Structural Equation Modeling

Using \texttt{gllamm}, \texttt{confa} and \texttt{gmm}

Stas Kolenikov

Department of Statistics
University of Missouri-Columbia
Joint work with Kenneth Bollen (UNC)

To be given: July 15, 2010
This draft: June 27, 2010
Goals of the talk

1. Introduce structural equation models
2. Describe Stata packages to fit them:
   - `confa`: a 5/8” hex wrench
   - `gllamm`: a Swiss-army tomahawk
   - `gmm`: do-it-yourself kit
3. Give example(s)
   - Health: daily functioning in NHANES
   - Sociology: industrialization and political democracy
   - Psychology: Holzinger-Swineford data
First, some theory

Introduction

Structural equation models
  Formulation
  Path diagrams
  Identification

Stata tools for SEM
  gllamm
  confa
  gmm+sem4gmm

NHANES daily functioning

Outlets

References
 Structural equation modeling (SEM)

- Standard multivariate technique in social sciences
- Incorporates constructs that cannot be directly observed:
  - psychology: level of stress
  - sociology: quality of democratic institutions
  - biology: genotype and environment
  - health: difficulty in personal functioning
- Special cases:
  - linear regression
  - confirmatory factor analysis
  - simultaneous equations
  - errors-in-variables and instrumental variables regression
Origins of SEM

Path analysis of Sewall Wright (1918)

Causal modeling of Hubert Blalock (1961)

Factor analysis estimation of Karl Jöreskog (1969)

Econometric simultaneous equations of Arthur Goldberger (1972)
Structural equations model

Latent variables:

$$\eta = \alpha_\eta + B\eta + \Gamma \xi + \zeta$$  \hspace{1cm} (1)

Measurement model for observed variables:

$$y = \alpha_y + \Lambda_y \eta + \varepsilon$$  \hspace{1cm} (2)
$$x = \alpha_x + \Lambda_x \xi + \delta$$  \hspace{1cm} (3)

$\xi$, $\zeta$, $\varepsilon$, $\delta$ are uncorrelated with one another

Implied moments

Denoting

\[
\begin{align*}
\mathbb{V}[\xi] &= \Phi, & \mathbb{V}[\zeta] &= \Psi, & \mathbb{V}[\varepsilon] &= \Theta_\varepsilon, & \mathbb{V}[\delta] &= \Theta_\delta, \\
R &= \Lambda_y (I - B)^{-1}, & z &= (x', y')'
\end{align*}
\]

obtain

\[
\begin{align*}
\mu(\theta) &\equiv \mathbb{E}[z] = \\
&= \begin{pmatrix}
\alpha_y + \Lambda_y (I - B)^{-1} \Gamma \mu_\xi \\
\alpha_x + \Lambda_x \mu_\xi
\end{pmatrix} \\
\Sigma(\theta) &\equiv \mathbb{V}[z] = \\
&= \begin{pmatrix}
\Lambda_x \Phi \Lambda_x' + \Theta_\delta & \Lambda_x \Phi \Gamma' R' \\
R \Gamma \Phi \Lambda_x' & R (\Gamma \Phi \Gamma' + \Psi) R' + \Theta_\varepsilon
\end{pmatrix}
\end{align*}
\]
Path diagrams

\[
\begin{align*}
&\xi_1 \\
&\quad \delta_1 \langle \theta_1 \rangle \\
&\quad \delta_2 \langle \theta_2 \rangle \\
&\quad \delta_3 \langle \theta_3 \rangle \\
&\quad x_1 \\
&\quad x_2 \\
&\quad x_3 \\
\end{align*}
\]

\[
\begin{align*}
&\eta_1 \\
&\quad \lambda_1 \langle \theta_1 \rangle \\
&\quad \lambda_2 \\
&\quad \lambda_3 \\
&\quad \beta_{11} \\
&\quad \phi_{12} \\
&\quad \beta_{12} \\
&\quad z_1 \\
&\quad \langle \phi_{11} \rangle \\
&\quad \langle \phi_{22} \rangle \\
\end{align*}
\]

\[
\begin{align*}
&\zeta_1 \\
&\quad \zeta_1 \langle \sigma_1 \rangle \\
\end{align*}
\]

\[
\begin{align*}
&\lambda_5 \\
&\lambda_6 \\
&\beta_{11} \\
&\phi_{12} \\
&\eta_1 \\
&y_1 \\
&\langle \theta_4 \rangle \\
&\epsilon_1 \\
\end{align*}
\]

\[
\begin{align*}
&y_2 \\
&\langle \theta_5 \rangle \\
&\epsilon_2 \\
\end{align*}
\]

\[
\begin{align*}
&y_3 \\
&\langle \theta_6 \rangle \\
&\epsilon_3 \\
\end{align*}
\]
Identification

