Finite mixture models for linked survey and administrative data: estimation and post-estimation

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

Co-author: Fernando Rios-Avila (Levy Institute)
Email: friosavi@levy.org

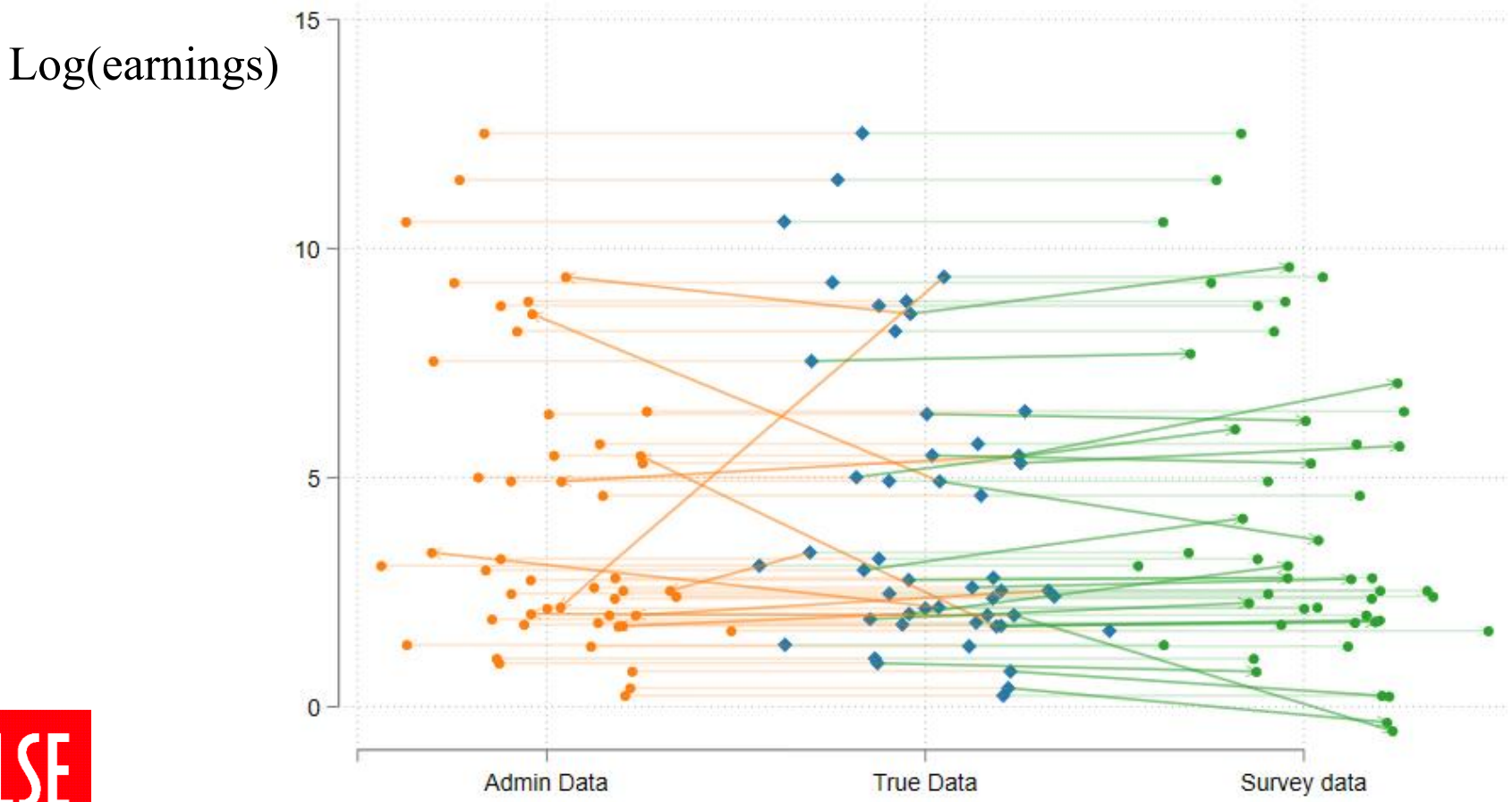
This talk is based on our paper: IZA Discussion Paper 14404,
with Stata programs at SSC (`ssc describe ky_fit`)
For our substantive application to UK data: see IZA Discussion Paper 14405

