INFO 7470/ILRLE 7400
Statistical Tools:
Basic Integrated Data Models

John M. Abowd and Lars Vilhuber

April 12, 2011

Outline

- What Are “Linked” or “Integrated” Data?
- The Relational Database Model
- Statistical Underpinnings of the Relational Database Model
- Graphical Representations of Integrated Data
- Estimating Models from Linked Data

What Are Linked Data?

- Observations used in the analysis are sampled from different universes of entities
- Observations from the different entities relate to each other according to a system of identifiers
- Integration of the observations requires specifying a universe for the result and a rule for associating data from entities belonging to other universes with the observations in the result.

Examples of Linked Data

- Hierarchies
  - Population census: block-household-resident
  - Economic census: enterprise-establishment
- Relations
  - Person-job-employer
  - Customer-item-supplier

The Relational Database Model

- Informal characterization (formal model , 2007 link)
- All data are represented as a collection of linked tables
- Each table has a unique key (primary key) that is defined for every entity in the table
- Each table may have data items defined for each entity in the table
- Each table may have items that refer to data from another table (foreign key)
- Views are created by specifying a reference table and gathering the values of data items based on the keys in the reference table and operations applied to the items retrieved by the foreign keys

Example of the Relational Database Model

- Table_Employer
  - Primary_key: Employer_ID
  - Foreign_key: NAICS
  - Items: Sales, Employees
- Table_Individual
  - Primary_key: Individual_ID
  - Foreign_key: Census_block
  - Items: Age, Education
- Table_Job
  - Primary_key: Job_ID
  - Foreign_key: Employer_ID
  - Foreign_key: Individual_ID
  - Items: Earnings
- Table_Industry
  - Primary_key: NAICS
  - Items: average_earnings

Example: Job View

- Select records from Table_Job (universe or sample)
- Look-up Sales, Employees, NAICS in Table_Employer using Employer_ID; compute sales_per_employee
- Look-up NAICS in Table_Industry; compute log_industry_average_earnings
- Look-up Age and Education in Table_Individual using Individual_ID; compute potential_experience
- Compute log_earnings
- Create Table_Output
  - Primary_key: Job_ID
  - Items: log_earnings, education, potential_experience, sales_per_employee, log_industry_average_earnings

Graphical Representation of Linked Data

- Graphs:
  - Nodes: list of entities
  - Edges: ordered (directed) or unordered (non-directed) pairs indicating a link between two nodes
- Example
  - Nodes: {Employer_IDs, Individual_IDs}
  - Edges: Ordered pairs (Individual_ID ‘works for’ Employer_ID)

Statistical Underpinnings of the Relational Database Model

- Tables are frames
- If every table is complete relative to its universe, then samples can be constructed by sampling records from the relevant table and linking data from the other tables
- If some tables are incomplete, then imputation of missing data is equivalent to imputing a link and estimating its items

Example: Industry View

- Select records from Table_Industry
- Look-up all Employer_IDs in NAICS in Table_Employer; compute variance_earnings
- Ouput Table_Output
  - Primary_key: NAICS
  - Item average_earnings, variance_earnings

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)

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

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

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

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

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

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

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.

INFO 7470/ILRLE 7400
Statistical Tools:
Graph-based Data Models

John M. Abowd and Lars Vilhuber

April 12, 2011

Outline

- Basic Graph Theory
- Integrated Labor Market Data
- Statistical Modeling
- Graph Theoretic Identification
- Estimation by Fixed-effects Methods
- Estimation by Mixed-effects Methods

BASIC GRAPH THEORY

What Is A Graph?

