Structural Equation Modeling Using `gllamm`, `confa` and `gmm`

Stas Kolenikov

Department of Statistics
University of Missouri-Columbia
The World Bank, Washington, DC

Joint work with Kenneth Bollen (UNC)

July 1, 2011

Goals of the talk

1. Introduce structural equation models
2. Describe Stata packages to fit them:
   - `confa`: a 13mm hex wrench
   - `gllamm`: a Swiss-army tomahawk
   - `gmm`: do-it-yourself kit
   - `sem`: the promised land?
3. Example 1: daily functioning in NHANES
4. Example 2: experimental ecology data set

First, some theory

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)

$\otimes$

Causal modeling of Hubert Blalock (1961)

$\otimes$

Factor analysis estimation of Karl Jöreskog (1969)

$\otimes$

Econometric simultaneous equations of Arthur Goldberger (1972)

Structural equations model

Latent variables:

$$\boldsymbol{\eta} = \boldsymbol{\alpha}_\eta + \boldsymbol{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \tag{1}$$

Measurement model for observed variables:

$$\boldsymbol{y} = \boldsymbol{\alpha}_y + \boldsymbol{\Lambda}_y\boldsymbol{\eta} + \boldsymbol{\varepsilon} \tag{2}$$

$$\boldsymbol{x} = \boldsymbol{\alpha}_x + \boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta} \tag{3}$$

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

Jöreskog (1973), Bollen (1989), Yuan & Bentler (2007)

Other re-expressions: Bentler & Weeks (1980), McArdle & McDonald (1984).

Implied moments

Denoting

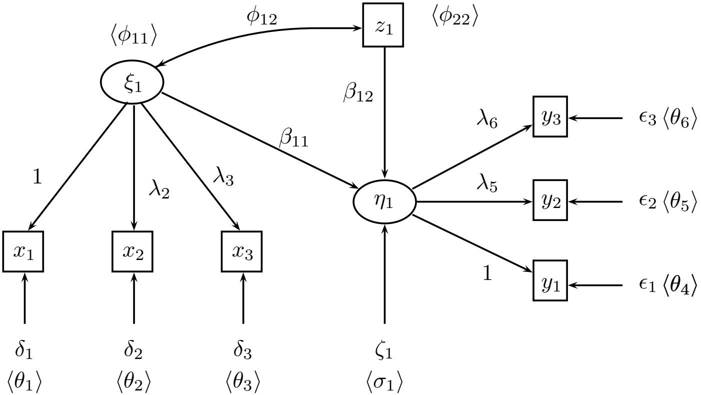
$$\mathbb{V}[\boldsymbol{\xi}] = \Phi, \quad \mathbb{V}[\boldsymbol{\zeta}] = \Psi, \quad \mathbb{V}[\boldsymbol{\varepsilon}] = \Theta_\varepsilon, \quad \mathbb{V}[\boldsymbol{\delta}] = \Theta_\delta,$$
$$R = \boldsymbol{\Lambda}_y (I - B)^{-1}, \quad z = \begin{pmatrix} x \\ y \end{pmatrix}$$

obtain

$$\mu(\boldsymbol{\theta}) \equiv \mathbb{E}\big[z\big] = \begin{pmatrix} \boldsymbol{\alpha}_y + \boldsymbol{\Lambda}_y R \boldsymbol{\mu}_\xi \\ \boldsymbol{\alpha}_x + \boldsymbol{\Lambda}_x \boldsymbol{\mu}_\xi \end{pmatrix} \tag{4}$$

$$\Sigma(\boldsymbol{\theta}) \equiv \mathbb{V}\big[z\big] = \begin{pmatrix} \boldsymbol{\Lambda}_x \Phi \boldsymbol{\Lambda}_x' + \Theta_\delta & \boldsymbol{\Lambda}_x \Phi \Gamma' R' \\ R \Gamma \Phi \boldsymbol{\Lambda}_x' & R(\Gamma \Phi \Gamma' + \Psi) R' + \Theta_\varepsilon \end{pmatrix} \tag{5}$$

Path diagrams



Identification

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

$$\mathbf{Pr}\{X : f(X,\theta) = f(X,\theta') \Rightarrow \theta = \theta'\} = 1$$

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 $\Rightarrow$ likelihood is the function of sufficient statistic $(\bar{z}, S)$:

$$
\begin{aligned}
-2\log L(\theta, Y, X) \sim n \ln \det\bigl(\Sigma(\theta)\bigr) + n\,\mathrm{tr}[\Sigma^{-1}(\theta)S] \\
+n(\bar{z} - \mu(\theta))'\Sigma^{-1}(\theta)(\bar{z} - \mu(\theta)) \to \min_{\theta}
\end{aligned}
\tag{6}
$$