Linked survey and administrative data

- Linked datasets: what are they?
  - Reports by respondents to a *household survey* on a variable such as earnings linked to reports on the same variable in an *administrative dataset* (e.g. personal income tax or social security data) *for the same respondents*
- Linked datasets can be used to tell us about data quality (measurement errors)
  - How much bias and spurious variation in the measures of interest?
  - Are errors are correlated with the 'true' (unobserved) measure?
    - Negative correlation means that low-earners over-report and high-earners under-report ('mean-reverting' errors); relevant to whether observed inequality > or < true inequality
- First generation studies (many) assume linked admin data provide error-free measures; all measurement errors arise in the survey reports
- Second generation studies (few) allow for errors in the administrative data as well, and errors in data linkage

Visualizing linked data on log(earnings)

- Two observations per worker: one from survey, one from admin data
- Unobserved: true value per worker
- Measurement error: differences in heights of the 3 points per worker
  - First Generation studies assume admin data = true data



Finite Mixture Models (FMMs)

- FMMs: useful for analyzing linked datasets because they allow you to succinctly describe both (i) the distribution of the 'true' (error-free) substantive variable of interest, as well as (ii) the distributions of the various types of measurement error
- FMM *latent classes* characterized by different combinations of error-ridden and/or error-free survey and administrative data observations
- FMM *probabilities of latent class membership* depend more fundamentally on the probabilities of the different types of error
- However, the 'structural' FMMs needed for this analysis cannot be fitted using readily-available software such as Stata's **`fmm`** suite
- We provide and illustrate Stata commands for various Second Generation models:
  1. *Estimation* of a general class of FMMs using linked data (**`ky_fit`**)
  2. *Post-estimation* commands for assessment of reliability, marginal effects, prediction of hybrid earnings variables that combine information from both data sources (**`ky_estat`**, **`ky_p`** with **`predict`** and **`margins`**, **`ky_sim`**), and data simulation (**`ky_sim`**)

LSE

We build on and extend earlier second generation models

- Kapteyn & Ypma (*JoLE* 2007, 'KY'): first to introduce FMM allowing for errors in linked admin data error, but only considered linkage error as its source
- Meijer, Rohwedder, and Wansbeek (*JBES* 2012, 'MRW'): derived formulae for a number of hybrid earnings predictors combining information from both survey and administrative data and showed that they were more reliable than either the survey or the administrative data measure (used KY's estimates)
- We extend KY's model: (1) to allow for measurement error in the (linked) admin as well as linkage error, and (2) MRW-type predictors; (3) allow measurement error distributions to vary with observed covariates

Admin earnings, r_i: 3 types of observation

Mixture of 3 types: (i) observations may be correctly linked (probability π_r) or mismatched, and (ii) correctly-linked cases may be error-free (probability π_v) or contain measurement error:

- ($R1$) r_i equals i's true earnings, ξ_i
- ($R2$) r_i contains mean-reverting measurement error
- ($R3$) mismatch: r_i is the earnings of someone else in the full admin dataset, ζ_i

$$r_i = \begin{cases} \xi_i \quad \text{with probability } \pi_r\pi_v & \text{(type } R1\text{)} \\ \xi_i + \rho_r(\xi_i - \mu_\xi) + v_i \quad \text{with probability } \pi_r(1-\pi_v) & \text{(type } R2\text{)} \\ \zeta_i \quad \text{with probability } (1-\pi_r) & \text{(type } R3\text{)} \end{cases} \qquad (1)$$

LSE

Survey earnings, s_i: 3 types of observation

Mixture: (i) observations with error-free earnings; (ii) with measurement error; with error plus contamination

- ($S1$) s_i equals true earnings, ξ_i, with probability π_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 $S2$ plus contamination error as well, with probability $(1-\pi_s)\pi_\omega$

where π_s: Pr(survey earnings error-free), and

π_ω: Pr(survey earnings include contamination too)