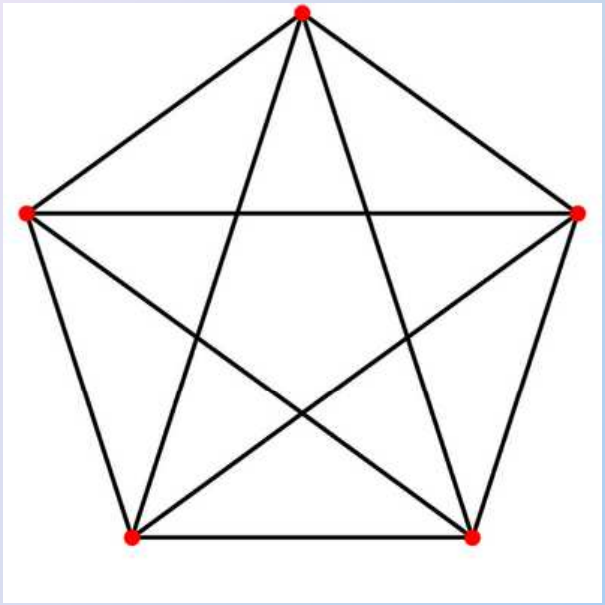
$$G = \{V^*, E^*\}$$

$$V = \{v_1, \ldots, v_N\}$$

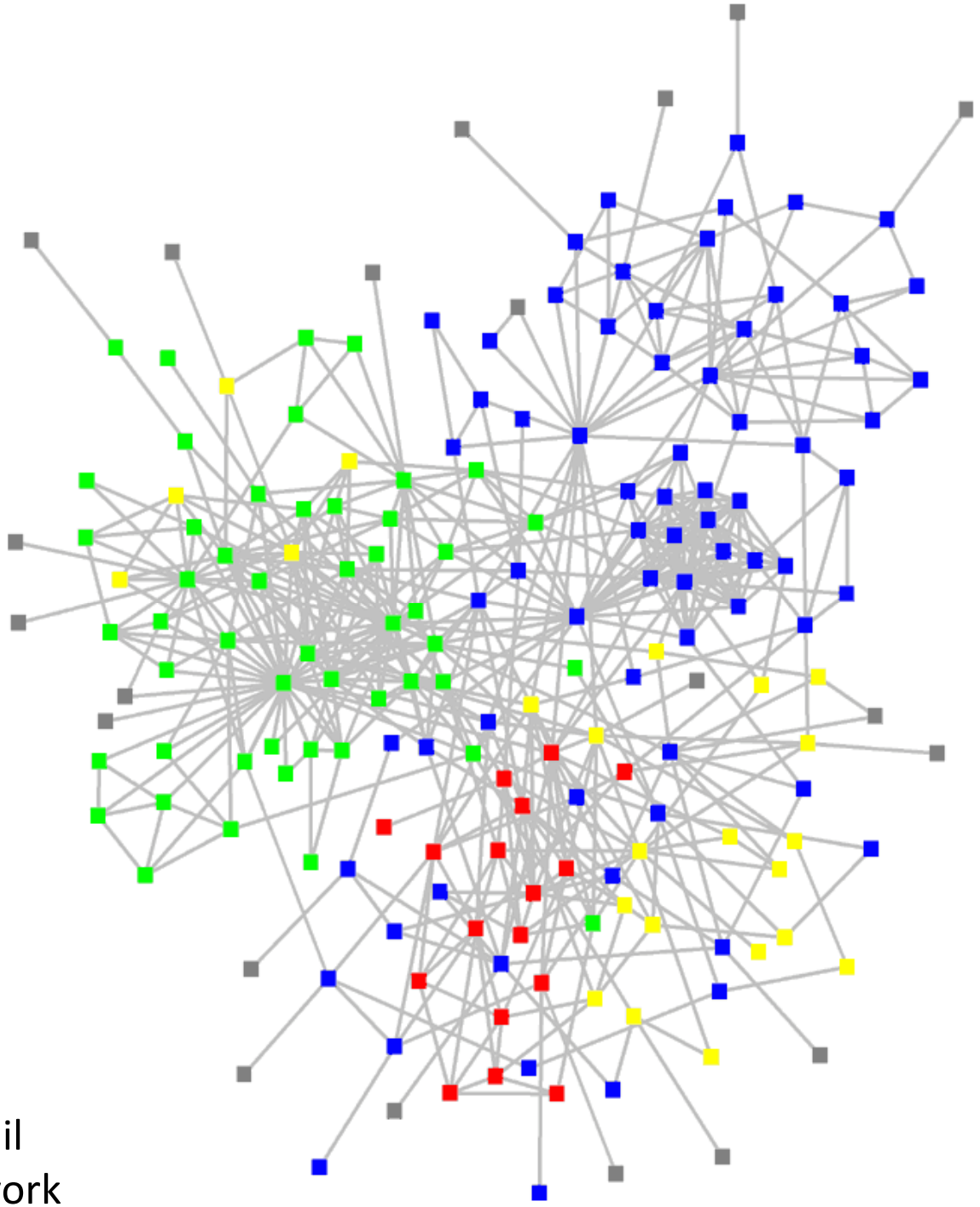
$$E = \{(i, j) \mid i, j \in \{1, \ldots, N\} \wedge i \neq j\}$$

