Modeling Integrated Data

John M. Abowd
April 2005
Outline

• Application of the grouping algorithm
• Further discussion of the methods
• Discussion of software
Characteristics of the Groups

Table 4: Results of Applying the Grouping Algorithm to the Pooled Data Set

<table>
<thead>
<tr>
<th></th>
<th>Largest Group</th>
<th>Second Largest Group</th>
<th>Average of All Other Groups</th>
<th>Total of All Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observations</td>
<td>285,402,315</td>
<td>90</td>
<td>4.3</td>
<td>287,241,891</td>
</tr>
<tr>
<td>Persons</td>
<td>64,441,382</td>
<td>38</td>
<td>1.5</td>
<td>68,329,212</td>
</tr>
<tr>
<td>Firms</td>
<td>3,200,067</td>
<td>8</td>
<td>1.1</td>
<td>3,662,974</td>
</tr>
<tr>
<td>Groups</td>
<td>1</td>
<td>1</td>
<td>430,529</td>
<td>430,531</td>
</tr>
<tr>
<td>Estimable Effects</td>
<td>67,641,448</td>
<td>45</td>
<td></td>
<td>71,992,185</td>
</tr>
</tbody>
</table>

Notes: The "pooled" data are comprised of annual observations from California, Florida, Illinois, Maryland, Minnesota, North Carolina, and Texas over the period 1986-2000. No single state contributed observations for all years. See Table 1. Sources: Author's calculations using the LEHD Program data base.
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 $\Delta$ is chosen to precondition the normal equations.
- The data matrices and parameter vectors are redefined as shown.

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

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

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

- The goal is to find $\delta$ 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.

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

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

$$\tau_{-1} = 0$$
$$d_{-1} = 0$$
$$\delta_{0} = 0$$
$$e_{0} = y - Z\delta_{0}$$
$$g_{0} = Z' e_{0} = Z' y - Z' Z\delta_{0}$$
$$d_{0} = g_{0}$$
$$\rho_{0} = g_{0}' g_{0}$$
$$\lambda_{0} = 0$$
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 $\delta$ 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 $\ell = 0,1,2,3,…$

\[
\begin{align*}
  d_\ell &= g_\ell + \tau_{\ell-1} d_{\ell-1} \\
  q_\ell &= Zd_\ell \\
  \lambda_\ell &= \rho_\ell / (q_\ell ' q_\ell) \\
  \delta_{\ell+1} &= \delta_\ell + \lambda_\ell d_\ell \\
  e_{\ell+1} &= e_\ell - \lambda_\ell q_\ell \\
  g_{\ell+1} &= Z' e_{\ell+1} \\
  \begin{bmatrix} 
    \beta_{\ell+1} \\
    \theta_{\ell+1} \\
    \psi_{\ell+1}
  \end{bmatrix} &= \Delta^{-1/2} \delta_{\ell+1}
\end{align*}
\]
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.

\[
\begin{bmatrix}
X \beta_{\ell+1} \\
D \theta_{\ell+1} \\
F \psi_{\ell+1}
\end{bmatrix} e_{\ell+1} \begin{bmatrix} c \\ c \\ c \end{bmatrix} < 0 \text{, stop } \delta = \delta_{\ell+1} \\
\text{else, continue}
\]

\[
\rho_{\ell+1} = (g_{\ell+1}', g_{\ell+1}) \\
\tau_{\ell} = \rho_{\ell+1} / \rho_{\ell}
\]

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

\[
S = \left( y - Z \hat{\delta} \right) \left( y - Z \hat{\delta} \right)
\]

© John M. Abowd 2005, all rights reserved
Mixed Effects Assumptions

\[ \Lambda = \begin{bmatrix} \Sigma_1 & 0 & \ldots & 0 \\ 0 & \Sigma_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \Sigma_N \end{bmatrix} \]

\[ E\left[ \begin{bmatrix} \theta \\ \psi \end{bmatrix} \mid X \right] = 0 \quad V\left[ \begin{bmatrix} \theta \\ \psi \end{bmatrix} \mid X \right] = \Omega \]

- 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

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

- Solve the mixed effects equations
- Techniques: Bayesian EM, Restricted ML
Relation Between Fixed and Mixed Effects Models

\[ \Lambda = \sigma^2 \epsilon I_{N^*} \quad \left| \begin{array}{c} \Omega \\ \rightarrow \infty \end{array} \right. \]

- 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 \( \Omega \) is diagonal.

© John M. Abowd 2005, all rights reserved
Software

- SAS: proc mixed
- ASREML
- aML
- SPSS: Linear Mixed Models
- STATA: xtreg, gllamm, xtmixed
- R: the lme() function
- S+: linear mixed models
- Gauss
- Matlab
- Genstat: REML