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

$$
-2\log L(\theta, Y, X) \sim \sum_{i=1}^{n} \ln \int f(y_i, x_i | \xi, \zeta; \theta) \mathrm{d}F(\xi, \zeta | \theta) \tag{7}
$$

Bartholomew & Knott (1999), Skrondal & Rabe-Hesketh (2004)

Estimation methods

- Normal theory MLE
- Weighted least squares:

$$s = \operatorname{vech} S, \quad \sigma(\theta) = \operatorname{vech} \Sigma(\theta)$$

$$F = (s - \sigma(\theta))' V_n (s - \sigma(\theta)) \to \min_\theta \tag{8}$$

  where V_n is weighting matrix:
  - Optimal $\hat{V}_n^{(1)} = \hat{\mathbb{V}}[s - \sigma(\theta)]$ (Browne 1984)
  - Simplistic: least squares $V_n^{(2)} = I$
  - Diagonally weighted least squares: $\hat{V}_n^{(3)} = \operatorname{diag} \hat{\mathbb{V}}[s - \sigma]$
- Model-implied instrumental variables limited information estimator (Bollen 1996)
- Bounded influence/outlier-robust methods (Yuan, Bentler & Chan 2004, Moustaki & Victoria-Feser 2006)
- Empirical likelihood

Goodness of fit

- The estimated model $\Sigma(\hat{\theta})$ is often related to the "saturated" model $\Sigma \equiv S$ and/or independence model $\Sigma_0 = \operatorname{diag} S$
- Likelihood formulation $\Rightarrow$ LRT test, asymptotically χ^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 χ^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

sem?

As announced earlier this week, Stata 12 will be released on 25 July 2011 and will have a full-fledge `sem` estimation routine.

sem

Viewer - view sem2.smcl

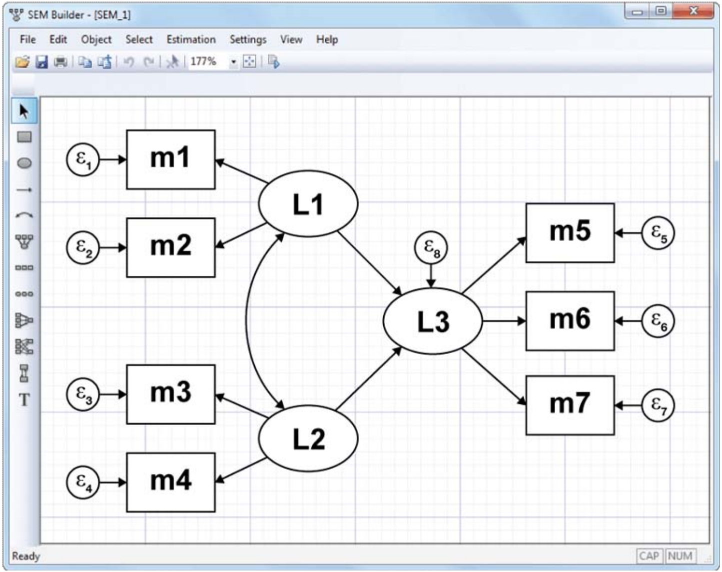
File Edit History Help

view sem2.smcl

Dialog | Also See | Jump To

```
Endogenous variables

Measurement:  anomia67 pwless67 anomia71 pwless71 educ66 occstat66
Latent:       Alien67 Alien71

Exogenous variables

Latent:       SES

Structural equation model                       Number of obs      =       932
Estimation method  = ml
Log likelihood     = -15246.469

 ( 1)  [anomia67]Alien67 = 1
 ( 2)  [anomia71]Alien71 = 1
 ( 3)  [educ66]SES = 1

------------------------------------------------------------------------------
             |                 OIM
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  Alien71 <- |
     Alien67 |   .7046345   .0533512    13.21   0.000     .6000681    .8092008
         SES |  -.1744151   .0542489    -3.22   0.001     -.280741   -.0680891
  -----------+----------------------------------------------------------------
  Alien67 <- |
```

sem


SEM Builder - [SEM_1]
File
Edit
Object
Select
Estimation
Settings
View
Help
177%
m1
m2
m3
m4
m5
m6
m7
L1
L2
L3
ε1
ε2
ε3
ε4
ε5
ε6
ε7
ε8
Ready
CAP
NUM