$$V^* \subseteq V$$

$$E^* \subseteq E$$



Fully connected graph



E-mail network

The Bipartite Labor Market Graph

$$G = \{V^*, E^*\}$$

$$V^I = \{v_1, \ldots, v_I\}; V^F = \{u_1, \ldots, u_J\}$$

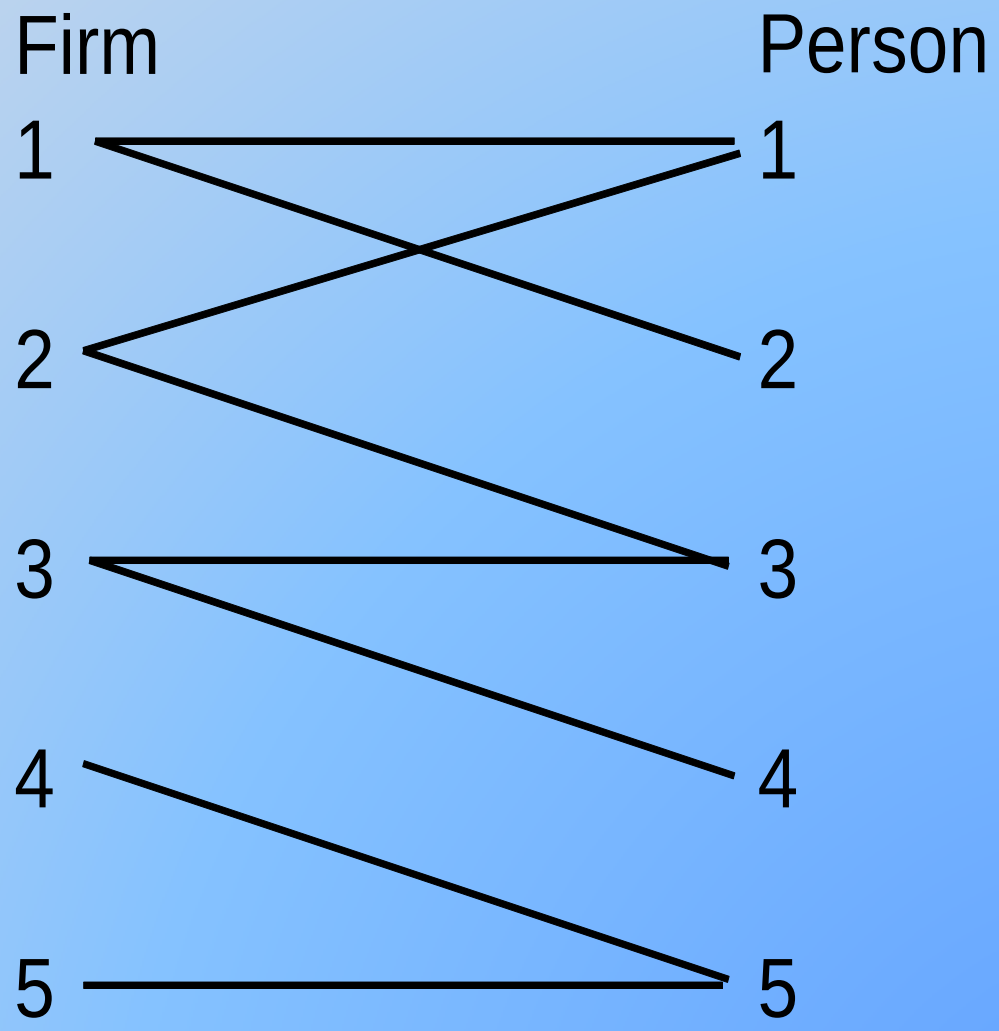
$$V = V^I \cup V^F$$

$$E = \{(i, j) | i \in \{1, \ldots, I\} \wedge j \in \{1, \ldots, J\}\}$$

$$V^* \subseteq V; E^* \subseteq E$$

Labor Market Graph

Firm	Person
1	1
1	2
2	1
2	3
3	3
3	4
4	5
5	5



Adjacency Matrices

$$X = \left[x_{ij} = \begin{cases} 1, \text{if } (i, j) \vee (j, i) \in E^* \\ 0, \text{otherwise} \end{cases} \right]$$

$$B = \left[b_{ij} = \begin{cases} 1, \text{if } (i, j) \in E^* \wedge G \text{ bipartite} \\ 0, \text{otherwise} \end{cases} \right]$$

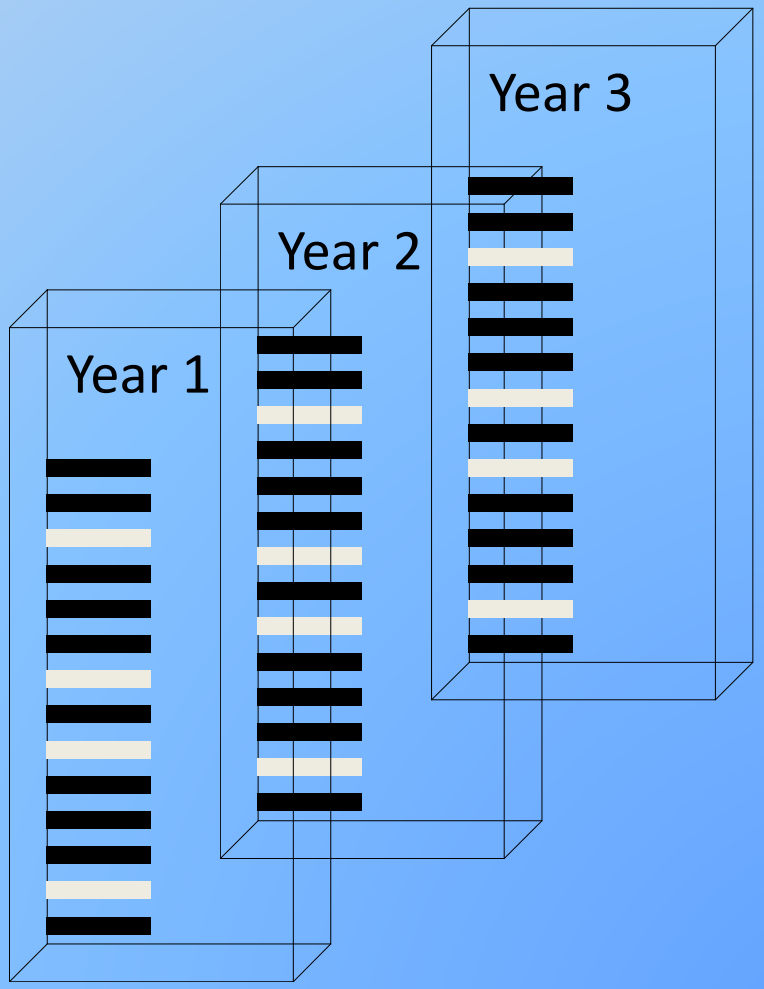
INTEGRATED LABOR MARKET DATA

Longitudinal Integrated Data

- Matched longitudinal data on employers and employees allows us to identify the separate effects of unobserved personal and firm heterogeneity.
- Large samples are required with simple random sampling because the identification is based on within sample mobility of workers between firms.
- For multi-stage sample surveys, geographic clustering permits identification of employer effects because the within sample employment mobility is generally within the same geographic area.

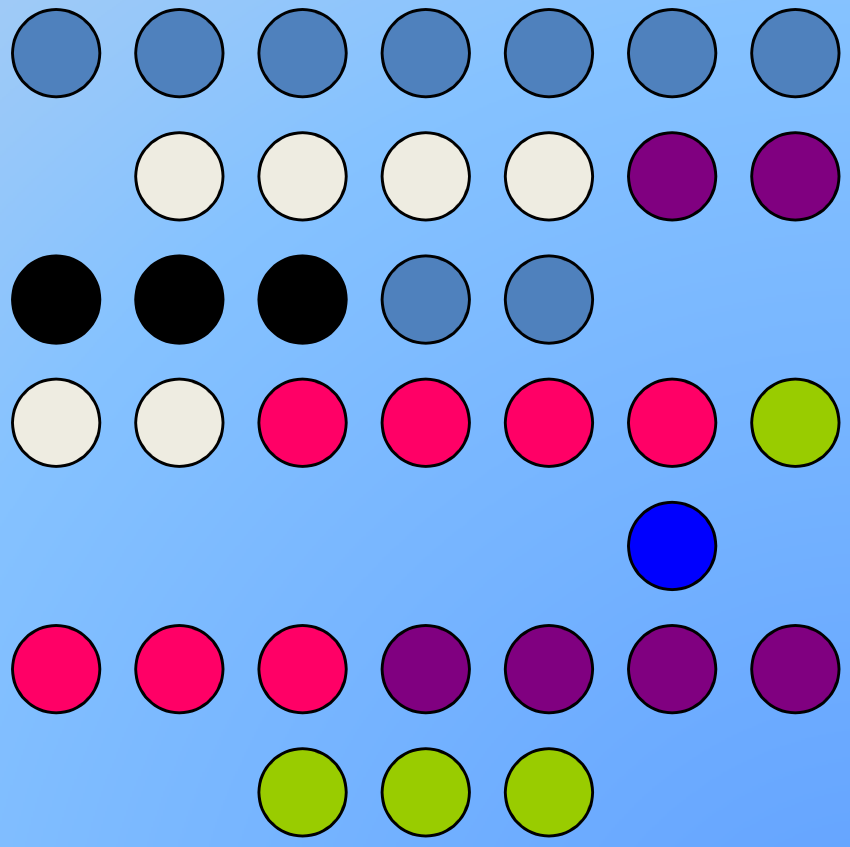
Longitudinal Sampling Plan for Individuals