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

LSE

General model

- Observations in the linked dataset are a **mixture of nine types (latent classes *j* = 1, …, 9)**, i.e., depending on the possible combinations of the 3 administrative and the 3 survey observation types
- **Latent class probabilities are π_j, $j = 1, \ldots, 9$**
  - E.g., group 1 contains observations with the combination ($R1$, $S1$) with probability $\pi_1 = \pi_r\pi_\upsilon\pi_s$,
  - E.g., group 2 contains observations with the combination ($R1$, $S2$) with probability $\pi_2 = \pi_r\pi_\upsilon(1-\pi_s)(1-\pi_\omega)$, etc., etc.
- FMM specification is completed by assumptions about the latent class earnings densities, $\boldsymbol{f_j(r_i, s_i)}$ for each $j = 1, \ldots, 9$.
- We assume that true earnings (ξ_i), mismatched earnings (ζ_i), and errors (υ_i, η_i, ω_i) are each **normally distributed** with the exception that true earnings and contamination errors (ω_i) are bivariate normal
  - We assume normality (as other researchers do) to fit models by maximum likelihood (see below) and because it facilitates post-estimation derivations
- We **allow distributions to vary with observed characteristics** by writing transformations of model parameters as linear indices of characteristics, i.e.,

$$G(\gamma_i) = \alpha\gamma + \beta\gamma X_i \quad \text{for generic parameter } \gamma$$

  - $G(.)$: identity function for means (μ); log function for SDs (σ); logistic function for probabilities (π), Fisher's Z transformation for correlations (ρ)

LSE

Estimation by ML (fit using `ml`)

- Log-likelihood function in general:

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

- But simplifies here because: (i) for class 1, $r_i = s_i$, so distribution degenerates to a univariate normal distribution with mean μ_ξ and variance σ_ξ^2; (ii) class membership is known for observations in this group
- Hence, log-likelihood function becomes:

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

- Identification of model parameters relies on
  1. (conditional) normality;
  2. how many obs are in group 1, i.e., for how many obs are r_i and s_i considered to be near-enough equal ('completely labelled' fraction)?
  3. NB latent class probabilities depend on only three underlying parameters (not nine)

LSE

Estimation using **ky_fit**

ky_fit fits the general FMM and special cases of it including the KY model:

```
ky_fit r_var s_var [cl_var] [if] [in] [fw pw aw iw]
[, model(#) options]
```

- **r_var** and **s_var** are required variables corresponding to the admin measure r_i and survey measure s_i
- **cl_var** is a binary variable that identifies observations belonging to Class 1 ('completely labelled')
  - If **cl_var** is not declared, **ky_fit** creates a binary indicator variable named **__ll__** equal to one for observations for which **abs(r_var-s_var)<= #d.** The default value of **#d** is 0, but other values can be declared using **delta(#d)**
- **model(#)** specifies which of 8 possible FMMs is fitted, ranging from very basic to the general one shown above, and including KY's (our model #4)
- Other **options** for controlling maximization and making parameters functions of covariates

Post-estimation: `ky estat`

`ky estat`: post-estimation command to derive get summary statistics for the model parameters, as well as assessment of hybrid data measures proposed by MRW

`estat [pr_{t|i|sr|all} rel xirel, sim reps(# 50)]`

`pr_{t|i|sr|all}`: summary statistics for latent class probabilities

`rel`: reliability statistics, where x is survey or admin var, and e is true var

- **`R1`**: Cov(x, e)/Var(x) slope coeff from regression of true on observed
- **`R2`**: Cov(x, e)^2/ [Var(e)Var(r)] squared correlation true and observed

`xirel`: reliability statistics for hybrid measures

`sim`: request numerical estimation for reliability statistics, with 50 Reps as default

Predictions and marginal effects: `ky p`

`ky p` derives predicted values and marginal effects for selected parameters of interest in their original scales

`predict` and **`margins`**: all distribution parameters, latent class moments, and class probabilities

`predict`: posterior class probabilities, and Bayesian classification

`predict prefix, star`: MRW-type hybrid/bias-corrected measures / predictions

Includes predictions assuming only survey data available

Simulate linked data using `ky_sim`

- Simulate data with user-provided parameters, which may those include from previously-fitted models
- Useful for analyzing data properties and for creation of synthetic data

1. **`ky_sim,[model(#)nobs(#)parameters]`**
   - Simulates data based on set of parameters (no covariates)
2. **`ky_sim,[est_sto(name)est_sav(name) prefix(str)]`**
   - Simulates data from parameters of models previously fitted that have been stored in memory or saved

Illustration: simulate from KY estimates, refit their model, and do MRW predictions

