Robust SEM for Non-Normal and Missing Data Using WebSEM

Zhiyong Zhang and Ke-Hai Yuan

UNIVERSITY OF NOTRE DAME

AMERICAN PSYCHOLOGICAL ASSOCIATION
ANNUAL CONVENTION
HONOLULU, HAWAI'I · JULY 31–AUGUST 4, 2013

Warning!

- WebSEM is free.

Warning!

- WebSEM is free.
- If you are not comfortable with this, we'd be happy to charge you to make you feel better.
- WebSEM is tested but comes without warranty.

Warning!

- WebSEM is free.
- If you are not comfortable with this, we'd be happy to charge you to make you feel better.
- WebSEM is tested but comes without warranty.

Warning!

- WebSEM is free.
- If you are not comfortable with this, we'd be happy to charge you to make you feel better.
- WebSEM is tested but comes without warranty.
- This talk is suspicious of self-promotion of WebSEM.

Outline

- Motivation of non-normal and missing data analysis
- An example on robust Cronbach's alpha and McDonald's omega
- Technical backgrounds for robust SEM
- WebSEM through examples
    - What is WebSEM?
    - Examples
- Q & A

Motivation – Non-normal data

- Practical data are often not normally distributed.
- Micceri, T. (1989). The Unicorn, The Normal Curve, and Other Improbable Creatures. *Psychological Bulletin*, *105*, 156–166.
  - 440 large-sample achievement and psychometric measures and all to be significantly nonnormal at $\alpha = 0.01$.
- Common sources
  - Longer or shorter tails
  - Skewness
  - Outlying observations

Influence of non-normal data

- Replication
- Type of data
  - Normal
  - Non-normal but satisfies certain requirements such as elliptical distribution or existence of certain moments
  - Non-normal data with outlying observations
- Evaluation criterion
  - Bias
  - Efficiency
  - Test statistics

- Methods
  - Normal distribution based methods (NML)
  - Distribution free methods (WLS, robust s.e.)
  - Robust methods (WebSEM)
- Comparison under asymptotic theory (large sample)

	Normal			Non-normal			Outlying		
	θ	s.e.	χ^2	θ	s.e.	χ^2	θ	s.e.	χ^2
NML	✔	✔	✔	✔	✗	✗	✗	✗	✗
Distribution free	✔	✔	✔	✔	✔	✔	✗	✗	✗
Robust	✔	✔	✔	✔	✔	✔	✔	✔	✔

Note. ✔ OK ✗ incorrect

Motivation – Missing data

- Practical data often include missing data.
- Variables used in the current model - $Y = (y_1, \ldots, y_p)$
- Missing data indicating variables - $M = (m_1, \ldots, m_p)$
- Auxiliary variables collected in a study not directly used in the current model - $A = (A_1, \ldots, A_s)$

	y_1	$\ldots$	y_p	m_1	$\ldots$	m_p	A_1	$\ldots$	A_s
1	O	O	O	0	0	0	O	O	O
2	-	O	O	1	0	0	O	O	O
3	O	O	O	0	0	0	O	-	O
$\vdots$	O	-	-	0	1	1	-	O	O
N	-	-	O	1	1	0	O	O	O

Missing mechanisms

- MCAR

$$\Pr(M|Y_{obs}, Y_{miss}, A, \boldsymbol{\theta}) = \Pr(M|\boldsymbol{\theta})$$

  - $\boldsymbol{\theta}$ represents unknown model parameters.
  - Missing data Y_{miss} are a simple random sample of Y.
  - The missingness is not related to D_{obs} or A.

- MAR

$$\Pr(M|Y_{obs}, Y_{miss}, A, \boldsymbol{\theta}) = \Pr(M|Y_{obs}, \boldsymbol{\theta})$$

  - The probability that a datum is missing is related to the data actually observed D_{obs} but not to the missing data D_{miss} or A.

- MNAR
  - The missing probability of a datum is related to the missing data D_{miss} or A, and
  - A are not included in the data analysis.
- Missing data methods and techniques in general assume that missing data are MCAR or MAR.
- If missingness is only related to A and A are observed and included in the data analysis, then the overall missing mechanism becomes MAR.

