Modelling errors in survey and administrative data on employment earnings: sensitivity to the fraction assumed to have error-free earnings

Stephen P. Jenkins
(LSE)

Fernando Rios-Avila
(The Levy Institute)

20 May 2020

Abstract

Kapteyn and Ypma (Journal of Labour Economics 2007) is an influential study of errors in survey and administrative data on employment earnings. To fit their mixture models, Kapteyn and Ypma assume a specific fraction of their sample have error-free earnings. Using a new UK dataset, we assess the sensitivity of model estimates and post-estimation statistics to variations in this fraction and find some lack of robustness.

Keywords: measurement error, misclassification error, labour earnings, Kapteyn-Ypma model

JEL classification codes: C81, C83, D31

Acknowledgements: Thanks to the FRS team at the UK Department of Work and Pensions 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 for data management and estimation is available from the authors.

Correspondence

Jenkins: Department of Social Policy, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK. Email: s.jenkins@lse.ac.uk

Rios-Avila: The Levy Economics Institute of Bard College, Blithewood, Annandale-on-Hudson NY 12504-5000, USA. Email: friosavi@levy.org

1. Introduction

Kapteyn and Ypma's (2007) study of error in survey and administrative data on employment earnings was pioneering because its modelling did not assume that administrative data represent the truth.[1] Using a sample of Swedish individuals aged 50+ years in the early 2000s, Kapteyn and Ypma (KY hereafter) concluded that simple econometric models had 'potentially very substantial biases' (2007: 513) and the 'common finding of substantial mean reversion in survey data largely goes away once [they] allow for a richer error structure' (2007: 513). Meijer et al. (2013), using KY's data and model estimates, showed that KY's survey data on earnings were more reliable than the linked administrative data. In this paper, using a new UK dataset, we assess the sensitivity of these conclusions to KY's assumption that the fraction of observations with error-free earnings was 14.8%.

KY's Full Model is a bivariate mixture model of six distributions. The first distribution is for the group containing the individuals for whom survey and administrative earnings data are correct, i.e. 'completely labelled'. KY assume that observations are completely labelled if the difference between their survey and administrative earnings are less than 1,000 SEK per year, thereby identifying 14.8% of their sample (2007: 524). This fraction seems large to us because, with all earnings prone to error, we would expect a smaller fraction than this to be truly error-free. KY state that '[i]n principle, this broader definition of "equal" observations affects the consistency of our estimates. However, we expect these effects to be minor' (KY: 540). They do not discuss the issue further.

Intuitively, the larger the fraction of observations that is completely labelled, the smaller is the proportion of total variation in survey or administrative earnings that a model will attribute to error, and there will also be implications for other aspects such as the relative importance of different kinds of errors and the reliabilities of survey and administrative earnings. Here we analyze in detail whether effects are 'minor' by fitting KY Full Models using completely labelled sample fractions ranging from 0.25% to 17%.

We confirm KY's conjecture that effects are minor if one focuses on error distribution means, the extent of regression to the mean, or the relative reliabilities of survey and

[1] Recent papers also assuming that administrative data do not represent the truth are Abowd and Stinson (2013), Bollinger et al. (2018), and Hyslop and Townsend (2020).

administrative earnings data, but there are more substantive effects on error variances and mixture group (latent class) membership probabilities. Thus, conclusions about whether the effects of changing the completed labelled fraction are 'minor' depend on which measurement feature is focused on.

2. The Linked FRS-P14 Dataset

This section introduces the UK Linked FRS-P14 Dataset on gross labour earnings for employees. For further details, see Jenkins and Rios-Avila (2020).

The survey data are from the Family Resources Survey (FRS) for financial year 2011/12 (Department for Work and Pensions 2013). The FRS is the UK's main income survey with an annual sample of around 20,000 private households, and the source for the DWP's annual report on low income prevalence, *Households Below Average Income*, and other oft-cited statistics about the UK income distribution.