Before proceeding to estimation, the researcher needs to verify that the SEM is identified:

\[ \Pr\{X : f(X, \theta) = f(X, \theta') \Rightarrow \theta = \theta'\} = 1 \]

Different parameter values should give rise to different likelihoods/objective functions, either globally, or locally in a neighborhood of a point in a parameter space.
Likelihood

- Normal data ⇒ likelihood is the function of sufficient statistic \( (\bar{z}, S) \):

\[
-2 \log L(\theta, Y, X) \sim n \ln \det(\Sigma(\theta)) + n \text{tr}[\Sigma^{-1}(\theta)S] \\
+n(\bar{z} - \mu(\theta))'\Sigma^{-1}(\theta)(\bar{z} - \mu(\theta)) \to \min_{\theta}
\]

- Generalized latent variable approach for mixed response (normal, binomial, Poisson, ordinal, within the same model):

\[
-2 \log L(\theta, Y, X) \sim \sum_{i=1}^{n} \ln \int f(y_i, x_i | \xi, \zeta; \theta)dF(\xi, \zeta | \theta)
\]

Estimation methods

- (quasi-)MLE
- Weighted least squares:

\[
s = \text{vech } S, \quad \sigma(\theta) = \text{vech } \Sigma(\theta)
\]

\[
F = (s - \sigma(\theta))' V_n (s - \sigma(\theta)) \rightarrow \min_{\theta}
\]  

where \( V_n \) is weighting matrix:

- Optimal \( \hat{V}_n^{(1)} = \hat{V} [s - \sigma(\theta)] \) (Browne 1984)
- Simplistic: least squares \( V_n^{(2)} = I \)
- Diagonally weighted least squares: \( \hat{V}_n^{(3)} = \text{diag } \hat{V} [s - \sigma] \)

- Model-implied instrumental variables limited information estimator (Bollen 1996)
- Empirical likelihood
Goodness of fit

- The estimated model \( \Sigma(\hat{\theta}) \) is often related to the “saturated” model \( \Sigma = S \) and/or independence model \( \Sigma_0 = \text{diag} S \)

- Likelihood formulation \( \Rightarrow \) LRT test, asymptotically \( \chi^2_k \)

- Non-normal data: LRT statistic \( \sim \sum_j w_j \chi^2_1 \), can be Satterthwaite-adjusted towards the mean and variance of the appropriate \( \chi^2_k \) (Satorra & Bentler 1994, Yuan & Bentler 1997)

- Analogies with regression \( R^2 \) attempted, about three dozen fit indices available (Marsh, Balla & Hau 1996)

- Reliability of indicators: \( R^2 \) in regression of an indicator on its latent variable

- Signs and magnitudes of coefficient estimates
Now, some tools

1. Introduction

1. Structural equation models
   Formulation
   Path diagrams
   Identification
   Estimation

2. Stata tools for SEM
   - gllamm
   - confa
   - gmm+sem4gmm

3. NHANES daily functioning

4. Outlets

5. References

- Exploits commonalities between latent and mixed models
- Adds GLM-like links and family functions to them
- Allows heterogeneous response (different exponential family members)
- Allows multiple levels
- Maximum likelihood via numeric integration of random effects and latent variables (Gauss-Newton quadrature, adaptive quadrature); hence one of the most computationally demanding packages ever
• One observation per dependent variable × observation
• Requires \texttt{reshape long} transformation of indicators for latent variable models
• Measurement model: \texttt{eq()} option
• Structural model: \texttt{geq()} \texttt{bmatrix()} options
• Families and links: \texttt{family()} \texttt{fv()} \texttt{link()} \texttt{lv()}
• Tricks that Stas commonly uses:
  • make sure the model is correctly specified: \texttt{trace noest options}
  • good starting values speed up convergence: \texttt{from()} option
  • number of integration points gives tradeoff between speed and accuracy: \texttt{nip()} option
  • get an idea about the speed: \texttt{dot option}
• CONfirmatory Factor Analysis models, a specific class of SEM
• Maximum likelihood estimation
• Arbitrary # of factors and indicators; correlated measurement errors
• Variety of standard errors (OIM, sandwich, distributionally robust)
• Variety of fit tests (LRT, various scaled tests)
• Post-estimation:
  • fit indices;
  • factor scores (predictions)
New (as of Stata 11) estimation command \texttt{gmm}:

