# Modelling errors in survey and administrative data on employment earnings:

## sensitivity to the fraction assumed to have error-free earnings

Stephen P. Jenkins (LSE)
Email: s.jenkins@lse.ac.uk

Co-author: Fernando Rios-Avila (Levy Institute)

Family Finances Conference
2020-07-08

LSE

# Outline

1. Background and relevance
2. The FRS-P14 linked dataset for 2011/12: measures of employment earnings from FRS and HMRC for consenting FRS respondents
   - Secure data used via special contract with DWP
   - First study of its type for the UK (afaik)
3. The **KY** model of measurement errors: **K**apteyn & **Y**pma, *J Labor Econ*, 2007
4. Findings (headlines only)
5. Conclusions
6. Current project research

- Based on a paper now out in: *Economics Letters*, online first (preprint open-access version here)

LSE

# Background and relevance

- Accuracy of survey earnings data essential for accuracy of data about household income, inequality and poverty
  - Employment earnings ~ 2/3 of total weekly household income among all households, and ~ 80% for households with head aged 25–54 (FRS 2011/12)
- Earnings also of interest in own right: as outcome or as explanatory variable
- Often assumed that admin data are error-free or at least much more accurate than survey data
  - Hence too interest in survey data substitution (also to reduce burden)
- Often assumed that survey measurement errors are 'classical', so 'true' inequality under-estimated; and also that low earners over-report & high earners under-report (errors exhibit 'regression to the mean')
- Most empirical evidence for the USA; none for the UK

# FRS-P14 linked dataset for 2011/12 on gross employment earnings: **FRS**

- For each job, respondents are asked what the last amount received was, followed by a question about the period to which that amount refers (month, year, etc.)
- Responses converted to "£ p.w." pro rata by FRS data producers, which we convert to "£ p.a."
- 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)

LSE

# FRS-P14 linked dataset for 2011/12 on gross employment earnings: **P14**

For the FRS respondents in employment who gave their consent to link responses to P14 admin records (Q at very end of FRS questionnaire)

- "P14" label because compiled from employers' returns on P14 forms to HMRC about employees' wages and salaries paid and taxes and NIC withheld

- Around 60% of FRS respondents provided consent to data linkage
- DWP statisticians linked the FRS and P14 data deterministically
  - Match keys: first name, last name, postcode, sex, and date of birth
  - Linkage rate in 2010/11 was 82% (rate for 2011/12 not known to us)
  - So, overall, we have around 50% of FRS earners in the linked dataset
    - Potential bias issues? As in virtually all previous research on this topic, we ignore these issues
- 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)
- 5,971 linked obs (i.e. excluding 420 with imputed/edit earnings)
- Summary graphs in 'additional slides' at end

LSE

# Kapteyn-Ypma (KY) model: distribution of administrative earnings, $r_i$

Mixture of 2 types: P14 observations correctly matched with an FRS respondent, and P14 observations incorrectly matched:

- (*R*1) $r_i$ equals *i*'s true earnings, $\xi_i$, with probability $\pi_r$
- (*R*2) $r_i$ is the earnings of someone else in the full P14 dataset, $\zeta_i$, with probability $(1-\pi_r)$

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

LSE

# Kapteyn-Ypma (KY) model: distribution of survey earnings, $s_i$

Mixture of 3 types: observations with error-free earnings; with measurement error; with error and contamination

- ($S1$) $s_i$ equals true earnings, $\xi_i$, with probability $\pi_s$
- ($S2$) $s_i$ contains response error with a regression-to-the-mean component, with probability $(1-\pi_s)(1-\pi_\omega)$
- ($S3$) $s_i$ as per $S2$ but with additional contamination, with probability $(1-\pi_s)\pi_\omega$

where $\pi_s$: Pr(survey earnings error-free), and

$\pi_\omega$: Pr(survey earnings include contamination too)

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

# The model implies six latent classes

- Each sample observation may belong to one of 6 latent classes characterized by the combinations of cases $R1$, $R2$ with $S1$, $S2$, or $S3$
  - $(R1, S1)$, $(R1, S2)$, $(R1, S3)$, $(R2, S1)$, $(R2, S2)$, $(R2, S3)$

- Class membership probabilities $\pi_j$ for $j = 1, \ldots, 6$, depend on model probabilities:
  - $\pi_r$: Pr(admin data error-free, i.e. no mis-match)
  - $\pi_s$: Pr(survey earnings error-free)
  - $\pi_\omega$: Pr(survey earnings include contamination error)