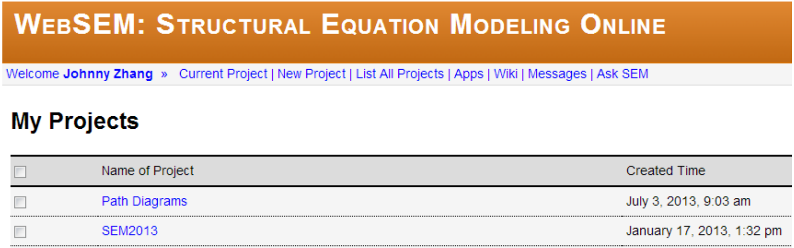
Methods dealing with missing data

- Listwise deletion
- Pairwise deletion
- Multiple imputation
- (Full information) Maximum likelihood method

	MCAR	MAR	MNAR	MNAR-A
Listwise	✔	✗	✗	-
Pairwise	✔	✗	✗	-
MI	✔	✔	✗	✔
FIML	✔	✔	✗	✔

Robust methods and WebSEM

- A robust procedure is developed to deal with both non-normal data and missing data simultaneously (e.g., Tong, Zhang, & Yuan, 2013; Yuan, 2013; Yuan, Tong, & Zhang, 2013; Yuan & Zhang, 2012a, 2012b; Zhang & Wang, 2012; Zhang & Yuan, 2013).
- The online software WebSEM is used to carry out the robust analysis (`https://websem.psychstat.org`).



Robust methods on reliability coefficients

- Given a test with p items with population mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})$. The sample covariance matrix is $\mathbf{S} = (s_{ij})$.
- Cronbach's alpha

$$\hat{\alpha} = \frac{p}{p-1}\left(1 - \frac{\sum_{i=1}^{p} s_{ii}}{\sum_{i=1}^{p}\sum_{j=1}^{p} s_{ij}}\right).$$

- McDonald's omega
  - Omega is defined on the factor model

$$y_{ij} = \mu_j + \lambda_j f_i + e_{ij}$$

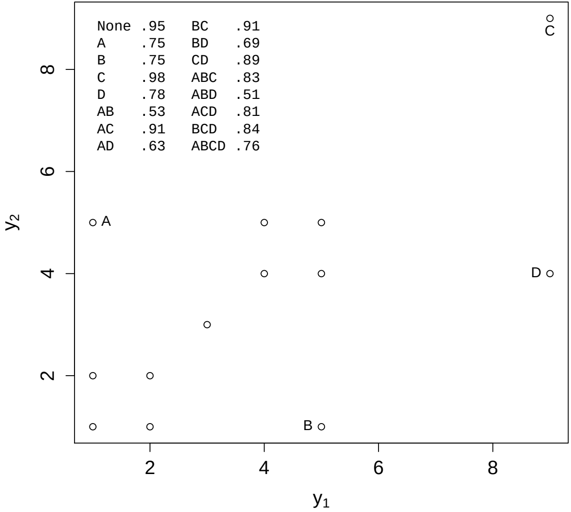
with $Var(e_{ij}) = \psi_j$.

▷

$$\hat{\omega} = \frac{(\sum_{k=1}^{p} \hat{\lambda}_j)^2}{(\sum_{k=1}^{p} \hat{\lambda}_j)^2 + (\sum_{k=1}^{p} \hat{\psi}_j)}.$$

- Alpha and omega are the same under tau-equivalent (McDonald, 1999).
- For non-tau-equivalent models, alpha and omega are often similar (e.g., Maydeu-Olivares et al., 2010).
- Both of them are influenced by outlying observations because the non-robustness of sample covariance matrix.

Influence of outlying observations on alpha



Types of outlying observations

- Invalid outlying observations
  - Erroneous observations that do not represent the underlying phenomena to be measured.
  - Data recording and input error is the most common cause.
- Valid outlying observations
  - Appear to be different from the majority of the data but truly represent the underlying phenomena.
  - Leverage observations
    - C has extremely large scores on both y_1 and y_2. The scores are extreme in the same direction.

- Common factor score shows extreme values.
- Good outlying observations - enlarge reliability and reduce s.e.

- Outliers
  - Show extremely values on certain items such as A and B.
  - Uniqueness factor scores show extreme values.
  - Bad outlying observations - reduce reliability and enlarge s.e..

