Finite Mixture Models for linked survey and administrative data

Estimation and post-estimation

Stephen P. Jenkins[1]   Fernando Rios-Avila [2]

[1]London School of Economics

[2]Levy Economics Institute of Bard College

Virtual Stata Conference, August 6, 2021
Feliz Dia Bolivia!

Table of Contents

Introduction: Why Linked data?

- In economics, as well as other sciences, we are often interested in analyzing income data of high quality. The truth. why? (poverty and inequality for once)
- More often than not, however, we may not have access to "the true" data, but proxies.
  - We usually have access to survey data. (Which may suffer from measurement errors)
  - But we may also have access to administrative data. Which is *almost* the truth.
- Each source has its strength and weaknesses for statistical analysis.
- Having access to data that links both sources allows us to investigate the quality of available data.

What is the problem with Admin and Survey data?

Survey Data:

- Surveys data comes along with other data of interest. Demographic characteristics, employment history, geographical data, etc.
- This allows us to make rich analysis across groups of interest.
- Income data, however, may be measure with error.

Administrative/Register data

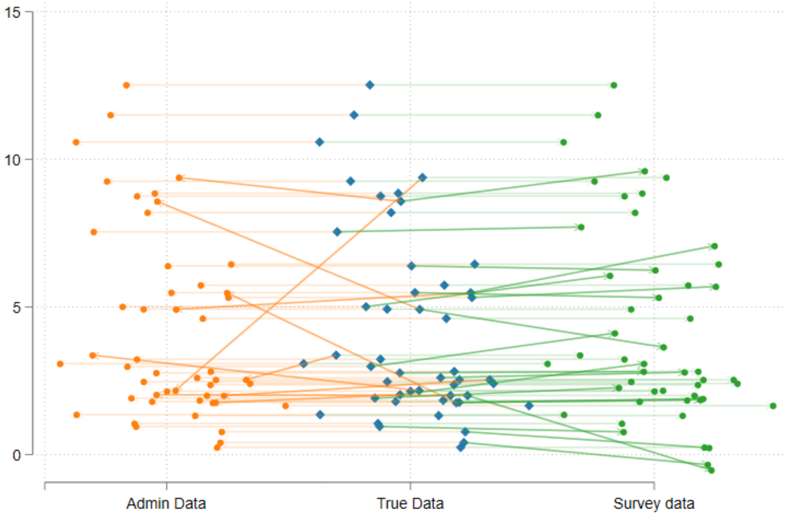
- Usually assumed to measure the "truth", or as close as possible to true data.
- It rarely has data on individual characteristics so, on its own, it is of limited used for further analysis.

The best of both worlds: If Survey and Admin Data data could be "linked", we could get better answers to problems of interest.

Linked Data, for a better Analysis

- First Generation Studies have used linked data to analyze the quality of Survey income data, assuming register data is error free.
  They conclude Survey income data may be biased. Classical measurement error and reversion to the mean (rtm) error.
- The problem, however, is that linked register data may not be error free.
  Linking data may be done through an statistical matching process. This may introduce errors in the data, linking errors incorrectly, generating an even greater measurement error problem.
- Second Generation Studies lift the "register-error-free" assumption, suggesting that as a whole, Survey income data, is still more reliable, than linked register data.
- And accounting for errors in Register data reduces other problems typically observed in survey data.

Visualizing Linked Data



Making better use of Linked Data

- Given the nature of the different sources of data, and the presence of unobserved errors, Finite Mixture Models are useful for analyzing linked data.
- Kapteyn and Ypma (2007) proposed a second generation model to analyze measurement errors in linked data, using structural FMM.
- Meijer, Rohwedder and Wansbeck(2012) extends KY (2007), and propose that linked Register and Survey data could be combined to obtain hybrid measures that are closer to the truth.
- BUT: Neither of their proposed strategies can be applied using readily available software (`fmm`).

Contribution

What do we contribute:

- Extend on KY, allowing for a richer measurement error structure in register data. (Jenkins and Rios-Avila, 2021b)
- Implement methods for data combination and earnings predictions proposed by MRW.
- Build a user friendly set of commands: for the estimation `ky_fit`, post-estimation `ky_estat`, and data simulation ky_sim for this type of models.

