Statistical Tools for Data Integration

John M. Abowd
April 2005

Outline

- Estimating Models from Linked Data
- Building Models with Heterogeneity for Linked Data
- Fixed Effect Estimation
- Identification
- Calculation
- Mixed Effect Estimation

Estimating Models from Linked Files

- Linked files are usually analyzed as if the linkage were without error
- Most of this class focuses on such methods
- There are good reasons to believe that this assumption should be examined more closely

Lahiri and Larsen

- Consider regression analysis when the data are imperfectly linked
- See JASA article March 2005 for full discussion

Setup of Lahiri and Larsen

$y_i = x_i\beta + \varepsilon_i$ where x_i is $(1\times p)$ and $i = 1,\ldots,n$

$y = X\beta + \varepsilon$ is the vector version (standard linear model)

$E\left[\varepsilon \middle| X\right] = 0, V\left[\varepsilon \middle| X\right] = \sigma^2 I$

Model for the matching error

$$z_i = \begin{cases} y_i \text{ w/ prob. } q_{ii} \\ y_j \text{ w/ prob. } q_{ij} \text{ for } j \neq i \text{ and } j = 1,\ldots,n \end{cases}$$

where $\sum_{j=1}^{n} q_{ij} = 1$

$q_i = (q_{i1},\ldots,q_{in})'$ and $Q = (q_1,\ldots,q_n)$

$w_i = q_i' X$ and $W = (w_1,\ldots,w_n)'$

Estimators

$$\hat{\beta}_N = \left(X'X\right)^{-1} X'z \text{ naive estimator}$$

$$\hat{\beta}_{SW} = \hat{\beta}_N - \left(X'X\right)^{-1} X'\hat{B}$$

$$B_i = \left(q_{ii} - 1\right)y_i + \sum_{j \neq i} q_{ij} y_j = q_i' y - y_i$$

$$\hat{\beta}_U = \left(W'W\right)^{-1} W'z$$

Problem: Estimating B

To estimate B one needs estimates of the q_{ij}

Fortunately, we have the Fellegi-Sunter model to use

Technique 1: estimate q_{ij} using the EM algorithm

(see lecture 10a)

Technique 2: estimate q_{ij} using mixture models

(see Larsen and Rubin JASA 2001)