Influences of outlying observations and missing data

- Data generation
  - 1000 sets of normal data on 6 items with N=100
  - Outlying observations
    - Outliers are generated by adding 4 from the first 3 items and subtracting 4 for the last three items for observations from 96 to 100.
    - Leverage observations are generated by subtracting on all items for observations from 96 to 100.
  - Missing data
    - Complete for the 1st and 4th item.
    - Missingness of the 2nd and 3rd items is related to the 4th item and missingness of the 5nd and 6rd items is related to the 1th item.

Results for outlying observations

- Population alpha and omega = 0.9.

	φ	alpha				omega			
		Est	s.e.	95% CI		Est	s.e.	95% CI	
Normal	0	.898	.015	.868	.928	.899	.016	.869	.929
	0.05	.898	.016	.867	.929	.899	.016	.868	.930
	0.1	.898	.016	.866	.930	.899	.016	.867	.931
outlier	0	.663	.109	.450	.875	.600	.101	.402	.798
	0.05	.863	.047	.770	.955	.862	.049	.766	.958
	0.1	.872	.033	.808	.936	.873	.033	.808	.938
Leverage	0	.972	.009	.954	.989	.972	.009	.954	.990
	0.05	.954	.023	.909	1.000	.955	.023	.909	1.000
	0.1	.948	.022	.905	.991	.948	.022	.905	.991

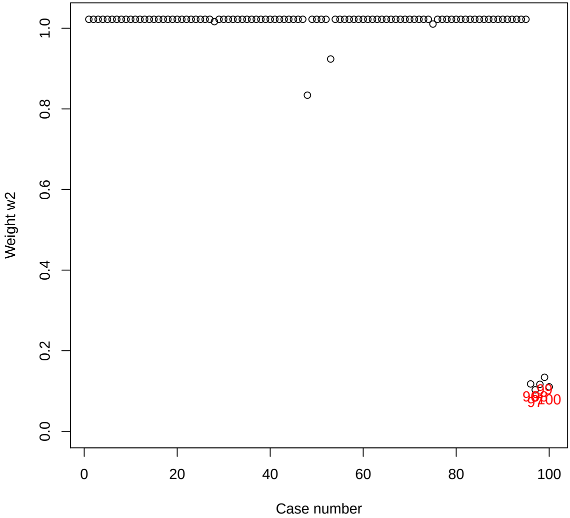
Results for missing data

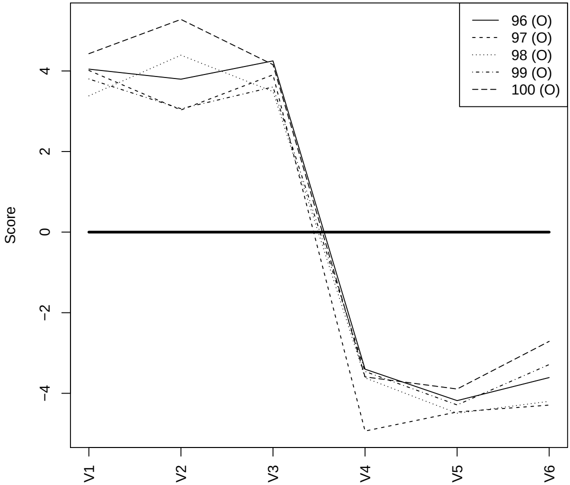
- Population alpha and omega = 0.9.

		alpha				omega			
	φ	Est	s.e.	CI		Est	s.e.	CI	
Deletion	0	.804	.036	.733	.875	.812	.037	.740	.884
	0.05	.804	.038	.729	.879	.812	.039	.736	.888
	0.1	.804	.039	.727	.880	.812	.039	.735	.889
ML	0	.898	.016	.867	.929	.899	.016	.868	.931
	0.05	.898	.016	.866	.930	.899	.016	.867	.932
	0.1	.898	.017	.865	.931	.899	.017	.866	.932

Why robust methods work?

- Smaller weights are given to outlying observations.