Table of Contents

Data Structure

For each individuals i, we have linked records for the register r_i and survey s_i data.

Both measures are proxies for the true income measure ε_i, which is unobserved.

$$(r_i, s_i, \varepsilon_i) \forall i = 1 ... N$$

All variables are measured in logs().

We assume that administrative data is a mixture of 3 distributions:

- Correctly linked data without measurement error.

$$R1 : r_i = \varepsilon_i ; \pi_{r1} = \pi_r \pi_v$$

- Correctly linked data with RTM measurement error, and noise

$$R2 : r_i = \varepsilon_i + \rho_r(\varepsilon_i - \mu_\varepsilon) + \nu_i ; \pi_{r2} = \pi_r(1 - \pi_v)$$

- Incorrectly linked data

$$R3 : r_i = \zeta_i ; \pi_{r3} = 1 - \pi_r$$

Survey data

We assume that survey data is a mixture of 3 distributions:

- Report true earnings.

$$S1 : r_i = \varepsilon_i ; \pi_{s1} = \pi_s$$

- Report earnings with RTM error

$$S2 : s_i = \varepsilon_i + \rho_r(\varepsilon - \mu_\varepsilon) + \eta_i ; \pi_{s2} = (1 - \pi_s)(1 - \pi_\omega)$$

- Report earnings with RTM error + Contamination (reference period error)

$$S3 : s_i = \varepsilon_i + \rho_r(\varepsilon - \mu_\varepsilon) + \eta_i + \omega_i ; \pi_{s3} = (1 - \pi_s)\pi_\omega$$

Combined data, and Latent Structure

The combination of Survey and Register Data generates 9 latent groups, (3R x 3S = 9L)

For the model identification, We assume that all unobserved latent factors (the error structures), follow normal (or conditionally normal if covariates are used) distributions such that:

$$\begin{matrix} \varepsilon_i \\ \omega_i \end{matrix} \sim N \left(\begin{matrix} \mu_\varepsilon \\ \mu_\omega \end{matrix}, \begin{matrix} \sigma_\varepsilon^2 & \rho_\omega \sigma_\varepsilon \sigma_\omega \\ \rho_\omega \sigma_\varepsilon \sigma_\omega & \sigma_\omega^2 \end{matrix} \eta \right)$$

$$\zeta_i \sim N(\mu_\zeta, \sigma_\zeta^2); \eta_i \sim N(\mu_\eta, \sigma_\eta^2); \nu_i \sim N(\mu_\nu, \sigma_\nu^2)$$

All parameters are allowed to vary with explanatory variables:

$$G(\gamma) = \alpha_\gamma + \beta_\gamma' X$$

Model Estimation

This is a complex model that can be estimated via Maximum Likelihood. Internally, we use -ml-.

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

However, because Latent Class 1 $(R1, S1)$ both survey and register data measures income data without error, the above expression turns to:

$$LogL(\Theta, \Pi) = \sum_{i \in C1} \pi_1 log(f_i(\varepsilon_i|\Theta)) + \sum_{i \notin C1} log \sum_{j=2}^{9} \pi_j f_j(r_i, s_i|\Theta)$$

Identification of the model relies on the (conditional) normality assumption, and the size of LC1 group.

Once parameters θ and Π are obtained, estimators for ϵ_i (see MRW), can be obtained

Table of Contents

Model estimation: `ky_fit`

We propose a the command `ky_fit`, as a command that allows you to estimate the proposed model, including its simplifications (see Jenkins and Rios-Avila 2021c). This includes KY model.

```
ky_fit r_var s_var [cl_var] [if in wgts] [,model(#) options]

r_var    : (log) register data
s_var    : (log) survey   data
cl_var   : Dummy for Class 1 data
model(#): Type of FMM model
options : Estimation options, and modeling of parameters.
* Covariates can be added as explanatory variables
for specific parameters
```

Model post-estimation: `ky_estat`

`ky_estat`, is a post-estimation command that allows you get summary statistics for the model parameters, as well as assessment of data hybrid 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
R1: Cov(x,e)/Var(x) ;
R2: Cov(x,e)^2/[Var(e)Var(r)]
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, works to obtain 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`: hybrid/bias-corrected measures predictions.