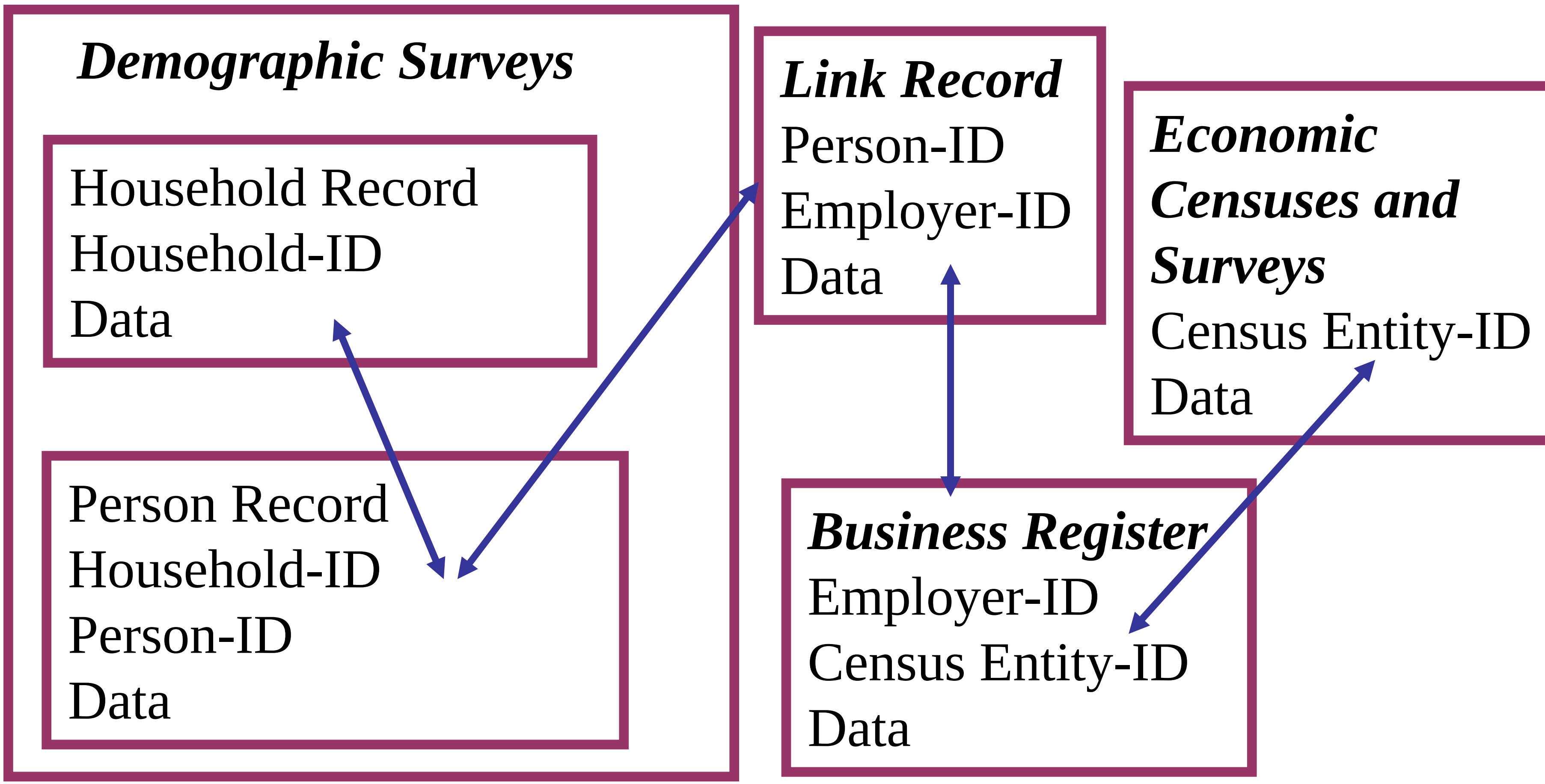
Does It Matter?

- Yes
- The bias from the naïve estimator is very large as the average q_{ii} goes away from 1.
- The SW estimator does better.
- The U estimator does very well, at least in simulations.

Building Linked Data

- Examples from the LEHD infrastructure files
- Analysis can be done using workers, jobs or employers as the basic observation unit
- Want to model heterogeneity due to the workers and employers for job level analyses
- Want to model heterogeneity due to the jobs and workers for employer level analyses
- Want to model heterogeneity due to the jobs and employers for individual analyses

The Longitudinal Employer - Household Dynamics Program



Basic model

$$y_{it} - \mu_y = \left(x_{it} - \mu_x\right)\beta + \theta_i + \psi_{\mathrm{J}(i,t)it} + \varepsilon_{it}$$

- The dependent variable is some individual level outcome, usually the log wage rate.
- The function J(*i,t*) indicates the employer of *i* at date *t*.
- The first component is the measured characteristics effect.
- The second component is the person effect.
- The third component is the firm effect.
- The fourth component is the statistical residual, orthogonal to all other effects in the model.

Matrix Notation: Basic Statistical Model

$$y = X\beta + D\theta + F\psi + \varepsilon$$

- All vectors/matrices have row dimensionality equal to the total number of observations.
- Data are sorted by person-ID and ordered chronologically for each person.
- *D* is the design matrix for the person effect: columns equal to the number of unique person IDs plus columns of u_i.
- *F* is the design matrix for the firm effect: columns equal to the number of unique firm IDs times the number of effects per firm.

Estimation by Fixed-effect Methods

- The normal equations for least squares estimation of fixed person, firm and characteristic effects are very high dimension.
- Estimation of the full model by either fixed-effect or mixed-effect methods requires special algorithms to deal with the high dimensionality of the problem.