- Begin with an administrative file of individuals
  - Sample individuals randomly
  - Select the same individuals in successive years



Structure of Resulting Data

- Continuously employed at the same employer
- Entry and a change of employers
- Exit and a change of employer
- Continuously employed with multiple employers
- ID errors
- Other sequences
- Example shows 7 person IDs and 7 different employers



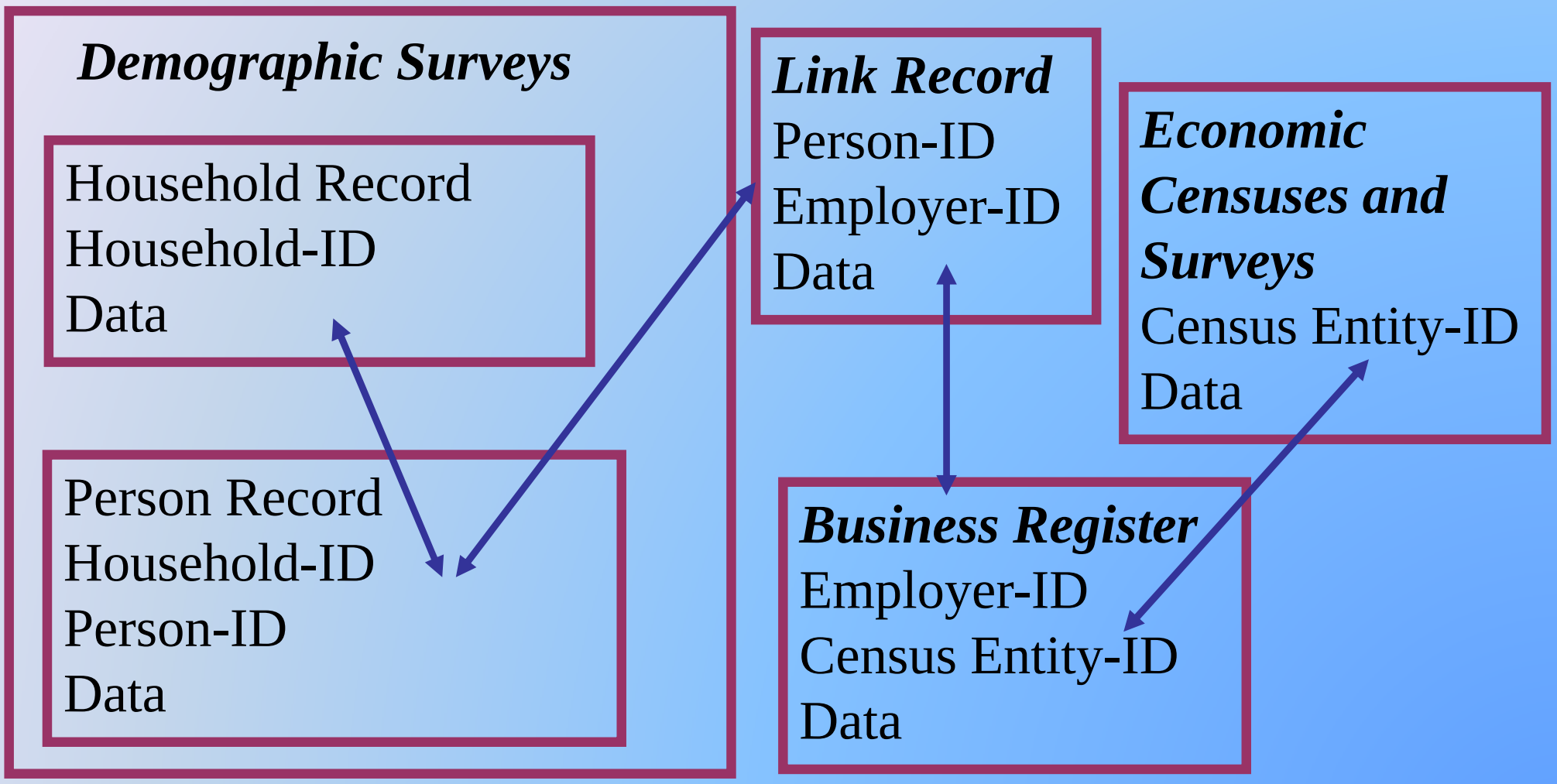
Other Sampling Plans

- Sample firms with probability proportional to employment; sample employees randomly from selected firms
- As above plus all employers every employing selected individuals
- As in first bullet, plus subsequent employers of selected individuals

Building Integrated Labor Market 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



STATISTICAL MODELING

Basic Statistical Model

$$y_{it} = \theta_i + \psi_{J(i,t)} + x_{it}\beta + \varepsilon_{it}$$

- The dependent variable is compensation
- The function J(i,t) indicates the employer of i at date t.
- The first component is the person effect.
- The second component is the firm effect.
- The third component is the measured characteristics effect.
- The fourth component is the statistical residual, orthogonal to all other effects in the model.