- KY's model (#4): admin data could be mismatched; survey data with possible mean-reverting measurement errors and contamination; possible admin-survey linkage error
- KY's parameter estimates (from data for 400 Swedish men and women aged 50+), including significant linkage error (~4%) but no significant mean-reversion in survey

```
global mean_e 12.283 ; global mean_t 9.187 ;
global mean_w (-0.304); global mean_n (-0.048) ;
global sig_e 0.717 ; global sig_t 1.807 ;
global sig_w 1.239 ; global sig_n 0.099 ;
global pi_r 0.959 ; global pi_s 0.152 ;
global pi_w 0.156 ; global rho_s (-0.013) ;
** Simulate data
ky_sim, nobs(400) model(4) seed(101) ///
 mean_e($mean_e)mean_t($mean_t)mean_w($mean_w)mean_n($mean_n) ///
 sig_e($sig_e) sig_t($sig_t) sig_w($sig_w) sig_n($sig_n) ///
 pi_r($pi_r) pi_s($pi_s) pi_w($pi_w) rho_s($rho_s) clear ///
 eststo m0
```

LSE

Illustration: summary statistics from simulated data match KY's

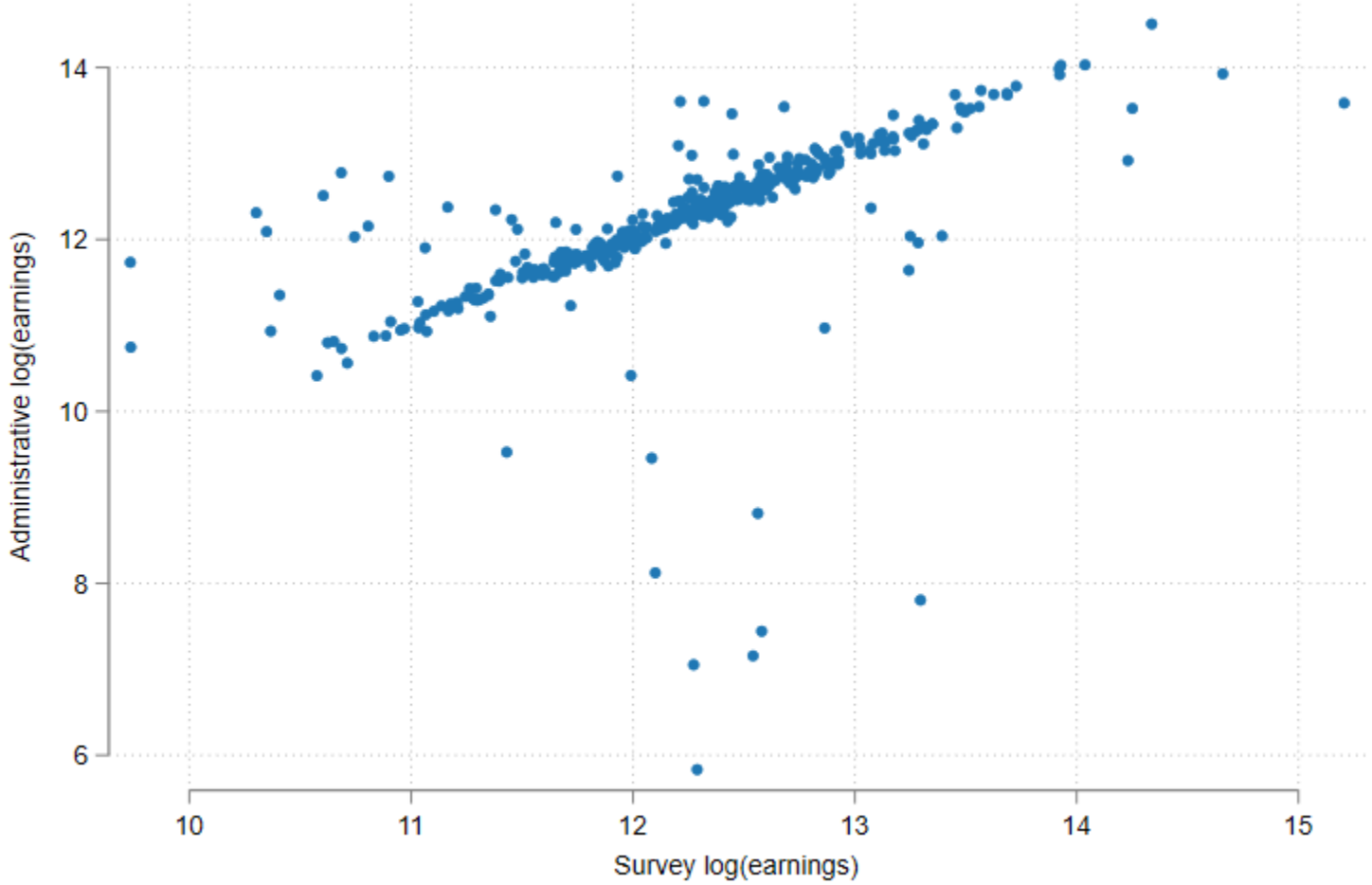
- True log earnings (e), measurement error (n), contamination (w), mismatched earnings (t), error probabilities, completely labelled fraction, types of observations, and latent classes (5 in total)

```
. summarize *, sep(0)
```