Least Squares Normal Equations

$$\begin{bmatrix} X'X & X'D & X'F \\ D'X & D'D & D'F \\ F'X & F'D & F'F \end{bmatrix} \begin{bmatrix} \beta \\ \theta \\ \psi \end{bmatrix} = \begin{bmatrix} X'y \\ D'y \\ F'y \end{bmatrix}$$

- The full least squares solution to the basic estimation problem solves these normal equations for all identified effects.

Identification of Effects

- Use of the decomposition formula for the industry (or firm-size) effect requires a solution for the identified person, firm and characteristic effects.
- The usual technique of eliminating singular row/column combinations from the normal equations won't work if the least squares problem is solved directly.

Identification by Grouping

- Firm 1 is in group g = 1.
- Repeat until no more persons or firms are added:
  - Add all persons employed by a firm in group 1 to group 1
  - Add all firms that have employed a person in group 1 to group 1
- For g= 2, ..., repeat until no firms remain:
  - The first firm not assigned to a group is in group g.
  - Repeat until no more firms or persons are added to group g:
    - Add all persons employed by a firm in group g to group g.
    - Add all firms that have employed a person in group g to group g.
- Identification of ψ: drop one firm from each group g.
- Identification of θ: impose one linear restriction

$$\sum_{\forall(i,t)} \theta_i = 0$$

Normal Equations after Group Blocking

$$\left[\begin{array}{c|cc|cc|c|cc} X'X & X'D_1 & X'F_1 & X'D_2 & X'F_2 & \cdots & X'D_G & X'F_G \\ \hline D_1'X & D_1'D_1 & D_1'F_1 & 0 & 0 & \cdots & 0 & 0 \\ F_1'X & F_1'D_1 & F_1'F_1 & 0 & 0 & \cdots & 0 & 0 \\ \hline D_2'X & 0 & 0 & D_2'D_2 & D_2'F_2 & \cdots & 0 & 0 \\ F_2'X & 0 & 0 & F_2'D_2 & F_2'F_2 & \cdots & 0 & 0 \\ \hline \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \hline D_G'X & 0 & 0 & 0 & 0 & \cdots & D_G'D_G & D_G'F_G \\ F_G'X & 0 & 0 & 0 & 0 & \cdots & F_G'D_G & F_G'F_G \end{array}\right] \left[\begin{array}{c} \beta \\ \hline \theta_1 \\ \psi_1 \\ \hline \theta_2 \\ \psi_2 \\ \hline \cdots \\ \hline \theta_G \\ \psi_G \end{array}\right] = \left[\begin{array}{c} X'y \\ \hline D_1'y \\ F_1'y \\ \hline D_2'y \\ F_2'y \\ \hline \cdots \\ \hline D_G'y \\ F_G'y \end{array}\right]$$

- The normal equations have a sub-matrix with block diagonal components.
- This matrix is of full rank and the solution for (β, θ, ψ) is unique.

Necessity of Identification Conditions

- For necessity, we want to show that exactly *N+J-G* person and firm effects are identified (estimable), including the grand mean μ_y. .
- Because *X* and *y* are expressed in deviations from the mean, all N effects are included in the equation but one is redundant because both sides of the equation have a zero mean by construction.
- So the grand mean plus the person effects constitute *N* effects.
- There are at most *N + J-1* person and firm effects including the grand mean.
- The grouping conditions imply that at most *G* group means are identified (or, the grand mean plus G-1 group deviations).
- Within each group *g*, at most N_g and J_g-1 person and firm effects are identified.
- Thus the maximum number of identifiable person and firm effects is:

$$N + J - G = \sum_g \left(N_g + J_g - 1 \right)$$

Sufficiency of Identification Conditions

- For sufficiency, we use an induction proof.
- Consider an economy with *J* firms and *N* workers.
- Denote by E[y_{it}] the projection of worker *i*'s wage at date *t* on the column space generated by the person and firm identifiers. For simplicity, suppress the effects of observable variables *X*

