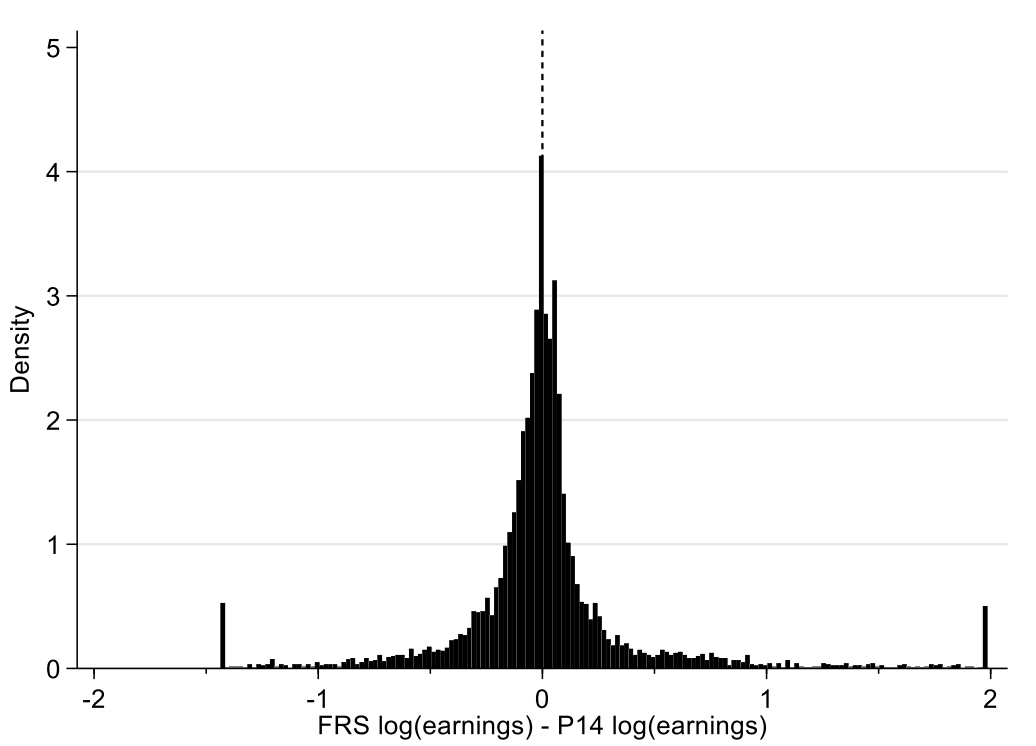
Source: Kapteyn and Ypma (2007, Table 4) and authors' calculations.

Figure A1. Histograms of difference between survey and register earnings (*s* – *r*)
(a) Swedish data



Source: Kapteyn and Ypma (2007, Figure 4a). Values were truncated at ± 0.6. Histogram bin-width not reported.

(b) UK data



Notes. Histogram (bin width 0.02). Values are bottom-coded at *p*1 (–1.44) and top-coded at *p*99 (1.97) for purposes of presentation.

Table A2. Estimates of KY model: UK dataset with completely labelled fraction = 3.4%

Parameter	Jenkins & Rios-Avila (UK)
μ_ξ	9.8105
σ_ξ	0.7565
μ_ζ	8.6211
σ_ζ	1.2881
μ_ω	–0.1239
σ_ω	0.6369
μ_η	–0.0091
σ_η	0.1142
π_r	0.9362
π_s	0.0365
π_ω	0.2606
ρ	–0.0192

Notes The UK estimates are for workers from across the full age range in 2011/12, from Jenkins and Rios-Avila (2020, Table 1), but use a completely labeled fraction of 3.4%, rather than 13.9% as in Table 1.

Table A3. Unconditional class membership probabilities (π_j) for UK earnings data model with completely labeled fraction = 3.4% and predictors of linear expressions *ar* + *bs* + *c*

				$\hat{\xi}_{\mathrm{U},j}$			$\hat{\xi}_j$		
Class (*j*)	*r*	*s*	π_j	*a*	*b*	*c*	*a*	*b*	*c*
1	R1	S1	0.03	0.5	0.5	0	0.5	0.5	0
2	R1	S2	0.67	1	0	0	1	0	0
3	R1	S3	0.24	1	0	0	1	0	0
4	R2	S1	0.00	0	1	0	0	1	0
5	R2	S2	0.05	0	1.02	–0.18	0	1.00	0.05
6	R2	S3	0.02	0	1.02	–0.06	0	0.58	4.21
				$\hat{\xi}_{\mathrm{U}}$:	0.92	0.08	–0.01	$\hat{\xi}$: 0.92	0.07	0.07
				$\tilde{\xi}_{\mathrm{L}}$:	0.30	0.60	0.99			