Matrix Notation: Basic Model

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

- 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.
- *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.

True Industry Effect Model

$$y_{it} = \theta_i + (\psi_{\mathrm{J}(i,t)} - \kappa_{\mathrm{K}(\mathrm{J}(i,t))}) + \kappa_{\mathrm{K}(\mathrm{J}(i,t))} + x_{it}\beta + \varepsilon_{it}$$

- The function K(*j*) indicates the industry of firm *j*.
- The first component is the person effect.
- The second component is the firm effect net of the true industry effect.
- The third component is the true industry effect, an aggregation of firm effects since industry is a property of the employer.
- The fourth component is the effect of personal characteristics
- The fifth component is the statistical residual.

Matrix Notation: True Industry Effect Model

$$y = D\theta + FA\kappa + (F\psi - FA\kappa) + X\beta + \varepsilon$$

- The matrix *A* is the classification matrix taking firms into industries.
- The matrix *FA* is the design matrix for the true industry effect.
- The true industry effect κ can be expressed as

$$\kappa = (A'F'FA)^{-1} A'F'F\psi$$

Raw Industry Effect Model

$$y_{it} = \kappa^{**}_{\mathrm{K}(i,t)} + x_{it}\beta^{**} + \varepsilon_{it}$$

- The first component is the raw industry effect.
- The second component is the measured personal characteristics effect.
- The third component is the statistical residual.
- The raw industry effect is an aggregation of the appropriately weighted average person and average firm effects within the industry, since both have been excluded from the model.
- The true industry effect is only an aggregation of the appropriately weighted average firm effect within the industry, as shown above.

Industry Effects Adjusted for Person Effects Model

$$y_{it} = \kappa^{*}_{\mathrm{K}(i,t)} + \theta^{*}_{i} + x_{it}\beta^{*} + \varepsilon_{it}$$

- The first component is the industry effect adjusted for person effects.
- The second component is individual effect (with firm effects omitted)
- The third component is the measured personal characteristics effect.
- The fourth component is the statistical residual.
- The industry effects adjusted for person effects are also biased.

Relation: True and Raw Industry Effects

$$\kappa^{**} = \kappa + (A'F'M_X FA)^{-1} A'F'M_X (M_{FA} F\psi + D\theta)$$

- The vector κ^{**} of industry effects can be expressed as the true industry effect κ plus a bias that depends upon both the person and firm effects.
- The matrix *M* is the residual matrix (column null space) after projection onto the column space of the matrix in the subscript. For example,

$$M_X \equiv I - X(X'X)^{-1}X'$$

Relation: Industry, Person and Firm Effects

$$\kappa^{**} = (A'F'M_X FA)^{-1} A'F'M_X D\theta + (A'F'M_X FA)^{-1} A'F'M_X F\psi$$

- The vector κ^{**} of raw industry effects can be expressed as a matrix weighted average of the person effects θ and the firm effects ψ.
- The matrix weights are related to the personal characteristics *X*, and the design matrices for the person and firm effects (see Abowd, Kramarz and Margolis, 1999).

IDENTIFICATION

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.

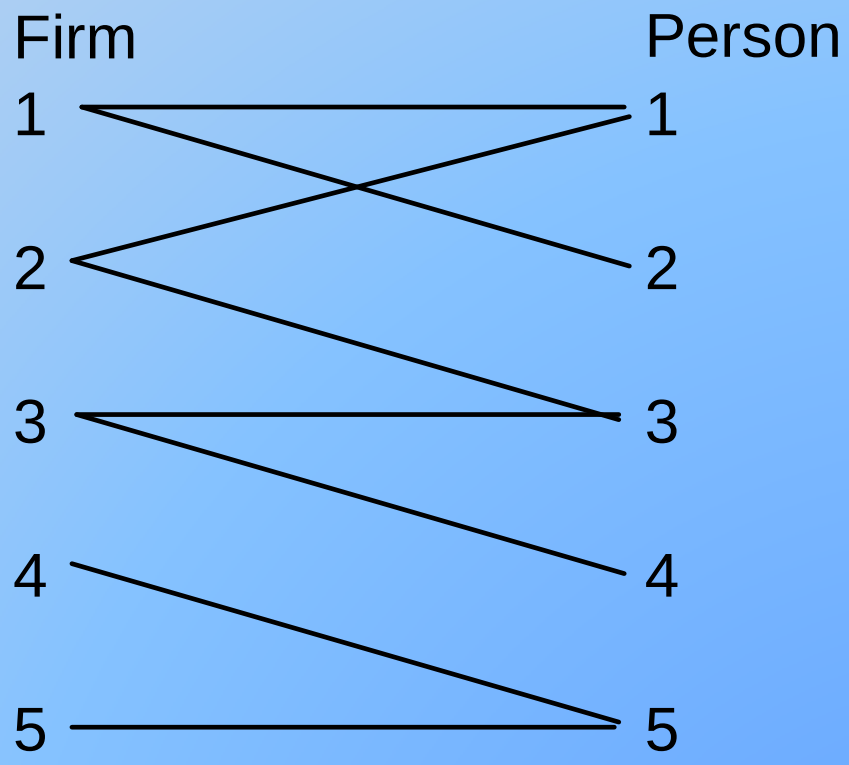
Identification by Finding connected Sub-graphs

- 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.
- Each group g is a connected subgraph
- Identification of ψ: drop one firm from each group g
- Identification of θ: impose one linear restriction