FRS information about gross earnings from employment is derived by asking what the last amount received was, followed by a question about the period to which that amount refers. The data producers convert responses to weekly GBP amounts pro rata – the originally-reported amounts are not released – which we have then converted to annual amounts. Respondents are asked about earnings for up to three jobs, but less than 5% of our sample report more than one. Our survey measure of earnings for each linked respondent i, s_i, 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 tax authorities (Her Majesty's Revenue and Customs, HMRC) and the DWP. We label this the P14 dataset because it is compiled from employers' returns on P14 forms to HMRC about wages and salaries paid to employees and taxes and National Insurance contributions withheld. (Around 60% of FRS respondents provided consent to data linkage.) DWP statisticians linked the FRS and P14 data deterministically using match keys constructed from information about first name, last name, postcode, sex, and date of birth. Our administrative measure of earnings for each linked respondent i, r_i, is the logarithm of total gross earnings per year (the sum across all spells

reported).[2]

The earnings measures s_i and r_i may differ for several reasons. The FRS earnings reference period is not the same as the P14 one; differences in s_i and r_i can arise through job or pay instability over the year. In the survey there are potential measurement errors in respondents' answers, interviewers' recording of them, or in subsequent survey processing. It is possible that there is incorrect linkage ('mismatch') between survey respondents and P14 records. All these features are incorporated in KY's Full Model.[3]

Our linked data set contains observations on r_i and s_i for 6,391 men and women. We follow common practice and drop observations with imputed or otherwise edited FRS earnings values, leaving an estimation sample of 5,971 individuals, substantially larger than KY's earnings sample with $N = 400$ (2007: Table 1). The means of r_i and s_i are 9.75 and 9.77 with standard deviations 0.842 and 0.813, respectively. The distribution of differences ($s_i - r_i$) has a mean of –0.02 with standard deviation 0.496.

3. The Kapteyn-Ypma Full Model

We restrict attention to KY's Full Model and, following KY and Meijer et al. (2013), we focus on models without covariates.[4]

The distribution of observed administrative earnings is a mixture of observations correctly matched with an FRS respondent and incorrectly matched observations. In the first case ($R1$), r_i equals i's true earnings ξ_i with probability π_r; in the second case ($R2$), r_i is the earnings of someone else in the full P14 dataset with probability $(1-\pi_r)$:

$$r_i = \begin{cases} \xi_i & \text{with probability } \pi_r \\ \zeta_i & \text{with probability } (1-\pi_r). \end{cases} \quad (1)$$

Survey observations are of three types. In case $S1$, s_i equals true earnings with probability

[2] P14 earnings spells cannot be linked to jobs in the survey and no individual characteristics besides sex are recorded.
[3] The administrative data may miss some earnings if some employment spells are not captured on P14 forms and there might be errors in employers' submissions. We are developing extensions to the KY Full Model to incorporate these kinds of error.
[4] We also fitted models in which mean true earnings depend on covariates, as KY did. Conclusions are similar to those reported below.

π_s. In case *S*2, s_i contains response error with probability $(1-\pi_s)(1-\pi_\omega)$ and in case *S*3 there are observations that also include contamination error with probability $(1-\pi_s)\pi_\omega$:

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

Thus, each sample observation may belong to one of six latent classes characterized by the combinations of cases *R*1, *R*2 with *S*1, *S*2, or *S*3.

For estimation, KY assume true earnings and errors are each independently and identically normally distributed: $\xi_i \sim N(\mu_\xi,\ \sigma_\xi^2)$, $\zeta_i \sim N(\mu_\zeta,\ \sigma_\zeta^2)$, $\eta_i \sim N(\mu_\eta,\ \sigma_\eta^2)$, and $\omega_i \sim N(\mu_\omega,\ \sigma_\omega^2)$. The mixture log-likelihood expression is given by KY (2007: Appendix B) and maximized contingent on setting the size of the first, i.e. completely labelled, group. Using the Linked FRS-P14 Dataset, we fit the Full Model separately for 8 values of the completely labelled fraction, ranging from values slightly greater than KY's to much smaller ones.

4. How do estimates change as the completely labelled fraction is varied?

We define completely labelled observations as those with $|r_i - s_i| \leq \delta$ with threshold δ taking values in the range [0, 0.025], implying sample fractions between 0.25% and 16.93%. Table 1 reports the 8 sets of estimates. All parameters are precisely estimated ($p < 0.01$).

Varying δ has virtually no impact on estimates of the mean and variance of the distributions of both true earnings (ξ) and mismatched earnings (ζ), or on the means of survey measurement and contamination errors (μ_η, μ_ω). Also, for all δ, estimated mean-reversion in response error ($\hat{\rho}$) is negative but close to zero, and the estimated probability of mismatch ($1 - \widehat{\pi_r}$) is around 7%. The stability of the latter is perhaps because mismatch is the only source of error in administrative earnings in this model.

Increasing the completely labelled fraction increases the probability of survey earnings being error-free by construction ($\widehat{\pi_s}$ ranges between 0.003 and 0.179). This variation is associated not only with a fall in the probability of survey contamination from 0.28 to 0.23 (–17%) but also with a marked increase in measurement and contamination error variances. The measurement error variance increases by 44% as δ is varied between 0 ($\widehat{\sigma_\eta}$ = 0.104) and 16.95%

(0.150), and the contamination error variance $\widehat{\sigma_\omega}$ increases by 32% (0.571 to 0.754).