96 (O)
97 (O)
98 (O)
99 (O)
100 (O)
Score
4
2
0
−2
−4
V1
V2
V3
V4
V5
V6

Robust SEM: Settings

- Let $\mathbf{y}$ represents a population of p random variables with $E(\mathbf{y}) = \boldsymbol{\mu}$ and $Cov(\mathbf{y}) = \boldsymbol{\Sigma}$. A sample $\mathbf{y}_i,\ i = 1, 2, \ldots, N$, from $\mathbf{y}$ with missing values is available.

- The vector $\mathbf{u}$ represents $q - p$ auxiliary variables with associated sample realization $\mathbf{u}_i,\ i = 1, 2, \ldots, N$.

- Let $\mathbf{x}$ represents all the variables that we are interested and those that are auxiliary (not of substantial interest). Then, $\mathbf{x} = (\mathbf{y}', \mathbf{u}')'$ with $E(\mathbf{x}) = \boldsymbol{\nu}$ and $Cov(\mathbf{x}) = \mathbf{V}$.

- Due to missing values, the vector $\mathbf{x}_i = (\mathbf{y}_i', \mathbf{u}_i')'$ only contains q_i marginal observations of $\mathbf{x}$. The mean vector and covariance matrix corresponding to the observations in $\mathbf{x}_i$ are denoted as $\boldsymbol{\nu}_i$ and $\mathbf{V}_i$, respectively.

Robust SEM: Step 1. Estimate the robust mean and covariance matrix

- Estimated through solving the following equations

$$\sum_{i=1}^{N} \omega_{i1}\left(d_i\right) \frac{\partial \boldsymbol{\nu}_i^{'}}{\partial \boldsymbol{\nu}} \mathbf{V}_i^{-1}\left(\mathbf{x}_i - \boldsymbol{\nu}_i\right) = 0$$

$$\sum_{i=1}^{N} \frac{\partial vec'\left(\mathbf{V}_i\right)}{\partial \boldsymbol{v}} \mathbf{W}_i vec\left[\omega_{i2}\left(d_i\right)\left(\mathbf{x}_i - \boldsymbol{\nu}_i\right)\left(\mathbf{x}_i - \boldsymbol{\nu}_i\right)^{'} - \omega_{i3}\left(d_i\right)\mathbf{V}_i\right] = 0$$

- d_i is the Mahalanobis distance (M-distance), defined by

$$d_i^2 = d^2(\mathbf{x}_i, \boldsymbol{\nu}_i, \mathbf{V}_i) = \left(\mathbf{x}_i - \boldsymbol{\nu}_i\right)^{'} \mathbf{V}_i^{-1}\left(\mathbf{x}_i - \boldsymbol{\nu}_i\right),$$

$\omega_{i1}\left(d_i\right)$, $\omega_{i2}\left(d_i\right)$ and $\omega_{i3}\left(d_i\right)$ are non-increasing weight functions of d_i.

- The tuning parameter $\varphi, 0 < \varphi < 1$. It is also the down-weighting rate, balancing the estimates' efficiency and protection against data contamination.

- The value of ρ_i is the $(1-\varphi)$ quantile corresponding to the chi-distribution with q_i degrees of freedom, χ_{q_i}. The Huber-type weight functions with missing data are given by

$$
\begin{aligned}
\omega_{i1}(d_i) &= \begin{cases} 1, & if\ d_i \le \rho_i \\ \rho_i/d_i, & if\ d_i > \rho_i \end{cases}, \\
\omega_{i2}(d_i) &= \left[\omega_{i1}(d_i)\right]^2/\kappa_i, \\
\omega_{i3}(d_i) &= 1,
\end{aligned}
$$

where κ_i is a constant defined by $E\left[\chi_{q_i}^2 \omega_{i1}^2\left(\chi_{q_i}^2\right)/\kappa_i\right] = q_i$.