Variable	Obs	Mean	Std. dev.	Min	Max
e_var	400	12.34898	.665869	10.4206	14.51099
n_var	400	-.0513431	.1030404	-.3312704	.2292065
w_var	400	-.3139371	1.128783	-3.336294	2.629267
t_var	400	9.012969	1.753307	4.315567	13.78396
pi_si	400	.135	.3421515	0	1
pi_wi	400	.1525	.3599551	0	1
pi_ri	400	.9725	.16374	0	1
r_var	400	12.23967	.9549137	5.839129	14.51099
s_var	400	12.25409	.7501207	9.732128	15.20382
l_var	400	.1325	.3394581	0	1
rclass	400	1.0275	.16374	1	2
sclass	400	1.985	.5053845	1	3
class	400	2.0675	.6958119	1	5

Illustration: simulated KY data

- Most observations lie close to the 45° line, but significant minority lie quite distant
- Different regions occupied by different latent classes (can show this)



Fit KY model and simpler variants

```
constraint 1 [mu_n]_cons = 0
//  Basic
ky_fit r_var s_var l_var, model(1)  constraint(1)
estimates store model1

//  No mismatch
ky_fit r_var s_var l_var, model(2)
estimates store model2

//  No contamination
ky_fit r_var s_var l_var, model(3)
estimates store model3

//  Full model
ky_fit r_var s_var l_var, model(4)
estimates store model4
```

Model estimates are very close to those reported by KY

	(1) Original		(2) Full mo~l		(3) No cont~n		(4) No mism~h		(5) Basic M~l	
mu_e	12.283	(.)	12.349	(0.034)	12.306	(0.038)	12.240	(0.048)	12.246	(0.037)
mu_n	-0.048	(.)	-0.061	(0.006)	-0.062	(0.006)	-0.059	(0.006)	0.000	(.)
mu_t	9.187	(.)	8.586	(0.678)	11.622	(0.256)				
mu_w	-0.304	(.)	-0.344	(0.148)			0.479	(0.284)		
ln_sig_e	-0.333	(.)	-0.406	(0.036)	-0.285	(0.036)	-0.047	(0.035)	-0.047	(0.035)
ln_sig_n	-2.313	(.)	-2.295	(0.048)	-2.270	(0.047)	-2.268	(0.046)	-0.449	(0.038)
ln_sig_t	0.592	(.)	0.501	(0.315)	0.622	(0.098)				
ln_sig_w	0.214	(.)	-0.026	(0.112)			0.731	(0.100)		
arho_s	-0.013	(.)	-0.022	(0.010)	-0.015	(0.010)	-0.026	(0.010)	-0.680	(0.054)
lpi_r	3.152	(.)	3.520	(0.335)	1.838	(0.159)				
lpi_s	-1.719	(.)	-1.844	(0.148)	-1.708	(0.150)	-1.879	(0.147)	-1.879	(0.147)
lpi_w	-1.688	(.)	-1.784	(0.189)			-1.683	(0.161)		
N			400		400		400		400	
ll			-543.028		-595.528		-695.498		-1041.749	

LSE

Use **margins** to transform parameter estimates to their natural metric

```
margins, predict(mean_e) predict(sig_e) ///
      predict(mean_t) predict(sig_t) ///
      predict(mean_w) predict(sig_w) ///
      predict(mean_n) predict(sig_n)  ///
      predict(pi_r) predict(pi_s) ///
      predict(pi_w) predict(rho_s)

[output partially omitted]
```