Let us compare our estimates for the case $\delta = 0.20$ (implying a completely labelled sample fraction of 13.9%) with KY's estimates (2007, Table C2). There are many similarities including the near-zero value for $\hat{\rho}$ and small mismatch rate. However, there are marked differences too, notably in the measurement and contamination error variances: our $\widehat{\sigma_\eta}$ is 0.15 compared with KY's 0.10, and our $\widehat{\sigma_\omega}$ is 0.74 rather than 1.24. In addition, our $\widehat{\pi_\omega}$ is substantially larger than KY's: 0.23 rather than 0.16. These differences may reflect differences in the sample composition (the UK data are not restricted to individuals aged 50+) or survey design.

Table 2 shows estimates of the class membership probabilities implied by the parameters reported in Table 1. Increasing δ raises the fraction in class 1 (completely labelled), of course, and there are corresponding decreases in the proportions predicted to belong to class 2 (R_1, S_2) and class 3 (R_1, S_3). (There are also changes in the probabilities of belonging to classes 4–6, but they are negligible in size.) Over the range of δ, the estimated fraction with error-free administrative earnings and measurement error in survey earnings ($\widehat{\pi_2}$) falls from 0.67 to 0.59 (–11%), whereas the fraction with error-free administrative earnings, survey measurement error, and contamination ($\widehat{\pi_3}$) falls from 0.26 to 0.18 (–31%). For the highest values of δ, ($\widehat{\pi_2} + \widehat{\pi_3}$) is approximately the same as KY's estimate (Meijer et al. 2013: Table 5), but we find relatively more in group 3 reflecting our higher estimate of $\widehat{\pi_\omega}$ (see earlier).

Table 3 displays estimates of the reliabilities of administrative and survey earnings, where reliability is the squared correlation between each measure and true earnings (Meijer et al. 2013). P14 earnings reliability is insensitive to variations in δ, around 0.52. FRS earnings reliability estimates are slightly less robust but their range is small (0.60 to 0.64). The survey data are more reliable than the administrative data, regardless of δ. The root cause is the non-zero chance of FRS-P14 mismatch.

Finally, we repeat KY's analysis of proportionate biases in OLS regression coefficients for bivariate regressions in which true earnings are the dependent variable ('ξ LHS') or the explanatory variable ('ξ RHS') but are replaced by either administrative earnings or survey earnings. Table 4 shows that when ξ is the dependent variable, biases are small – a maximum of 6% for *r* and only 2%–3% for *s*. Biases are markedly greater, however, when ξ is the explanatory variable and, again, bias is less for survey earnings (14%–18%) than for administrative earnings

(around 23% for all δ). Again, the differences can be traced back to the presence of mismatch error. Our ξ RHS bias estimates are similar to KY's (Table 5) but our estimated ξ RHS biases are smaller, especially for administrative data.

In sum, we have demonstrated that the choice of the completely labelled fraction is unimportant for some statistics but has notable impacts on others such as the relative importance of measurement and matching errors, and predicted class membership probabilities. The more general lesson is that researchers need to reflect carefully on the implications of whatever choice(s) they make.

References

Abowd, J. and Stinson, M. (2013). Estimating measurement error in annual job earnings: a comparison of survey and administrative data. Review of Economics and Statistics, 95, 1451–1467.

Bollinger, C. R., Hirsch, B. T., Hokayem, C. M., and Ziliak, J. P. (2018). The good, the bad and the ugly: measurement error, non-response and administrative mismatch in the CPS. Working Paper, Gatton College of Business, University of Kentucky.

Department for Work and Pensions (2013). Family Resources Survey United Kingdom, 2011/12. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/206887/frs_2011_12_report.pdf

Hyslop, D. R. and Townsend, W. (2020). Earnings dynamics and measurement error in matched survey and administrative data. Journal of Business and Economic Statistics, 38, 457–469.

Jenkins, S. P. and Rios-Avila, F. (2020). Measurement error in UK survey and administrative data on gross labour earnings. Unpublished paper, LSE.

Kapteyn, A. and Ypma, J. Y. (2007). Measurement error and misclassification: a comparison of survey and administrative data. Journal of Labor Economics, 25, 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, 191–201.