# Estimation

- For estimation, KY assume true earnings and errors are each independently and identically normally distributed:
  ( ), ( ), ( ), and ( )
- Maximize likelihood after fixing the size of the first group – an identification assumption about the fraction with error-free earnings ('completely labelled')
  - Obs with $r_i \approx s_i$ are the ones with observed earnings = true earnings
  - How close do $r_i$ and $s_i$ have to be in order to be judged 'equal'?
- KY use 1 completely labelled fraction (CLF): 14.8%
- Are findings sensitive to the choice of CLF?
- We fit models separately for 8 CLFs: range of values from slightly greater than KY's to much smaller ones
  - $|r_i - s_i| \leq \delta$ with threshold $\delta$ taking values in the range [0,0.025], implying sample fractions 0.25% through 16.93%
- All model estimates precisely estimated ($p < 0.01$)

# Selected estimates

(For details, see tables in 'additional slides' at end)

$\pi_s$: Pr(survey earnings error-free)

- Varies directly with choice of completely-labelled fraction (CLF)

$\pi_\omega$: Pr(survey earnings include contamination error)

- between 23% and 28% depending on CLF
- If you assume a lower % error-free $\Rightarrow$ greater % with error!

$\pi_r$: Pr(admin data error-free, i.e. no mis-match)

- Pr(mismatch) = $(1-\pi_r) \approx 6\%$, regardless of CLF

$\rho$: degree of 'regression to the mean' in survey error

- $\rho \approx 0$, as found by KY
- Earlier studies assuming admin data error-free found $\rho << 0$

Latent classes with highest probabilities:

- $(R1, S2)$ = correctly-matched admin data combined with survey error (range 67%–59%), and $(R1, S3)$ survey error plus contamination (range 26%–18%); value in range depends on CLF

# Reliability of earnings measures

- Reliability = squared correlation(measure, true)

| $\delta$ | Sample fraction (%) (completely-labelled fraction) | Reliability | |
|---|---|---|---|
| | | admin | survey |
| 0 | 0.25 | 0.686 | 0.843 |
| 0.001 | 1.00 | 0.688 | 0.835 |
| 0.002 | 1.74 | 0.689 | 0.831 |
| 0.005 | 3.43 | 0.691 | 0.825 |
| 0.010 | 7.74 | 0.692 | 0.815 |
| 0.015 | 11.14 | 0.692 | 0.810 |
| 0.020 | 13.87 | 0.692 | 0.807 |
| 0.025 | 16.93 | 0.692 | 0.804 |

NB unfortunately reliabilities reported by Jenkins and Rios-Avila (2020: Table 3) are incorrect, but conclusions are not affected

# Reliability estimates

(0.69 for *r*, 0.80–0.84 for *s*)

- FRS employee earnings data are more reliable than the P14 employee data!
  - KY had similar finding (for Swedish workers aged 50+, 2003) regarding survey data compared to register data
  - Even a small amount of mis-match has serious consequences for data quality
  - Raising the fraction of FRS observations that gives consent and is linked (as in recent years) doesn't necessarily ensure that the fraction *correctly linked (i.e. without mismatch*) increases commensurately, since linking 'technology' has remained the same, as far I know

LSE

# Conclusions

- From a reliability and data quality perspective, admin record data should not be treated as 'perfect' and survey data 'imperfect'
  - Substitution of survey earnings measure by an admin measure problematic
    - Relevance of the point not changed by changing from P14 data (part of the WPLS) to RTI data
  - Arguing that admin data are correct *de jure* (because they are used to administer) is a different argument: cf. Britton et al. (*JRSSA* 2019) using Student Loan Company records data on graduates' earnings
- Caveat: the nature and prevalence of measurement errors are contingent on model assumptions
  - Our paper examines one feature of the KY model (choice of CLF); but not the KY model per se
- Caveat: findings are re earnings, not e.g. cash benefits