- Estimation by minimization of

\[
g(X, \theta)' V_n g(X, \theta) \rightarrow \min_{\theta}
\]

- Evaluator vs. “regression+instruments”

- Variety of weight matrices \( V_n \)

- Homoskedastic/unadjusted or heteroskedastic/robust standard errors

- Overidentification (goodness of fit) \( J \)-test via \texttt{estat overid}
Least squares estimators can be implemented using \textit{gmm} (Kolenikov & Bollen 2010).

1. Compute the implied moment matrix $\Sigma(\theta)$ (user-specified Mata function \texttt{ParsToSigma()})

2. Form observation-by-observation contributions to the moment conditions $\text{vech}[(x_i - \bar{x})(x_i - \bar{x})' - \Sigma(\theta)]$ (Mata function \texttt{VechData()} provided by Stas)

3. Feed into \textit{gmm} using moment evaluator function \texttt{sem4gmm} (provided by Stas)

4. Enjoy!
LS family of estimators

- **Common part:**
  \[ gmm \text{ sem4gmm, parameters('pars')} \ldots \]
- **ULS:**
  \[ \ldots \text{ winit(id) onestep vce(unadj)} \]
- **DWLS:**
  \[ \ldots \text{ winit(unadj, indep) wmat(unadj, indep) twostep} \]
- **ADF:**
  \[ \ldots \text{ twostep | igmm} \]
## Comparison of functionality

<table>
<thead>
<tr>
<th></th>
<th><code>gllamm</code></th>
<th><code>confa</code></th>
<th><code>gmm+sem4gmm</code></th>
</tr>
</thead>
<tbody>
<tr>
<td>General SEM</td>
<td>…</td>
<td>–</td>
<td>√</td>
</tr>
<tr>
<td>Estimation</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Overall test</td>
<td>–</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Fit indices</td>
<td>–</td>
<td>…</td>
<td>–</td>
</tr>
<tr>
<td>Prediction</td>
<td>√</td>
<td>…</td>
<td>–</td>
</tr>
<tr>
<td>Ease of use</td>
<td>–</td>
<td>√</td>
<td>–</td>
</tr>
<tr>
<td>Speed</td>
<td>–</td>
<td>…</td>
<td>–</td>
</tr>
</tbody>
</table>
Finally, examples

1. Introduction
2. Structural equation models
   Formulation
   Path diagrams
   Identification
   Estimation
3. Stata tools for SEM
   gllamm
   confa
   gmm+sem4gmm
4. NHANES daily functioning
5. Outlets
6. References
NHANES data

- NHANES 2007–08 data
- Personal functioning section: “difficulty you may have doing certain activities because of a health problem”
- 17 questions: Walking for a quarter mile; Walking up ten steps; Stoop, crouching, kneeling; Lifting or carrying; House chore; Preparing meals; Walking between rooms on same floor; Standing up from armless chair; Getting in and out of bed; Dressing yourself; Standing for long periods; Sitting for long periods; Reaching up over head; Grasp/holding small objects; Going out to movies, events; Attending social event; Leisure activity at home
- Response categories: “No difficulty”, “Some difficulty”, “Much difficulty”, “Unable to do”
- Research questions: How to summarize these items? What’s the relation between individual demographics and health?
A multiple indicators and multiple causes (MIMIC) model
NHANES example using `confa`

Only the measurement model can be estimated with `confa`, as a preliminary step in gauging the performance of this part of the model.

```
. confa (difficulty: pfq*), from(iv)

. confa (difficulty: pfq*), from(iv)
> missing
```

Show results: estimates use confa_pwise, estimates use confa_fiml
Factor scores

Age at Screening Adjudicated - Recode

PF score, CFA model

-1 0 1 2 3

20 40 60 80
NHANES example via \texttt{gllamm}

Data management steps for \texttt{gllamm}:

1. **Rename** \texttt{pfq061b} $\rightarrow$ \texttt{pfq1}, \texttt{pfq061c} $\rightarrow$ \texttt{pfq2}, ... \texttt{pfq061s} $\rightarrow$ \texttt{pfq17}
2. **Reshape long** \texttt{pfq}, \texttt{i(seqn)} j(\texttt{item})
3. **Generate binary indicators** \texttt{q1}--\texttt{q17} of the items
4. **Produce binary outcome measures**: 
   \texttt{bpfq`k' = ![No difficulty] of pfq`k'}

Model setup steps:

1. **Define loading equations**: 
   \texttt{eq items: q1 q2 ...q17}
2. **Come up with good starting values**
NHANES example via `gllamm`

Syntax of `gllamm` command:
```
gllamm ///
bpfq ///   single dependent variable
q1 - q17, nocons ///  item-specific intercepts
i(seqn) ///  “common factor”
f(bin) l(probit) ///  link and family
eq(items) ///  loadings equation
from(...) copy  starting values
```

The “common factor” is a latent variable that is constant across the `i()` panel, but can be modified with loadings.

Show results in Stata: `est use cfa_via_gllamm; gllamm`
MIMIC model

Additional estimation steps:

1. Store the CFA results: `mat hs_cfa = e(b)`
2. Define the explanatory variables for functioning: `eq r1: female age splines`
3. Extend the earlier command: `gllamm ..., geq(r1) from(hs_cfa, skip)`

Parameter “complexity”: 

1. fixed effects
2. loadings
3. latent regression slopes
4. latent (co)variances

Show results in Stata: `est use mimic_bmi; gllamm; show the diagram again.`
NHANES example via \textit{gmm}

Full model:
- 1 latent variable $\Rightarrow$ 1 variance
- 17 indicators $\Rightarrow$ 17 loadings, 17 variances
- 7 explanatory variables $\Rightarrow$ $7 \cdot \frac{8}{2}$ covariances, 7 regression coefficients
- Total: 70 parameters, 300 moment conditions

Trimmed model:
- 1 latent variable $\Rightarrow$ 1 variance
- 5 indicators $\Rightarrow$ 5 loadings, 5 variances
- 4 explanatory variables $\Rightarrow$ $4 \cdot \frac{5}{2}$ covariances, 4 regression coefficients
- Total: 25 parameters, 45 moment conditions
NHANES example: syntax and results

Show syntax: nhanes-def-sem-reduced.do, nhanes-gmm-est-reduced.do

Show results:
foreach eres in r_uls_homosked r_uls_heterosked r_dwls_2step_heterosked r_effls_2step_heterosked r_effls_Igmm_heterosked {
  est use ‘eres’
  est store ‘eres’
}
estimates table, se stats(J)
# Main journals

<table>
<thead>
<tr>
<th>Journal title</th>
<th>Impact factor</th>
<th>$h$-index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structural Equation Modeling</td>
<td>2.4</td>
<td>15</td>
</tr>
<tr>
<td>Psychometrika</td>
<td>1.1</td>
<td>27</td>
</tr>
<tr>
<td>British Journal of Mathematical and Statistical Psychology</td>
<td>1.3</td>
<td>20</td>
</tr>
<tr>
<td>Multivariate Behavioral Research</td>
<td>1.8</td>
<td>30</td>
</tr>
<tr>
<td>Psychological Methods</td>
<td>4.3</td>
<td>52</td>
</tr>
<tr>
<td>Sociological Methodology</td>
<td>2.5</td>
<td>21</td>
</tr>
<tr>
<td>Sociological Methods and Research</td>
<td>1.2</td>
<td>24</td>
</tr>
<tr>
<td>JASA</td>
<td>2.3</td>
<td>74</td>
</tr>
<tr>
<td>Biometrika</td>
<td>1.3</td>
<td>48</td>
</tr>
<tr>
<td>J of Multivariate Analysis</td>
<td>0.7</td>
<td>24</td>
</tr>
<tr>
<td>Stata Journal</td>
<td>1.3</td>
<td>9</td>
</tr>
</tbody>
</table>

What I covered was…

1. **Introduction**

2. **Structural equation models**
   - Formulation
   - Path diagrams
   - Identification
   - Estimation

3. **Stata tools for SEM**
   - `gllamm`
   - `confa`
   - `gmm+sem4gmm`

4. **NHANES daily functioning**

5. **Outlets**

6. **References**
References


References II


References III


References IV