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

Connected Sub-graphs of the Labor Market

Firm	Person	Group
1	1	1
1	2	1
2	1	1
2	3	1
3	3	1
3	4	1
4	5	2
5	5	2



Normal Equations after Group Blocking

$$\begin{bmatrix} X'X & X'D_1 & X'F_1 & X'D_2 & X'F_2 & \cdots & X'D_G & X'F_G \\ D_1'X & D_1'D_1 & D_1'F_1 & 0 & 0 & \cdots & 0 & 0 \\ F_1'X & F_1'D_1 & F_1'F_1 & 0 & 0 & \cdots & 0 & 0 \\ D_2'X & 0 & 0 & D_2'D_2 & D_2'F_2 & \cdots & 0 & 0 \\ F_2'X & 0 & 0 & F_2'D_2 & F_2'F_2 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ D_G'X & 0 & 0 & 0 & 0 & \cdots & D_G'D_G & D_G'F_G \\ F_G'X & 0 & 0 & 0 & 0 & \cdots & F_G'D_G & F_G'F_G \end{bmatrix} \begin{bmatrix} \beta \\ \theta_1 \\ \psi_1 \\ \theta_2 \\ \psi_2 \\ \cdots \\ \theta_G \\ \psi_G \end{bmatrix} = \begin{bmatrix} X'y \\ D_1'y \\ F_1'y \\ D_2'y \\ F_2'y \\ \cdots \\ D_G'y \\ F_G'y \end{bmatrix}$$

- 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 $\mathrm{E}[y_{it}]$ 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*

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

- 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:

$$\sum_{j \in \{g\mathrm{roup}g\}} w_j \psi_j = 0$$

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

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

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

- 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

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

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

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

ESTIMATION BY FIXED-EFFECTS METHODS

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

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

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

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

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

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.

$$\hat{\delta} = \underset{\delta}{\operatorname{argmin}}\left[(y - Z\delta)'(y - Z\delta)\right]$$

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

$$\tau_{-1} = 0$$

$$d_{-1} = 0$$

$$\delta_0 = 0$$

$$e_0 = y - Z\delta_0$$

$$g_0 = Z'e_0 = Z'y - Z'Z\delta_0$$

$$d_0 = g_0$$

$$\rho_0 = g_0'g_0$$

$$\lambda_0 = 0$$

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.

$$\text{For } \ell = 0,1,2,3,\ldots$$

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

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

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

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

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

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

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

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.

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

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

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

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

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

ESTIMATION BY MIXED-EFFECTS METHODS

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.
- Software: ASREML, cgmixed

Estimation by Mixed Effects Methods

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

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

Bayesian ECM

- The algorithm is illustrated for the special case of uncorrelated residuals and uncorrelated random effects.
- The initial conditions are taken directly from the LS solution to the fixed effects problem.

$$\Lambda = \sigma_{\varepsilon}^2 I_{N*}$$

$$\Omega = \begin{bmatrix} \sigma_{\theta}^2 I_N & 0 \\ 0 & \sigma_{\psi}^2 I_J \end{bmatrix}$$

$$\begin{bmatrix} \beta_0 \\ \theta_0 \\ \psi_0 \end{bmatrix} = \begin{bmatrix} \hat{\beta} \\ \hat{\theta} \\ \hat{\psi} \end{bmatrix}$$

$$\sigma_{\varepsilon 0}^2 = \frac{S}{N*}, \sigma_{\theta 0}^2 = \frac{\hat{\theta}'\hat{\theta}}{N}, \sigma_{\psi 0}^2 = \frac{\hat{\psi}'\hat{\psi}}{J}$$

Bayesian ECM (II)

- At each loop of the algorithm the E step is used to compute the parameters
- The conditional M step is used to update the variances.

$$\beta_{\ell+1} = \left(X'X\right)^{-1} X'\left(y - D\theta_{\ell} - F\psi_{\ell}\right)$$

$$\sigma^2_{\varepsilon\,\ell+1} = \left(y - X\beta_{\ell+1} - D\theta_{\ell} - F\psi_{\ell}\right)'(.)/N *$$

$$\theta_{i\,\ell+1} = \frac{\sigma^2_{\theta\,\ell} \sum_{t\in\{n_{i1}\ldots n_{iT_i}\}} \left(y_{it} - x_{it}\beta_{\ell+1} - \psi_{\mathrm{J}(i,t)\,\ell}\right)/T_i}{\sigma^2_{\theta\,\ell} + \sigma^2_{\varepsilon\,\ell+1}/T_i}$$

$$\psi_{j\,\ell+1} = \frac{\sigma^2_{\psi_\ell} \sum_{(i,t)\in\{\mathrm{J}(i,t)=j\}} \left(y_{it} - x_{it}\beta_{\ell+1} - \theta_{i\,\ell}\right)/N_j}{\sigma^2_{\psi_\ell} + \sigma^2_{\varepsilon\,\ell+1}/N_j}$$

$$\sigma^2_{\theta\,\ell+1} = \frac{\hat{\theta}_\ell'\hat{\theta}_\ell}{N}, \sigma^2_{\psi_{\ell+1}} = \frac{\hat{\psi}_\ell'\hat{\psi}_\ell}{J}$$