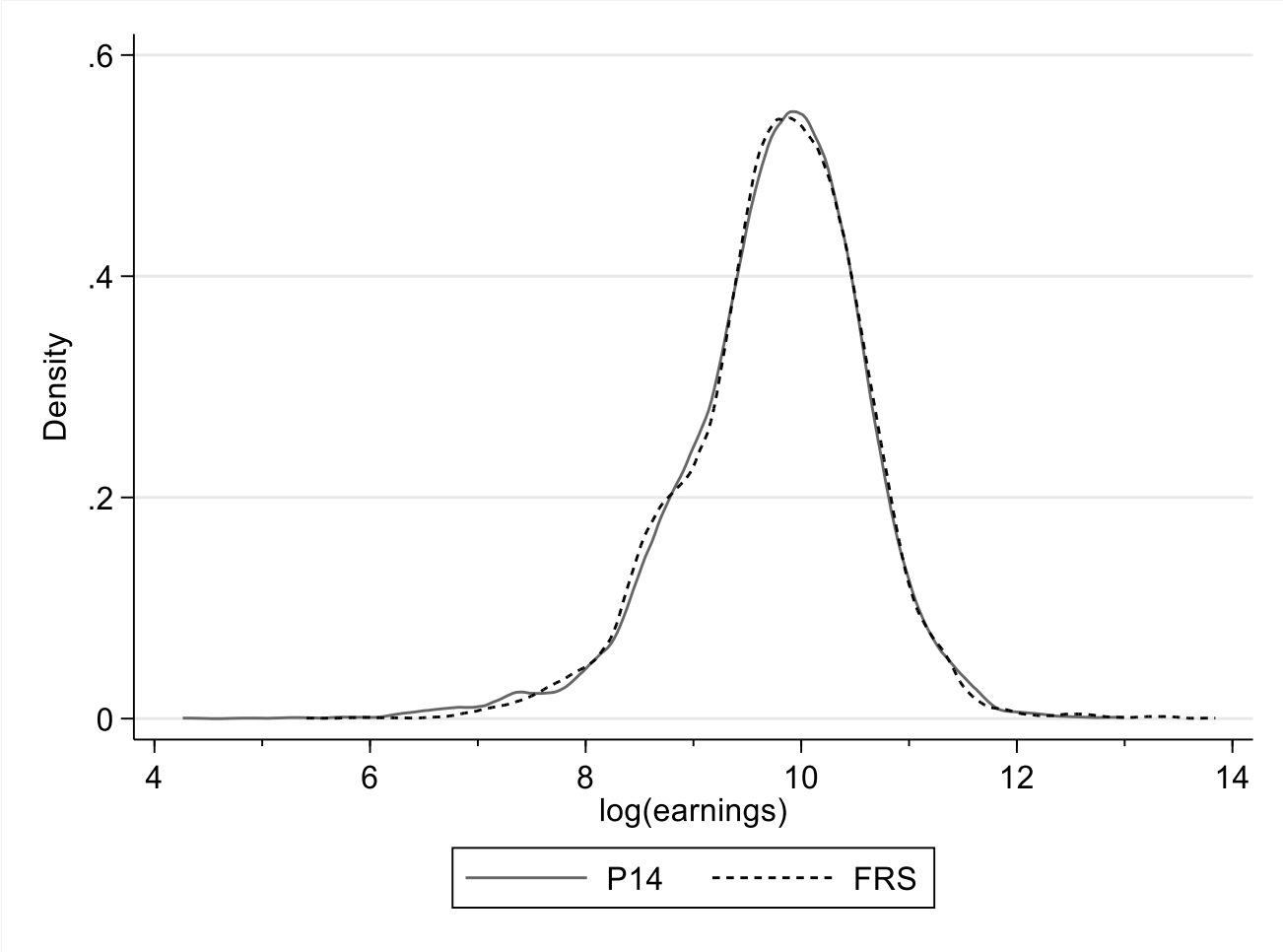
LSE

# Current project research

- Predicting 'true earnings' by combining information from both survey and admin sources
  - We have a paper replicating Meijer et al. (*J Bus & Econ Stats* 2012) who used KY's model and data; forthcoming IZA DP series
- Model extensions and variants
  - Extending the KY model, to allow for measurement error in admin data
  - Adding covariates
    - Preliminary result: measurement error variances lower for those with payslip
  - Alternative model in which survey and admin data are measures of different latent earnings concepts ('current' and 'annual') which are correlated
    - To directly address the 'differences in reference period' issue
    - NB KY model assumes each source provides a measure of same earnings concept