- For complete data,

$$\hat{\boldsymbol{\mu}} = \frac{1}{\sum_{i=1}^{n} w_1(d_i)} \sum_{i=1}^{n} w_1(d_i)$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} w_2(d_i)(\mathbf{y}_i - \hat{\boldsymbol{\mu}})(\mathbf{y}_i - \hat{\boldsymbol{\mu}})'$$

Robust SEM: Step 2. Fit SEM

- Fit $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ by any structural model. Let $\boldsymbol{\mu}(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ be the structural model satisfying $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ represents all the parameters in the model. The estimates $\hat{\boldsymbol{\theta}}$ are obtained by minimizing

$$
\begin{aligned}
F_{ML}(\boldsymbol{\theta}) &= [\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}(\boldsymbol{\theta})]' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) [\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}(\boldsymbol{\theta})] + tr\left[\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\right] \\
&\quad - log\left|\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\right| - p
\end{aligned}
$$

- Robust standard errors can be obtained.

$$
\hat{\boldsymbol{\Omega}} = \left(\hat{\dot{\boldsymbol{\delta}}}'\hat{\mathbf{W}}_{\delta}\hat{\dot{\boldsymbol{\delta}}}\right)^{-1}\left(\hat{\dot{\boldsymbol{\delta}}}'\hat{\mathbf{W}}_{\delta}\hat{\boldsymbol{\Gamma}}\hat{\mathbf{W}}_{\delta}\hat{\dot{\boldsymbol{\delta}}}\right)\left(\hat{\dot{\boldsymbol{\delta}}}'\hat{\mathbf{W}}_{\delta}\hat{\dot{\boldsymbol{\delta}}}\right)^{-1}
$$

- Robust test statistics

- Regular χ^2 statistic T_{ML}

$$T_{ML} = (N-1) \cdot F_{ML}\left(\hat{\boldsymbol{\theta}}\right) \sim \chi^2_{df}$$

- Mean corrected T_{RML}

$$T_{RML} = \hat{m} T_{ML} \sim \chi^2_{df}$$

- Mean and variance corrected T_{AML}

$$T_{AML} = \hat{m}_1 T_{ML} \sim \chi^2_{m_2}$$

- Corrected RADF (CRADF) statistic

$$T_{CRADF} = \frac{T_{RADF}}{1 + \mathbf{r}'\hat{\mathbf{Q}}\mathbf{r}} \sim \chi^2_{df}$$

- Residual-based F-statistic, T_{RF}

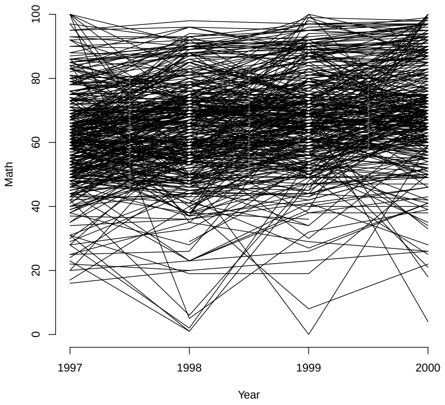
$$T_{RF} = \frac{(N-df)T_{RADF}}{(N-1)df} \sim F_{df,(N-df)}$$

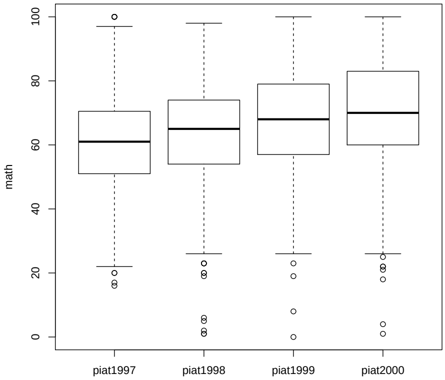
An example

- Longitudinal data from the National Longitudinal Survey of Youth 1997 Cohort (NLSY97) data on Peabody Individual Achievement Test (PIAT) mathematics test scores.
- N=399 school children are measured yearly from 1997 to 2000.