Table 1. Estimates of Kapteyn-Ypma Full Model, by definition of completely labelled group

Parameter	Definition of completely labelled group: $\lvert r_i - s_i \rvert \leq \delta$, where $\delta = \ldots$							
	0.000 (0.25%)*	0.001 (1.00%)	0.002 (1.74%)	0.005 (3.43%)	0.010 (7.74%)	0.015 (11.14%)	0.020 (13.87%)	0.025 (16.93%)
μ_ξ	9.8095	9.8099	9.8101	9.8105	9.8112	9.8118	9.8121	9.8125
	(0.0102)	(0.0102)	(0.0102)	(0.0102)	(0.0102)	(0.0102)	(0.0101)	(0.0101)
σ_ξ	0.7617	0.7594	0.7582	0.7565	0.7542	0.7530	0.7522	0.7515
	(0.0092)	(0.0082)	(0.0079)	(0.0078)	(0.0076)	(0.0076)	(0.0076)	(0.0075)
μ_ζ	8.7183	8.6774	8.6549	8.6211	8.5716	8.5469	8.5303	8.5147
	(0.1230)	(0.1073)	(0.1049)	(0.1042)	(0.1067)	(0.1093)	(0.1114)	(0.1139)
σ_ζ	1.2990	1.2964	1.2939	1.2881	1.2752	1.2664	1.2596	1.2530
	(0.0539)	(0.0557)	(0.0567)	(0.0582)	(0.0607)	(0.0621)	(0.0632)	(0.0643)
μ_ω	–0.1122	–0.1162	–0.1189	–0.1239	–0.1354	–0.1436	–0.1499	–0.1574
	(0.0192)	(0.0197)	(0.0203)	(0.0215)	(0.0241)	(0.0261)	(0.0277)	(0.0295)
σ_ω	0.5713	0.5968	0.6118	0.6369	0.6814	0.7096	0.7309	0.7542
	(0.0511)	(0.0352)	(0.0313)	(0.0279)	(0.0263)	(0.0267)	(0.0275)	(0.0287)
μ_η	–0.0075	–0.0080	–0.0084	–0.0091	–0.0105	–0.0117	–0.0129	–0.0141
	(0.0021)	(0.0021)	(0.0021)	(0.0022)	(0.0025)	(0.0027)	(0.0029)	(0.0031)
σ_η	0.1036	0.1068	0.1093	0.1142	0.1255	0.1342	0.1413	0.1497
	(0.0043)	(0.0033)	(0.0031)	(0.0029)	(0.0029)	(0.0030)	(0.0032)	(0.0034)
π_r	0.9311	0.9334	0.9346	0.9362	0.9381	0.9389	0.9393	0.9398
	(0.0076)	(0.0062)	(0.0059)	(0.0057)	(0.0056)	(0.0056)	(0.0056)	(0.0057)
π_s	0.0027	0.0107	0.0185	0.0365	0.0821	0.1181	0.1469	0.1793
	(0.0007)	(0.0014)	(0.0018)	(0.0025)	(0.0037)	(0.0043)	(0.0048)	(0.0052)
π_ω	0.2809	0.2729	0.2682	0.2606	0.2483	0.2417	0.2373	0.2328
	(0.0187)	(0.0145)	(0.0136)	(0.0128)	(0.0123)	(0.0124)	(0.0127)	(0.0130)
ρ	–0.0156	–0.0169	–0.0177	–0.0192	–0.0231	–0.0256	–0.0279	–0.0304
	(0.0034)	(0.0032)	(0.0032)	(0.0033)	(0.0036)	(0.0038)	(0.0041)	(0.0044)

Notes. Standard errors in parentheses. Sample N = 5,971. *: fraction of sample satisfying $\lvert r_i - s_i \rvert \leq \delta$ condition, where r is P14 earnings and s is FRS earnings. Parameters μ and σ refer to means and standard deviations respectively. ξ: true earnings. ζ: r if mismatch of FRS case with P14 case. ω: contamination error in s. η: measurement error in s. ρ: regression to the mean in s. π_r: Pr(FRS obs correctly matched with P14 obs). π_s: Pr(s reported correctly). π_ω: Pr(s contains contamination error). Source: authors' estimates from Linked FRS-P14 dataset.