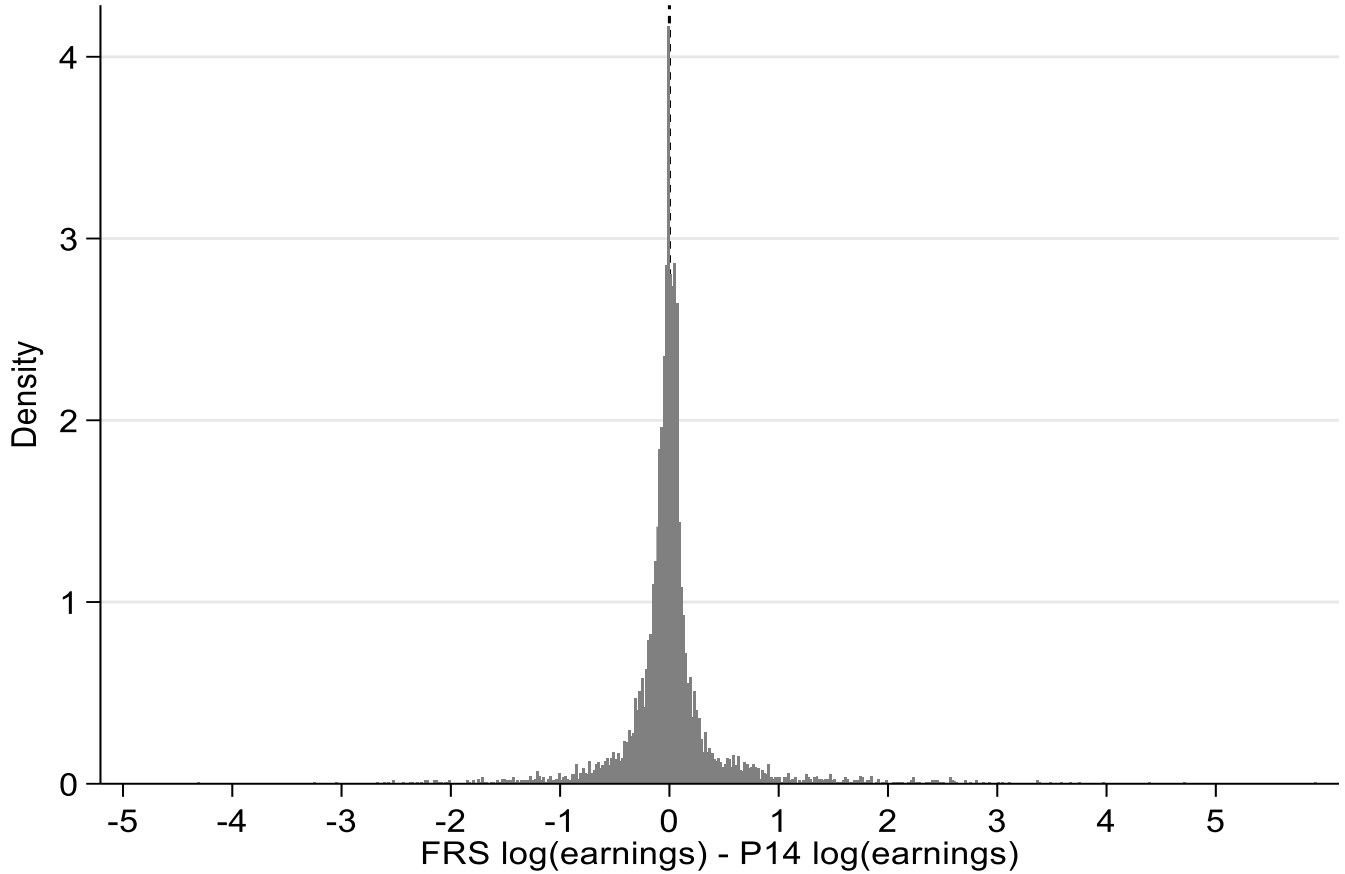
# Additional slides

# Distributions of log(earnings) in P14 data ($r_i$) and FRS data ($s_i$)




*Note:* The means of $r_i$ and $s_i$ are 9.75 and 9.77 with standard deviations 0.842 and 0.813, respectively. Densities shown are kernel density estimates.

# Distribution of log(earnings) differences: $s_i - r_i$




*Note* The distribution of differences ($s_i - r_i$) has a mean of −0.02 with standard deviation 0.496.

**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%)[a] | 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%) |
| $\mu_\xi$ | 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) |
| $\sigma_\xi$ | 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) |
| $\mu_\zeta$ | 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) |
| $\sigma_\zeta$ | 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) |
| $\mu_\omega$ | −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) |
| $\sigma_\omega$ | 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) |
| $\mu_\eta$ | −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) |
| $\sigma_\eta$ | 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) |
| $\pi_r$ | 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) |
| $\pi_s$ | 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) |
| $\pi_\omega$ | 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) |
| $\rho$ | −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 $N = 5971$. Parameters $\mu$ and $\sigma$ refer to means and standard deviations respectively. $\xi$: true earnings. $\zeta$: $r$ if mismatch of FRS case with P14 case. $\omega$: contamination error in $s$. $\eta$: measurement error in $s$. $\rho$: regression to the mean in $s$. $\pi_r$: Pr(FRS obs correctly matched with P14 obs). $\pi_s$: Pr($s$ reported correctly). $\pi_\omega$: Pr($s$ contains contamination error).
Source: authors' estimates from Linked FRS-P14 dataset.
[a]Fraction of sample satisfying $\lvert r_i - s_i \rvert \leq \delta$ condition, where $r$ is P14 earnings and $s$ is FRS earnings.

# Estimates of class probabilities, $\pi_j$

**Table 2**
Estimates of class probabilities ($\pi_j$) from Kapteyn–Ypma Full Model, by definition of completely labelled group.

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

Notes. Standard errors in parentheses, derived by delta method. $R_1$: 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.
[a]Fraction of sample satisfying $\lvert r_i - s_i \rvert \leq \delta$ condition, where $r$ is P14 earnings and $s$ is FRS earnings.

LSE