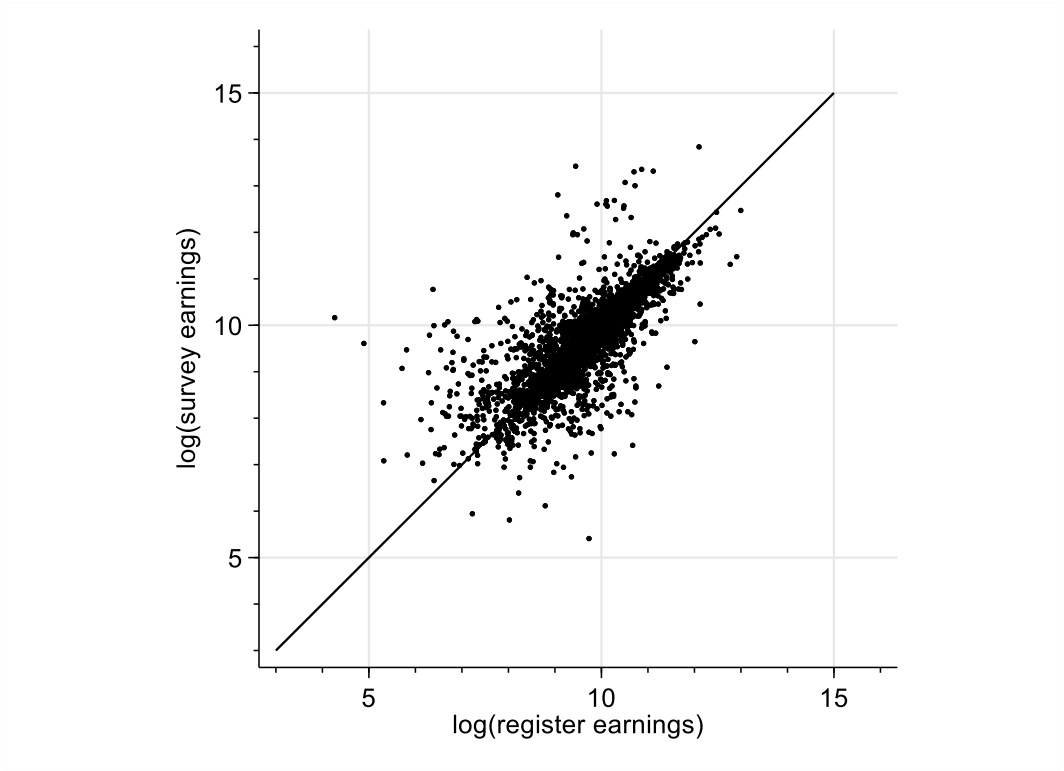
Notes Authors' calculations from KY model estimates reported by Jenkins and Rios-Avila (2020), but using a completely labeled fraction of 3.4%, rather than 13.9% as in Table 4. The π_j do not sum to 1 because of rounding.

Table A4. Precision of the predictors: UK earnings data model with completely labeled fraction = 3.9%

Label	Predictor	Reliability	MSE	Bias	Variance
Administrative data	r	0.68	0.24	–0.08	0.23
Survey	s	0.82	0.12	–0.04	0.12
Weighted (unconditional), unbiased	$\hat{\xi}_{\mathrm{U}}$	0.72	0.20	–0.07	0.20
Weighted (unconditional)	$\hat{\xi}$	0.72	0.20	–0.07	0.20
Weighted (conditional), unbiased	$\hat{\xi}^{*}_{\mathrm{U}}$	0.96	0.02	0.00	0.02
Two-stage, unbiased	$\hat{\xi}^{(2)}_{\mathrm{U}}$	0.95	0.03	0.00	0.03
Two-stage	$\hat{\xi}^{(2)}$	0.95	0.03	0.00	0.03
System-wide linear	$\tilde{\xi}_{\mathrm{L}}$	0.87	0.07	0.00	0.07
System-wide	$\tilde{\xi} = \hat{\xi}^{*}$	0.96	0.02	0.00	0.02