Table 2. Estimates of class probabilities (π_j) from Kapteyn-Ypma Full Model, by definition of completely labelled group

Class, j	r	s	π_j	Definition of completely labelled group: $\lvert r_i - s_i \rvert \leq \delta$, where $\delta = \ldots$							
				0.000 (0.25%)*	0.001 (1.00%)	0.002 (1.74%)	0.005 (3.43%)	0.010 (7.74%)	0.015 (11.14%)	0.020 (13.87%)	0.025 (16.93%)
1	R_1	S_1	$\pi_r\pi_s$	0.0025	0.0100	0.0173	0.0342	0.0770	0.1108	0.1380	0.1685
				(0.0006)	(0.0013)	(0.0017)	(0.0023)	(0.0034)	(0.0041)	(0.0045)	(0.0048)
2	R_1	S_2	$\pi_r(1-\pi_s)(1-\pi_\omega)$	0.6677	0.6715	0.6713	0.6669	0.6472	0.6279	0.6112	0.5917
				(0.0204)	(0.0147)	(0.0133)	(0.0120)	(0.0108)	(0.0105)	(0.0103)	(0.0102)
3	R_1	S_3	$\pi_r(1-\pi_s)\pi_\omega$	0.2608	0.2520	0.2460	0.2351	0.2139	0.2002	0.1902	0.1796
				(0.0165)	(0.0133)	(0.0125)	(0.0118)	(0.0110)	(0.0108)	(0.0106)	(0.0106)
4	R_2	S_1	$(1-\pi_r)\pi_s$	0.0002	0.0007	0.0012	0.0023	0.0051	0.0072	0.0089	0.0108
				(0.0001)	(0.0001)	(0.0002)	(0.0003)	(0.0005)	(0.0007)	(0.0009)	(0.0011)
5	R_2	S_2	$(1-\pi_r)(1-\pi_s)(1-\pi_\omega)$	0.0494	0.0479	0.0470	0.0455	0.0427	0.0409	0.0395	0.0379
				(0.0050)	(0.0045)	(0.0043)	(0.0042)	(0.0040)	(0.0039)	(0.0039)	(0.0038)
6	R_2	S_3	$(1-\pi_r)(1-\pi_s)\pi_\omega$	0.0193	0.0180	0.0172	0.0160	0.0141	0.0130	0.0123	0.0115
				(0.0029)	(0.0020)	(0.0018)	(0.0016)	(0.0013)	(0.0012)	(0.0011)	(0.0011)

Notes. Standard errors in parentheses, derived by delta method. *: fraction of sample satisfying $\lvert r_i - s_i \rvert \leq \delta$ condition, where r is P14 earnings and s is FRS earnings. R_1: correct match between FRS and P14 datasets. R_2: mismatch. S_1: no survey measurement error. S_2: measurement error. S_3: measurement error plus contamination. Calculations based on estimates shown in Table 1.

Table 3. Reliabilities of P14 (*r*) and FRS (*s*) earnings, by definition of completely labelled group

δ	Sample fraction (%)	P14 data, *r*	FRS data, *s*
0.000	0.25	0.523	0.642
0.001	1.00	0.523	0.651
0.002	1.74	0.522	0.630
0.005	3.43	0.522	0.624
0.010	7.74	0.522	0.615
0.015	11.14	0.521	0.610
0.020	13.87	0.521	0.607
0.025	16.93	0.520	0.604

Notes. Reliability: squared correlation of true earnings ξ and measure y, where y is r or s. Calculations based on estimates shown in Table 1. Observations are completely labelled if $|r_i - s_i| \leq \delta$.

Table 4. Proportionate biases in OLS estimates resulting from using P14 or FRS data

δ	Sample fraction (%)	ξ LHS		ξ RHS	
		P14, r	FRS, s	P14, r	FRS, s
0.000	0.25	0.931	0.984	0.737	0.856
		(0.008)	(0.003)	(0.017)	(0.018)
0.001	1.00	0.933	0.984	0.737	0.849
		(0.006)	(0.003)	(0.018)	(0.013)
0.002	1.74	0.935	0.983	0.738	0.845
		(0.006)	(0.003)	(0.018)	(0.012)
0.005	3.43	0.936	0.981	0.738	0.840
		(0.006)	(0.003)	(0.018)	(0.011)
0.010	7.74	0.938	0.979	0.737	0.833
		(0.006)	(0.003)	(0.018)	(0.010)
0.015	11.14	0.939	0.977	0.737	0.829
		(0.006)	(0.003)	(0.018)	(0.010)
0.020	13.87	0.939	0.976	0.737	0.827
		(0.006)	(0.003)	(0.018)	(0.010)
0.025	16.93	0.940	0.975	0.737	0.824
		(0.006)	(0.004)	(0.018)	(0.011)

Notes. Table entries show proportional asymptotic biases in OLS estimates were true earnings ξ replaced by r (P14 data) or s (FRS data), with 1 meaning no bias. Calculations use the formulae in KY's equations (15), (16), (19), (21) and estimates reported in Table 1. Standard errors in parentheses derived by delta method. Observations are completely labelled if $|r_i - s_i| \leq \delta$.