Relation Between Fixed and Mixed Effects Models

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

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

Correlated Random Effects vs. Orthogonal Design

$X'D = 0$ orthogonal personal characteristics and person-effect design

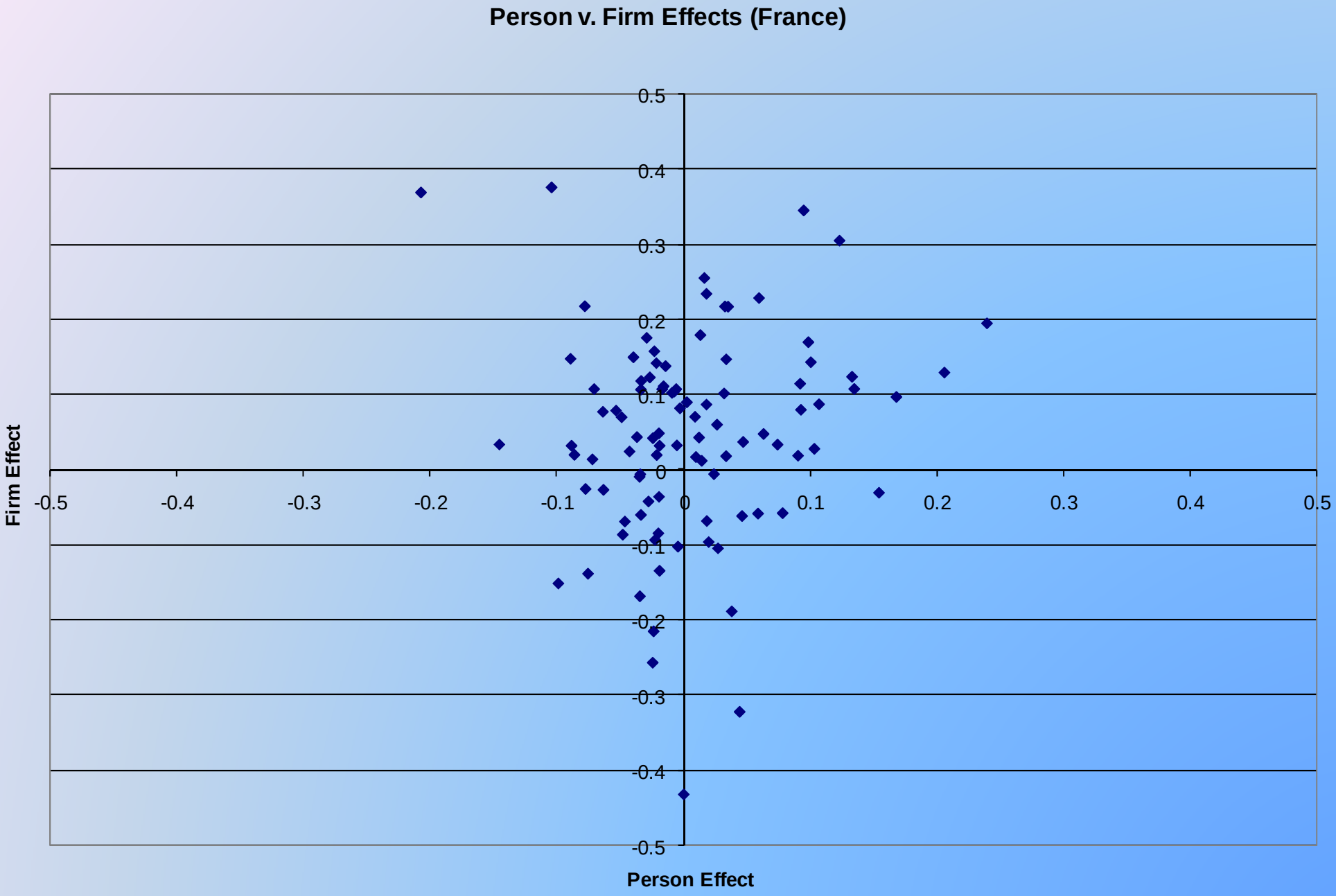
$X'F = 0$ orthogonal personal characteristics and firm-effect design

$D'F = 0$ 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

Software

- SAS: proc mixed
- ASREML
- aML
- SPSS: Linear Mixed Models
- STATA: xtreg, gllamm, xtmixed
- R: the lme() function
- S+: linear mixed models
- Matlab
- Genstat: REML
- Grouping (connected sub-graphs)
- Custom software on the VRDC


Person v. Firm Effects (France)
Firm Effect
Person Effect
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5