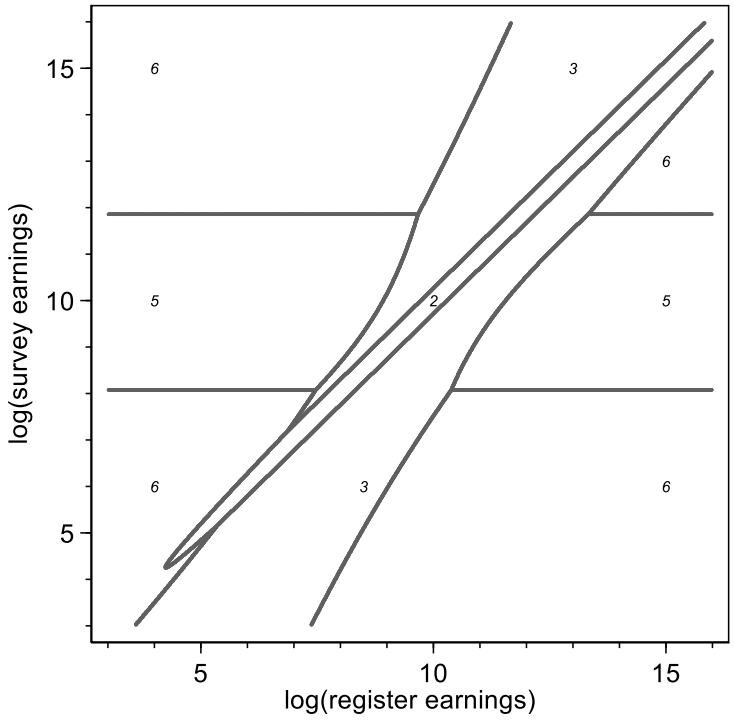
Notes Authors' calculations from KY model estimates reported by Jenkins and Rios-Avila (2020), but using a completely labeled fraction of 3.4%, rather than 13.9% as in Table 5.

Appendix Figure A2. Scatter plot of *r* and *s*: UK FRS-P14 Linked Dataset



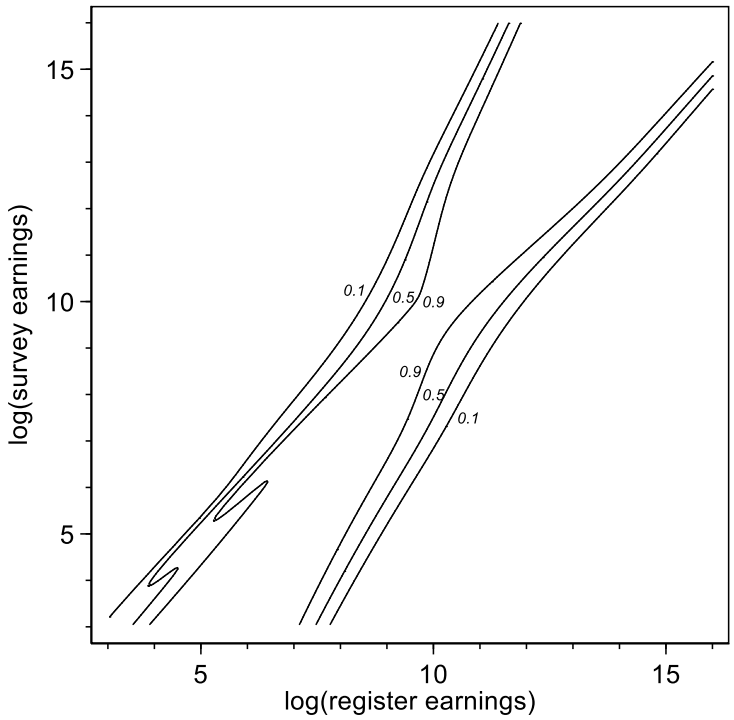
Notes Mean of r = 9.75; mean of s = 9.77.

Appendix Figure A3. Predicted class membership $\hat{j}$ for each combination of register and survey data (excluding observations with $r = s$): KY model applied to UK earnings data



Note Completely labeled fraction = 3.4% (cf. Figure 5, 13.9%).

Appendix Figure A4. Relative weight of *r* in the conditionally weighted predictor $\tilde{\xi}$: KY model applied to UK earnings data



Notes The lines connect points with the same relative weight given to *r*, calculated as $\tilde{a}(\boldsymbol{y}) \,/\, \left[\tilde{a}(\boldsymbol{y}) + \tilde{b}(\boldsymbol{y})\right]$. Completely labeled fraction = 3.4% (cf. Figure 5, 13.9%)