Includes predictions assuming only survey data is available

Data Simulation `ky_sim`

This command allows being able to simulate data based on provided parameters, or previously estimated models.
Useful for analyzing data properties, and creation of synthetic data.

```
Opt1: ky_sim, [model(#) nobs(#) parameters]
Simulates data based on set of parameters (no covariates)

Opt2: ky_sim, [ est_sto(name) est_sav(name) prefix(str)]
Simulates data based on estimated models.
Previously estimated, stored in memory, or saved.
```

Example: Setup KY 2007

Defining data parameters following KY 2007

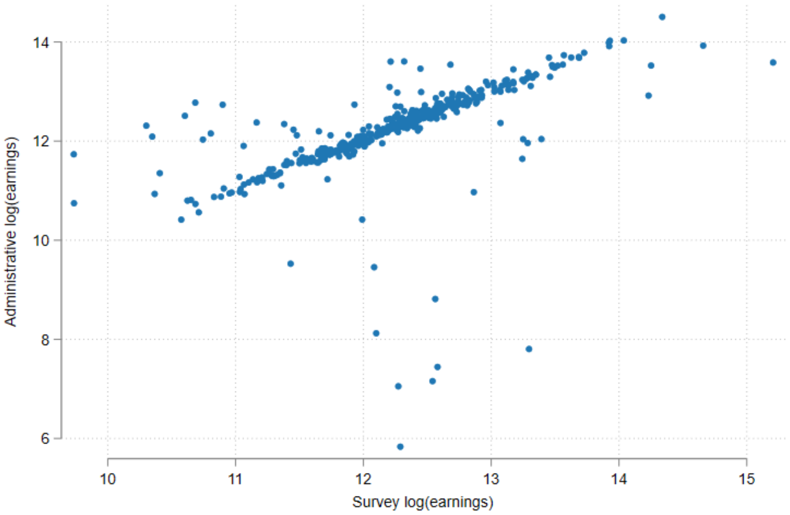
```
global mean_e  12.283  ; global mean_t   9.187
global mean_w  (-0.304); global mean_n  (-0.048)
global sig_e  0.717    ; global sig_t  1.807
global sig_w  1.239    ; global sig_n  0.099
global pi_r   0.959    ; global pi_s  0.152
global pi_w   0.156    ; global rho_s  (-0.013)
** Simulate data:
#4 Admin data could be mismatched.
   Survey data with RTM and contamination.
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
est sto m0
```

Example: Summary Statistics

```
. 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
```

Example: KY 2007



Example: Model estimations

```
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
```

Example: Model estimations

	(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	

Example: Model PostEstimation

```
. 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
```

Example: Model PostEstimation

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

            Rel1      Rel2       MSE    E(Bias)  Var(Bias)
r_var     0.4885    0.4793    0.4937    -0.1099     0.4821
s_var     0.7601    0.7476    0.1539    -0.0944     0.1451
  e_1     0.8772    0.7879    0.1014    -0.0275     0.1007 Wgt unc
  e_2     0.7837    0.7854    0.1214    -0.0011     0.1215 Wgt unc unbi
  e_3     1.0154    0.8229    0.0784    -0.0036     0.0784 Wgt con
  e_4     0.7858    0.7764    0.1246    -0.0058     0.1246 Wgt con unb
  e_5     0.8896    0.7860    0.1000    -0.0006     0.1000 2-step
  e_6     0.7673    0.7615    0.1372    -0.0237     0.1367 2-step unb
  e_7     0.9915    0.7476    0.1118     0.0012     0.1119 Sys-wide
```

Conclusions

- We introduce a new set of commands to facilitate estimation of FMMs for application to linked survey and administrative data on earnings or similar variables.
  `ssc install ky_fit`
- The FMM specifications are those proposed by Jenkins and Rios-Avila (2021b) that extend the ones proposed by KY.
- We also provide a suite of post-estimation commands for simulation, assessing reliability, and deriving highly-reliable hybrid earnings predictors of latent true earnings.

Thank you! Questions or comments?

friosavi@levy.org
Or
friosa@gmail.com

References

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, forthcoming.

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.

References

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 Economic Statistics, 30 (2): 191–201.