$$\mathrm{E}\left[y_{it}\right] = \mu_y + \theta_i + \psi_{\mathrm{J}(i,t)}$$

- The firms are connected into *G* groups, then all effects ψ_j, in group g are separately identified up to a constraint of the form:

$$\sum_{j \in \{\text{group } g\}} w_j \psi_j = 0$$

Sufficiency of Identification Conditions II

- Suppose *G*=1 and *J*=2.
- Then, by the grouping condition, at least one person, say 1, is employed by both firms and we have

$$w_1\psi_1 + w_2\psi_2 = 0$$

$$\mathrm{E}\left\lfloor y_{1t_1} \right\rfloor - \mathrm{E}\left\lfloor y_{1t_2} \right\rfloor = \psi_1 - \psi_2$$

- So, exactly N+2-1 effects are identified.

Sufficiency of Identification Conditions III

- Next, suppose there is a connected group g with J_g firms and exactly J_g -1 firm effects identified.
- Consider the addition of one more connected firm to such a group.
- Because the new firm is connected to the existing J_g firms in the group there exists at least one individual, say worker 1 who works for a firm in the identified group, say firm J_g, at date 1 and for the supplementary firm at date 2. Then, we have two relations

$$\sum_{g \le J_g} w_g \psi_g + w_{J_g+1} \psi_{J_g+1} = 0$$

$$\mathrm{E}\left\lfloor y_{1t_1} \right\rfloor - \mathrm{E}\left\lfloor y_{1t_2} \right\rfloor = \psi_{J_g} - \psi_{J_g+1}$$

- So, exactly J_g effects are identified with the new information.

Estimation by Direct Solution of Least Squares

- Once the grouping algorithm has identified all estimable effects, we solve for the least squares estimates by direct minimization of the sum of squared residuals.
- This method, widely used in animal breeding and genetics research, produces a unique solution for all estimable effects.

Least Squares Conjugate Gradient Algorithm

- The matrix Δ is chosen to precondition the normal equations.
- The data matrices and parameter vectors are redefined as shown.

$$\Delta = \text{diagonal elements of} \begin{bmatrix} X'X & X'D & X'F \\ D'X & D'D & D'F \\ F'X & F'D & F'F \end{bmatrix}$$

$$y = \left[X \mid D \mid F\right]\Delta^{-1/2}\Delta^{1/2}\begin{bmatrix} \beta \\ \theta \\ \psi \end{bmatrix} + \varepsilon \equiv Z\delta + \varepsilon$$

$$Z \equiv \left[X \mid D \mid F\right]\Delta^{-1/2} \text{ and } \delta \equiv \Delta^{1/2}\begin{bmatrix} \beta \\ \theta \\ \psi \end{bmatrix}$$

LSCG (II)

- The goal is to find δ to solve the least squares problem shown.
- The gradient vector *g* figures prominently in the equations.
- The initial conditions for the algorithm are shown.
  - *e* is the vector of residuals.
  - *d* is the direction of the search.

$$\hat{\delta} = \arg\min_{\delta}\left[\left(y - Z\delta\right)'\left(y - Z\delta\right)\right]$$

$$0 = \frac{1}{2}\frac{\partial\left(y - Z\delta\right)'\left(y - Z\delta\right)}{\partial\delta} = Z'\left(y - Z\delta\right) \equiv g$$

$$\tau_{-1} = 0$$
$$d_{-1} = 0$$
$$\delta_0 = 0$$
$$e_0 = y - Z\delta_0$$
$$g_0 = Z'e_0 = Z'y - Z'Z\delta_0$$
$$d_0 = g_0$$
$$\rho_0 = g_0'g_0$$
$$\lambda_0 = 0$$

LSCG (III)

- The loop shown has the following features:
  - The search direction *d* is the current gradient plus a fraction of the old direction.
  - The parameter vector δ is updated by moving a positive amount in the current direction.
  - The gradient, *g*, and residuals, *e*, are updated.
  - The original parameters are recovered from the preconditioning matrix.