`gllamm`

Generalized Linear Latent And Mixed Models (Skrondal & Rabe-Hesketh 2004, Rabe-Hesketh, Skrondal & Pickles 2005, Rabe-Hesketh & Skrondal 2008)

- 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

`gllamm`

- One line of data per dependent variable $\times$ unit
- Requires `reshape long` transformation of indicators for latent variable models
- Measurement model: `eq()` option
- Structural model: `geq()` `bmatrix()` options
- Families and links: `family()` `fv()` `link()` `lv()`
- Tricks that Stas commonly uses:
  - make sure the model is correctly specified: `trace noest` options
  - good starting values speed up convergence: `from()` option
  - number of integration points gives tradeoff between speed and accuracy: `nip()` option
  - get an idea about the speed: `dot` option

`confa` package

- 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)
  - Bollen & Stine (1992) bootstrap

gmm

Estimation command `gmm` introduced in Stata 11:

- Estimation by minimization of

$$g(X,\theta)' V_n \, g(X,\theta) \to \min_{\theta}$$

- Evaluator vs. "regression+instruments"
- Variety of weight matrices V_n
- Asy efficient estimator: $V_n = \widehat{\mathbb{V}} g(X, \hat\theta)$
- Homoskedastic/`unadjusted`, heteroskedastic/`robust`, `cluster`'ed and HAC-consistent standard errors
- Overidentification (goodness of fit) J-test via `estat overid`

One possible set up for SEM

1. Write a program to compute the implied moment matrix $\Sigma(\theta)$
2. Form observation-by-observation contributions to the moment conditions
$g(X,\theta) = \mathrm{vech}\big[(x_i-\bar{x})(x_i-\bar{x})' - \Sigma(\theta)\big]$
3. Feed into `gmm` using moment evaluator function

Some of these steps were simplified by the author's `sem4gmm` which will be obsolete in Stata 12.

Another possible set up for SEM

- Rather than relying on covariance representation of SEM, one can use regression representation instead
- Latent variables are measured with error $\Rightarrow$ need to use the techniques to account for that
- Observed indicators of latent variables are endogenous variables in the model

Account for endogeneity by instrumental variables

- Econometric technique of instrumental variables adapted to SEM by Bollen (1996)
- An instrumental variable:
  - correlated with regressors
  - not correlated with the error term
- Single equation: `ivregress`
- Simultaneous equations: all earlier determined variables can serve as instruments
- Full structural equation model: tracing rules Bollen & Bauer (2004)
- Can be implemented using the "interactive" version of `gmm`
- Tests of model specification: by equation and for the system as a whole

Comparison of functionality

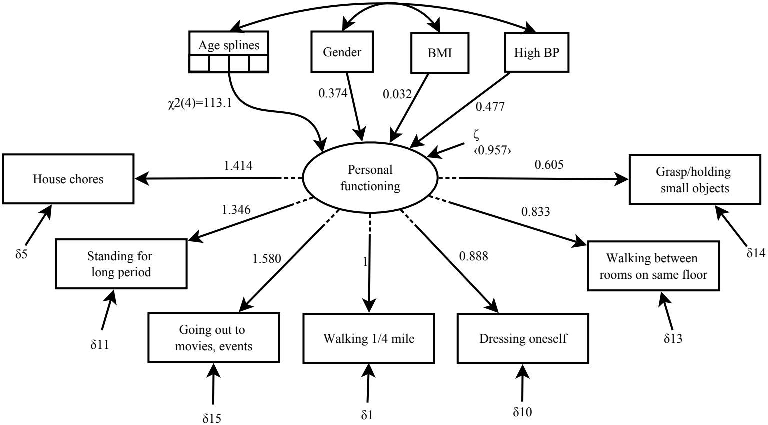
	gllamm	confa	gmm $+\Sigma(\theta)$	gmm + IV
General SEM	$\ldots$	–	$\surd$	$\ldots$
Estimation	$\surd$	$\surd$	$\surd$	$\ldots$
Overall test	–	$\surd$	$\surd$	$\surd$
Fit indices	–	$\ldots$	–	–
Prediction	$\surd$	$\ldots$	–	–
Ease of use	–	$\surd$	–	$\ldots$
Speed	–	$\ldots$	–	$\surd$

Finally, examples

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; Stooping, 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?