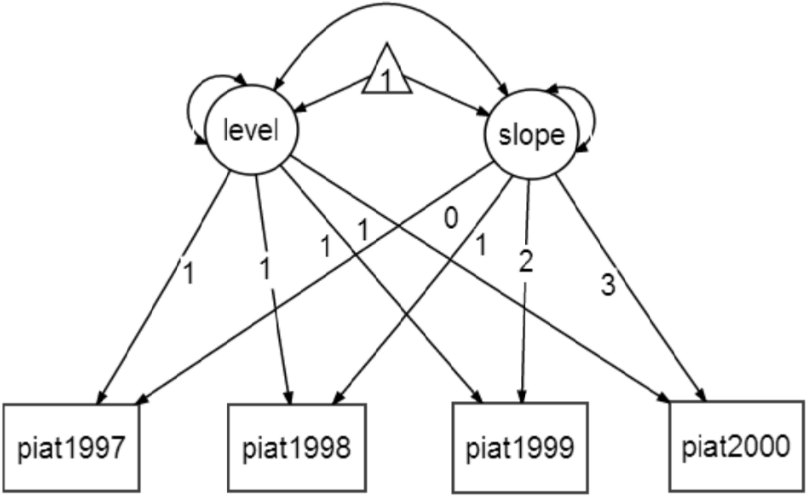
Year	N_C	Mean	SD	Missing rate
1997	375	61.160	15.887	6.015%
1998	377	63.271	17.219	5.514%
1999	357	67.557	16.649	10.526%
2000	350	69.689	17.605	12.281%
Family income	234	17.473	14.844	41.353%
Father's Education	275	12.244	2.860	31.078%
Mother's education	362	12.017	2.615	9.273%

- Plot of the data





A growth curve model



Results

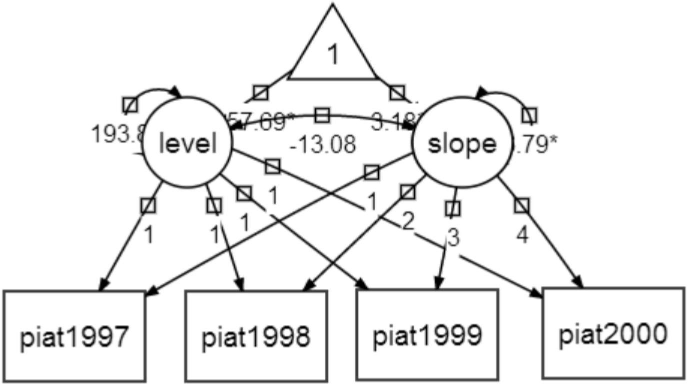
- Fit statistics

	2-stage NML ($\varphi = 0\%$)		2-stage Robust ($\varphi = 10\%$)	
	statistic	p-value	statistic	p-value
T_{ML}	20.282	.001	12.386	.030
T_{RML}	14.124	.015	9.181	.102
T_{AML}	11.448	.023	8.179	.111
T_{CRADF}	11.672	.040	7.948	.159
T_{RF}	2.381	.038	1.606	.157

- Parameter estimates and standard errors

	Robust ($\varphi = .1$)			NML		
θ	$\hat{\theta}$	SE	z	$\hat{\theta}$	SE	z
τ_1	60.865	0.784	77.622	60.645	0.790	76.72
τ_2	3.177	0.251	12.637	3.1	0.272	11.404
ϕ_{11}	174.45	19.254	9.060	177.49	24.59	7.218
ϕ_{21}	-6.290	4.904	-1.283	-4.938	7.644	-0.644
ϕ_{22}	6.791	2.746	2.473	6.994	4	1.748
ψ_{11}	62.406	13.87	4.499	87.576	25.896	3.382
ψ_{22}	77.177	9.105	8.476	103.477	14.275	7.249
ψ_{33}	73.794	9.818	7.516	90.147	14.391	6.264
ψ_{44}	72.463	17.173	4.22	109.889	27.734	3.962

- The path diagram



WebSEM

- Integration of R, LaTeX, PHP, Javascript, etc to conduct SEM analysis online.
- SPSS-like interface for typical data analysis.
- AMOS-like interface with R robust SEM support.
- Accessible through a web browser.
- More suitable for big data.
- The essential features of WebSEM will be illustrated using the growth curve model.

Registration

- URL: `https://websem.psychstat.org`
- Registration is required except for some WebSEM apps so that
  - A user can save and retrieve analysis online.
  - A user can share analysis with others.
  - A user's data can be protected.
  - The abuse of WebSEM can be avoided.
  - Users can better communicate with each other.
- Registration information is verified manually and can be turned down if no sufficient information is provided.
- After registration, one can log in to use WebSEM.