For $\ell = 0,1,2,3,\dots$

$$d_\ell = g_\ell + \tau_{\ell-1} d_{\ell-1}$$

$$q_\ell = Z d_\ell$$

$$\lambda_\ell = \rho_\ell / \left(q_\ell' q_\ell\right)$$

$$\delta_{\ell+1} = \delta_\ell + \lambda_\ell d_\ell$$

$$e_{\ell+1} = e_\ell - \lambda_\ell q_\ell$$

$$g_{\ell+1} = Z' e_{\ell+1}$$

$$\begin{bmatrix} \beta_{\ell+1} \\ \theta_{\ell+1} \\ \psi_{\ell+1} \end{bmatrix} = \Delta^{-1/2} \delta_{\ell+1}$$

LSCG (IV)

- Verify that the residuals are uncorrelated with the three components of the model.
  - Yes: the LS estimates are calculated as shown.
  - No: certain constants in the loop are updated and the next parameter vector is calculated.

$$\begin{bmatrix} X\beta_{\ell+1} & D\theta_{\ell+1} & F\psi_{\ell+1} \end{bmatrix}' e_{\ell+1} \begin{cases} < \begin{bmatrix} c \\ c \\ c \end{bmatrix}, \text{stop } \hat{\delta} = \delta_{\ell+1} \\ \text{else, continue} \end{cases}$$

$$\rho_{\ell+1} = \left(g_{\ell+1}' g_{\ell+1} \right)$$

$$\tau_{\ell} = \rho_{\ell+1} / \rho_{\ell}$$

$$\begin{bmatrix} \hat{\beta} \\ \hat{\theta} \\ \hat{\psi} \end{bmatrix} = \Delta^{-1/2} \hat{\delta}$$

$$S = \left(y - Z\hat{\delta} \right)' \left(y - Z\hat{\delta} \right)$$

Mixed Effects Assumptions

$$\Lambda = \begin{bmatrix} \Sigma_1 & 0 & \cdots & 0 \\ 0 & \Sigma_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \Sigma_N \end{bmatrix} \qquad \mathrm{E}\left[\begin{matrix} \theta \\ \psi \end{matrix} \middle| X \right] = 0 \qquad \mathrm{V}\left[\begin{matrix} \theta \\ \psi \end{matrix} \middle| X \right] = \Omega$$

- The assumptions above specify the complete error structure with the firm and person effects random.
- For maximum likelihood or restricted maximum likelihood estimation assume joint normality.

Estimation by Mixed Effects Methods

$$\begin{bmatrix} X'\Lambda^{-1}X & X'\Lambda^{-1}\left[D \mid F\right] \\ \begin{bmatrix} D' \\ -- \\ F' \end{bmatrix}\Lambda^{-1}X & \begin{bmatrix} D' \\ -- \\ F' \end{bmatrix}\Lambda^{-1}\left[D \mid F\right] + \Omega^{-1} \end{bmatrix} \begin{bmatrix} \beta \\ \theta \\ \psi \end{bmatrix} = \begin{bmatrix} X'\Lambda^{-1}y \\ \begin{bmatrix} D' \\ -- \\ F' \end{bmatrix}\Lambda^{-1}y \end{bmatrix}$$

- Solve the mixed effects equations
- Techniques: Bayesian EM, Restricted ML

Relation Between Fixed and Mixed Effects Models

$$\Lambda = \sigma_{\varepsilon}^{2} I_{N*} \qquad \left|\Omega\right| \to \infty$$

- Under the conditions shown above, the ME and estimators of all parameters approaches the FE estimator

Correlated Random Effects vs. Orthogonal Design

$$X'D = 0 \text{ orthogonal personal characteristics and person-effect design}$$
$$X'F = 0 \text{ orthogonal personal characteristics and firm-effect design}$$
$$D'F = 0 \text{ orthogonal person-effect and firm-effect designs}$$

- Orthogonal design means that characteristics, person design, firm design are orthogonal.
- Uncorrelated random effects means that Ω is diagonal.