Path diagram



A multiple indicators and multiple causes (MIMIC) model

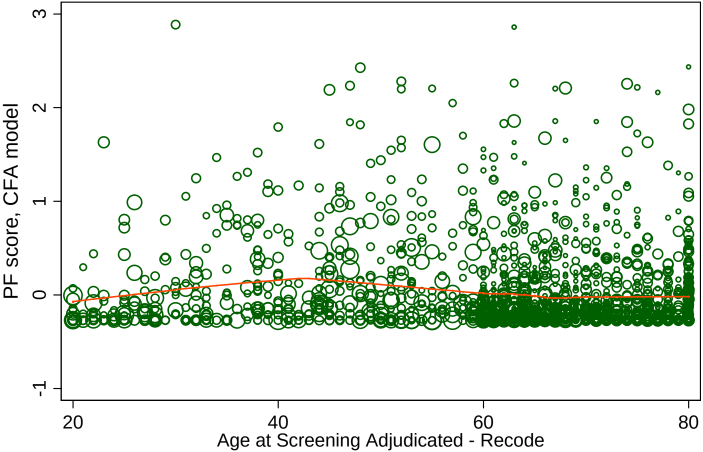
NHANES example using confa

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 cfa; cfa_miss_fromcfa; cfa_miss_fromiv`

Factor scores



NHANES example via `gllamm`

Data management steps for `gllamm`:

1. Rename `pfq061b`↦`pfq1`, `pfq061c`↦`pfq2`, `...pfq061s`↦`pfq17`
2. `reshape long pfq, i(seqn) j(item)`
3. Generate binary indicators `q1-q17` of the items
4. Produce binary outcome measures: `bpfq`k'` = !(“No difficulty”) of `pfq`k'`

Model setup steps:

1. Define loading equations: `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 bmi highbp` *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 gmm

Full model:

- 1 latent variable $\Rightarrow$ 1 variance
- 17 indicators $\Rightarrow$ 17 loadings, 17 variances
- 7 explanatory variables $\Rightarrow 7 \cdot 8/2$ 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 $\Rightarrow 4 \cdot 5/2$ 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:

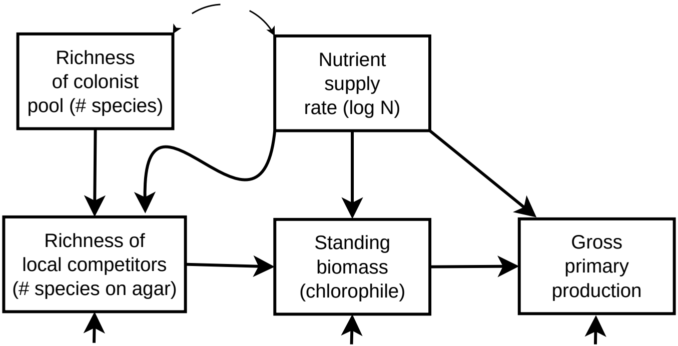
```
foreach eres in r_uls_homosked
r_uls_heterosked r_dwls_2step_heterosked
r_effls_2step_heterosked
r_effls_igmm_heterosked {
    est use `eres'
    gmm
    est store `eres'
}
estimates table, se stats(J)
```

Ecology example: observed variables

SEM in ecology

- Truly continuous variables, rather than Likert scales
- Observed and/or composite variables
- Small sample sizes (you're lucky if you have a few dozen)
- Methodology is at early stages of adoption
- Existing textbooks: Shipley (2000), Pugesek, Tomer & von Eye (2002)

Richness vs. productivity



Cardinale, Bennett, Nelson & Gross (2009)

First step: `regress`

```
regress ///
    dependent var ///
    its predictors from the path diagram
```

Account for endogeneity: `ivregress`

```
ivregress 2sls ///
   dependent var ///
   its exogenous predictors ///
      from the path diagram ///
   (its sl endogenous predictors = ///
      variables before them ///
      in the path model)
```

```
estat overid
```

Systemwide estimation: `reg3`

```
reg3 ///
    (depvar1 explvars1) ///
    (depvar2 explvars2) ///
```

Stata figures out the instrumental variables as all exogenous variables.
It will also implicitly correlate the errors to improve efficiency.

Systemwide estimation: gmm

gmm ///
    (explicit equation for first regression) ///
    (explicit equation for first regression) ///
    ... ///
    , winitial(id) wmatrix(robust) [igmm] ///
    instruments(1: instruments for first regression)
///
    ...

estat overid

Mediation, direct and indirect effects

- Is the effect of N on production mediated by biomass?
- Direct effect: regression coefficient
- Indirect effect: influence of N propagates through its effects on richness of local competition and biomass
- Algebraic expressions available, so this is the job for `nlcom`

What I covered was. . .