	Delta-method				
	Margin	Std. Err.	z	P>|z|	[95% Conf. Interval]
_predict					
1	12.34936	.0335341	368.26	0.000	12.28364 12.41509
2	.6659948	.023718	28.08	0.000	.6195083 .7124813
3	8.586231	.6782982	12.66	0.000	7.256791 9.915671
4	1.650615	.5192742	3.18	0.001	.6328562 2.668374
5	-.3435237	.1479331	-2.32	0.020	-.6334672 -.0535803
6	.9747349	.1089581	8.95	0.000	.7611809 1.188289
7	-.0608566	.0063531	-9.58	0.000	-.0733084 -.0484048
8	.1007999	.0048806	20.65	0.000	.091234 .1103657
9	.9712426	.0093542	103.83	0.000	.9529088 .9895765
10	.1365808	.0174403	7.83	0.000	.1023985 .1707632
11	.1437948	.0233102	6.17	0.000	.0981077 .1894819
12	-.0220813	.0097204	-2.27	0.023	-.041133 -.0030297

- If you specify a model in which parameters depend on observed covariates, **margins** can be used to obtain marginal mean estimates of the parameters
  - For example, suppose your **ky_fit** command specifies that the log of the survey measurement error SD depends on a binary indicator variable for the respondent's sex using the option **ln_sig_v(i.sex)**:

```
margins sex, predict(sig_v)
```

LSE

Hybrid variables (combine survey and admin data information), à la MRW

- 7 hybrid variables from combining data in different ways
- Estimates close to those that MRW report (using KY estimates)
- Hybrid variables, esp. 3, 4, 5, 6, have higher reliability and lower MSE than observed survey or admin data

```
. estat xirel
Rel Statistics for 'e' predictions

          Rel1       Rel2        MSE    E(Bias)  Var(Bias)
r_var   0.4955     0.4806     0.4945    -0.1060     0.4833
s_var   0.7569     0.7439     0.1583    -0.0970     0.1489
  e_1   0.5440     0.5267     0.4079    -0.1032     0.3973 Wgt unc
  e_2   0.5437     0.5281     0.4077    -0.1024     0.3973 Wgt unc unbi
  e_3   0.9987     0.9873     0.0056     0.0003     0.0056 Wgt con
  e_4   0.9907     0.9845     0.0069     0.0003     0.0069 Wgt con unb
  e_5   0.9911     0.9850     0.0066    -0.0009     0.0066 2-step
  e_6   0.9871     0.9838     0.0072    -0.0013     0.0072 2-step unb
  e_7   0.9917     0.7893     0.0938    -0.0009     0.0938 Sys-wide
```

Conclusions

- We introduce a new set of commands for estimation and post-estimation of FMMs for applications to linked survey and administrative data on earnings or similar variables.

```
ssc install ky_fit
```

- The FMM specifications (8 model versions) are those proposed by Jenkins and Rios-Avila (2021b), extending KY's
- The suite includes post-estimation commands for simulation, assessing reliability, and deriving highly reliable hybrid predictors of latent true earnings
- Our substantive applications of our models and software have been to UK linked data; some others are currently looking at US and Austrian data

Thank you! Questions or comments?

LSE

References (selected)

- Jenkins, S. P. and Rios-Avila, F. (2020). Modelling errors in survey and administrative data on labour earnings: sensitivity to the fraction assumed to have error-free earnings. Economics Letters, 192: 109253
- Jenkins, S. P. and Rios-Avila, F. (2021a). Measurement error in earnings data: replication of Meijer, Rohwedder, and Wansbeek's mixture model approach to combining survey and register data, Journal of Applied Econometrics, online first
- Jenkins, S. P. and Rios-Avila, F. (2021b). Reconciling reports: modelling employment earnings and measurement errors using linked survey and administrative data. IZA Discussion Paper 14405. Submitted
- Jenkins, S. P. and Rios-Avila, F. (2021c). Finite mixture models for linked survey and administrative data: estimation and post-estimation. IZA Discussion Paper 14404. Submitted to The Stata Journal
- [KY] 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
- [MRW] 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 & Economic Statistics, 30 (2): 191–201

LSE