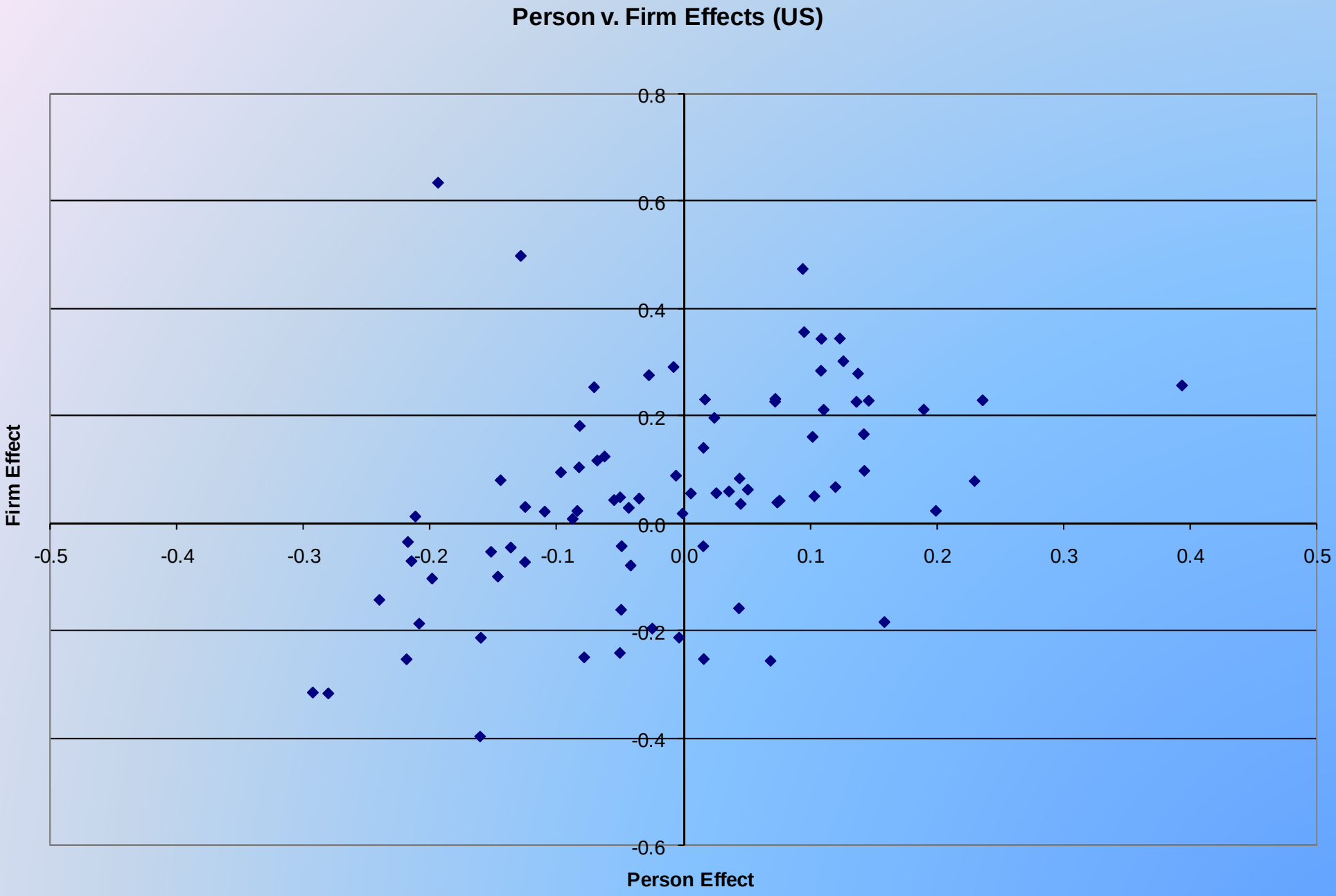

Person v. Firm Effects (US)
Firm Effect
Person Effect
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
0 0
0.1
0.2
0.3
0.4
0.5

Table 1: US, Winners and Losers

SIC	Industry	Raw Industry Wage Differential	Industry Average Person Effect	Industry Average Firm Effect
62	Security brokers, dealers, exchanges	0.659	0.393	0.258
46	Pipelines, except natural gas	0.625	0.094	0.475
29	Petroleum and coal products	0.478	0.109	0.345
81	Legal services	0.471	0.236	0.230
48	Communications	0.460	0.123	0.346
49	Electric, gas, and sanitary services	0.445	0.095	0.358
13	Oil and gas extraction	0.437	0.108	0.285
38	Instruments and related products	0.411	0.137	0.280
89	Miscellaneous services	0.409	0.136	0.227
28	Chemicals and allied products	0.407	0.126	0.303
83	Social services	-0.296	-0.199	-0.102
53	General merchandise stores	-0.303	-0.050	-0.241
79	Amusement and recreation services	-0.312	-0.079	-0.249
72	Personal services	-0.371	-0.160	-0.213
70	Hotels, rooming houses, camps, and lodging	-0.385	-0.209	-0.186
23	Apparel and other textile products	-0.402	-0.240	-0.142
58	Eating and drinking places	-0.554	-0.161	-0.397
07	Agricultural services	-0.568	-0.280	-0.316
01	Agriculture-crops	-0.570	-0.293	-0.315
88	Private households	-0.643	-0.219	-0.252

Table 2: France, Winners and Losers

NAP	Industry	Raw Industry Wage Differential	Industry Average Person Effect	Industry Average Firm Effect
05	Crude petroleum and natural gas extraction	0.490	0.239	0.196
27	Office and accounting machines	0.472	0.094	0.346
72	Air transportation	0.466	0.123	0.306
76	Financial holding companies	0.378	0.205	0.130
42	Tobacco products manufacture	0.337	0.016	0.256
07	Distribution of Gas	0.319	0.059	0.229
33	Aircraft and parts manufacture	0.313	0.167	0.098
04	Coal mining	0.312	0.098	0.170
94	Health care, non-market	0.301	-0.105	0.376
17	Basic chemical manufacture	0.289	0.100	0.144
67	Hotels, motels, bars and restaurants	-0.167	-0.047	-0.068
62	Retail specialty and neighborhood food	-0.177	-0.048	-0.085
90	Public administration, general	-0.218	-0.035	-0.167
38	Bakery products	-0.233	-0.019	-0.133
87	Miscellaneous commercial services	-0.239	-0.076	-0.137
95	Social services, non-market	-0.241	-0.099	-0.150
96	Recreational, cultural, and sporting, non-market	-0.252	-0.024	-0.214
82	Commercial education services	-0.259	0.044	-0.321
97	Miscellaneous public services, non-market	-0.288	-0.025	-0.255
92	Teaching, non-market	-0.414	0.000	-0.431

Table 3: Correlation of Industry Effects

	France			***US***		
	Raw Industry Wage Differential	*Industry Average Person Effect*	*Industry Average Firm Effect*	*Raw Industry Wage Differential*	*Industry Average Person Effect*	*Industry Average Firm Effect*
France			***French Industry Weights***			
Raw Industry Wage Differential	1.0000	0.6110	0.9046	0.5783	0.4444	0.5689
Industry Average Person Effect	0.6110	1.0000	0.2549	0.4810	0.6337	0.2880
Industry Average Firm Effect	0.9046	0.2549	1.0000	0.3792	0.1655	0.4564
US			***US Industry Weights***			
Raw Industry Wage Differential	0.4914	0.3652	0.3630	1.0000	0.8167	0.9214
Industry Average Person Effect	0.2662	0.4631	0.0630	0.8167	1.0000	0.5382
Industry Average Firm Effect	0.5500	0.2024	0.5065	0.9214	0.9214	1.0000