References I

Bartholomew, D. J. & Knott, M. (1999), *Latent Variable Models and Factor Analysis*, Vol. 7 of *Kendall's Library of Statistics*, 2nd edn, Arnold Publishers, London.

Bentler, P. M. & Weeks, D. G. (1980), 'Linear structural equations with latent variables', *Psychometrika* **45**, 289–308.

Blalock, H. M. (1961), 'Correlation and causality: The multivariate case', *Social Forces* **39**(3), 246–251.

Bollen, K. A. (1989), *Structural Equations with Latent Variables*, Wiley, New York.

Bollen, K. A. (1996), 'An alternative two stage least squares (2SLS) estimator for latent variable models', *Psychometrika* **61**(1), 109–121.

Bollen, K. A. & Bauer, D. J. (2004), 'Automating the selection of model-implied instrumental variables', *Sociological Methods Research* **32**(4), 425–452.

References II

Bollen, K. A. & Stine, R. (1992), ‘Bootstrapping goodness of fit measures in structural equation models’, *Sociological Methods and Research* **21**, 205–229.

Browne, M. W. (1984), ‘Asymptotically distribution-free methods for the analysis of the covariance structures’, *British Journal of Mathematical and Statistical Psychology* **37**, 62–83.

Cardinale, B. J., Bennett, D. M., Nelson, C. E. & Gross, K. (2009), ‘Does productivity drive diversity or vice versa? a test of the multivariate productivity-diversity hypothesis in streams’, *Ecology* **90**(5), 1227–1241.

Goldberger, A. S. (1972), ‘Structural equation methods in the social sciences’, *Econometrica* **40**(6), 979–1001.

Jöreskog, K. G. (1969), ‘A general approach to confirmatory maximum likelihood factor analysis’, *Psychometrika* **34**(2), 183–202.

References III

Jöreskog, K. G. (1973), A general method for estimating a linear structural equation system, *in* A. S. Goldberger & O. D. Duncan, eds, 'Structural Equation Models in the Social Sciences', Academic Press, New York, pp. 85–112.

Marsh, H. W., Balla, J. R. & Hau, K.-T. (1996), An evaluation of incremental fit indices: A clarification of mathematical and empirical properties, *in* G. Marcoulides & R. Schumaker, eds, 'Advanced Structural Equation Modeling Techniques', Erlbaum, Mahwah, NJ, pp. 315–353.

McArdle, J. J. & McDonald, R. P. (1984), 'Some algebraic properties of the reticular action model for moment structures.', *The British Journal of Mathematical and Statistical Psychology* **37**, 234–251.

Moustaki, I. & Victoria-Feser, M.-P. (2006), 'Bounded-influence robust estimation in generalized linear latent variable models', *Journal of the American Statistical Association* **101**(474), 644–653.

References IV

Pugesek, B. H., Tomer, A. & von Eye, A., eds (2002), *Structural Equation Modeling: Applications in Ecological and Evolutionary Biology*, Cambridge University Press.

Rabe-Hesketh, S. & Skrondal, A. (2008), 'Classical latent variable models for medical research', *Statistical Methods in Medical Research* **17**(1), 5–32.

Rabe-Hesketh, S., Skrondal, A. & Pickles, A. (2005), 'Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects', *Journal of Econometrics* **128**(2), 301–323.

Satorra, A. & Bentler, P. M. (1994), Corrections to test statistics and standard errors in covariance structure analysis, *in* A. von Eye & C. C. Clogg, eds, 'Latent Variable Analysis', Sage, Thousands Oaks, CA, chapter 16, pp. 399–419.

Shipley, B. (2000), *Cause and correlation in Biology: A user's guide to path analysis, structural equations and causal inference*, Cambridge Unversity Press, Cambridge, UK.

References V

Skrondal, A. & Rabe-Hesketh, S. (2004), *Generalized Latent Variable Modeling*, Chapman and Hall/CRC, Boca Raton, Florida.

Wright, S. (1918), 'On the nature of size factors', *Genetics* **3**, 367–374.

Yuan, K.-H., Bentler, P. & Chan, W. (2004), 'Structural equation modeling with heavy tailed distributions', *Psychometrika* **69**(3), 421–436.

Yuan, K.-H. & Bentler, P. M. (1997), 'Mean and covariance structure analysis: Theoretical and practical improvements', *Journal of the American Statistical Association* **92**(438), 767–774.

Yuan, K.-H. & Bentler, P. M. (2007), Structural equation modeling, *in* C. Rao & S. Sinharay, eds, 'Handbook of Statistics: Psychometrics', Vol. 26 of *Handbook of Statistics*, Elsevier, chapter 10.