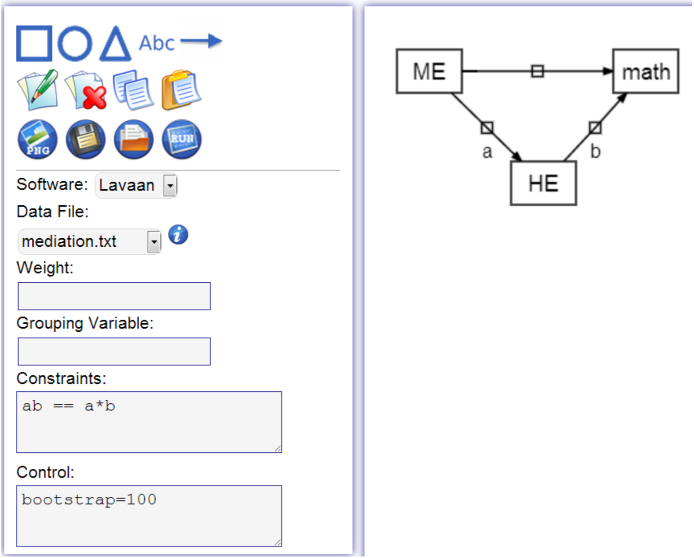
Use WebSEM

- Build a path diagram directly
  - Click the Path Diagram button.
  - The data feature
- Generate a path diagram using equations
  - The Diagram It button.
- Save the path diagram
- Edit a path diagram
- Run the analysis
- Read the output

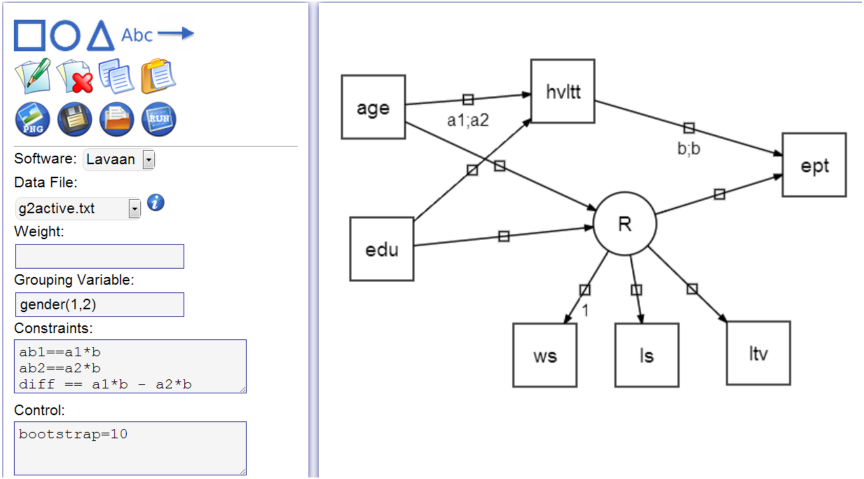
Examples

- Robust Cronbach's alpha and McDonald's Omega `http://www.youtube.com/watch?v=rdj1x_N3Rp4`
- Robust growth curve analysis `https://www.youtube.com/watch?v=GaRk3PmrBDo`

- Mediation analysis using bootstrap https://www.youtube.com/watch?v=1bAsPum98DY



- Multiple group analysis https://www.youtube.com/watch?v=kLLNri-THy0



An incomplete list of WebSEM features

- Drawing path diagrams
  - Interactive drawing
  - Generate from equations
  - Generate from dot (graphviz) file
  - Save, export, and edit
- SEM analysis through rsem and Lavaan
  - Missing data and non-normal data simultaneously
  - Automatic bootstrap
  - Categorical SEM
  - Multiple group analysis

- Mediation analysis

- Other features
  - Sharing
  - WebDav
  - SPSS-like interface for simple data analysis and graphs
  - Edit and run R online
  - Edit and run LaTeX online
  - Wiki and Questions & Answers

Road map

- Robust multiple group analysis
- Robust categorical data analysis
- Scalable vector graphs
- Separated web server, storage server, and computing server
- Incorporation of dropbox, google drive, etc

Q & A

- For more information: `https://websem.psychstat.org/wiki/workshop/index`
- We appreciate any form of feedback.
  - `https://websem.psychstat.org/wiki/workshop/feedback`
  - Contact: Zhiyong Zhang (zzhang4@nd.edu); Ke-Hai Yuan (kyuan@nd.edu).
- Thanks to Institute for Scholarship in the Liberal Arts, Center for Creative Computing, and Center for Research Computing at the University of Notre Dame for support.