Research Journal of Applied Sciences, Engineering and Technology 7(20): 4182-4186, 2014
DOI:10.19026/rjaset.7.785
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: June 28, 2012 Accepted: August 28, 2012 Published: May 20, 2014

Research Article

An Overview on R Packages for Structural Equation Modeling

[1]Haibin Qiu, [2]Yanan Song and [1]Tingdi Zhao
[1]School of Reliability and Systems Engineering, Beihang University, Beijing 100191, China
[2]Department of Industrial Engineering, Tsinghua University, Beijing 100084, China

Abstract: The aim of this study is to present overview on R packages for structural equation modeling. Structural equation modeling, a statistical technique for testing and estimating causal relations using an amalgamation of statistical data and qualitative causal hypotheses, allow both confirmatory and exploratory modeling, meaning they are matched to both hypothesis testing and theory development. R project or R language, a free and popular programming language and computer software surroundings for statistical computing and graphics, is popularly used among statisticians for developing statistical computer software and data analysis. The major finding is that it is necessary to build excellent and enough structural equation modeling packages for R users to do research. Numerous packages for structural equation modeling of R project are introduced in this study and most of them are enclosed in the Comprehensive R Archive Network task view Psychometrics.

Keywords: Psychometrics, R project, structural equation modeling

INTRODUCTION

Structural Equation Modeling (SEM) is a statistical technique for testing and estimating causal relations using a amalgamation of statistical data and qualitative causal hypotheses, which allow both confirmatory and exploratory modeling, meaning they are matched to both theory testing and theory development (Fornell and Larcker, 1981; Sobel, 1982; Hayduk, 1987; Hatcher, 1994; Bollen, 1998; Tomas *et al*., 1999; Schumacker and Lomax, 2004). R project or R language, free and popular programming language and computer software environment for statistical computing and graphics, consists of a lot of packages created by users. It is widely used among statisticians for developing statistical software and data analysis (Ripley, 2001; Ihaka and Gentleman, 1996).

R packages can be browse and installed from CRAN Task Views. The applications of these "Task views" are widely including Finance, Genetics, Machine Learning, Medical Imaging, Social Sciences and Spatial statistics, etc. Psychometricians have also worked collaboratively with the Comprehensive R Archive Network (CRAN) task view Psychometrics (psychometric models and methods), concerned with the design and analysis of investigation and the measurement of human characteristics, in the field of statistics and quantitative methods to develop ameliorated ways to organize and analyze data. There are numerous packages for structural equation modeling in the CRAN task view Psychometrics.

The objective of this study is to review R packages for structural equation modeling and related studies. It is a necessary work for building excellent and enough SEM packages for R users to do research.

STEP BY STEP: HOW TO USE R PACKAGES FOR STRUCTURAL EQUATION MODELING

R package SEM contains functions for fitting general linear structural equation models (with observed and unobserved variables) using the Revised Analog Method (RAM) (Waelbroeck *et al*., 1998) for fitting structural equations in the models of observed variables by two-stage least squares (Fox, 2006). SEM is an R package for complementary structural-equation models (Fox, 2006). The package upholds general structural equation models with latent variables, fit by maximum likelihood assuming multi-normality and single-equation estimation for observed-variable models by two-Stage Least Squares (2SLS) (Bollen, 1996; Chen and Portnoy, 1996; Bollen, 1998; Fox, 2006). The SEM package is available on the Comprehensive R Achieve Network (CRAN) (Fox, 2006). The SEM package fits overall (i.e., latent-variable) structural equation models by Full Information Maximum Likelihood (FIML) (Enders, 2001) and structural equations in observed-variable models by 2SLS (Fox, 2006). An interface between the EQS computer software for structural equation modeling (Bentler, 1985; Bentler and Wu, 1993;

Corresponding Author: Tingdi Zhao, School of Reliability and Systems Engineering, Beihang University, Beijing 100191, China

Bentler, 1995; Mueller, 1996) and R project is granted by the REQS package (Mair *et al*., 2010).

Model specification using R packages: Model specification is the most important step in the structural equation modeling analysis (Hayduk, 1987; Bollen, 1998; Schumacker and Lomax, 2004). When structural equation modeling is used, the model must be specified correctly based on the sufficient theoretical basis (Hayduk, 1987). The model is consistent with two types of variables, named exogenous and endogenous variables. Endogenous variables regress on exogenous variables. However, both endogenous and exogenous can be used as a variable to be regressed on (Bollen, 1998). Structural equation modeling consists of two main components (Hayduk, 1987; Bollen, 1998; Schumacker and Lomax, 2004):

- The structural model:

$$\eta = B\eta + \Gamma\xi + \zeta$$

- The measurement model:

$$x = \Lambda_x\xi + \delta$$

$$y = \Lambda_y\eta + \varepsilon$$

where,

B and Γ	: Both the path coefficient
B	: The relationship between the endogenous latent variables
Γ	: The effluence from the exogenous latent variables to endogenous latent variables ζ is the error
x	: Exogenous variable
y	: Endogenous variable
Λ_x and Λ_y	: Factors loading matrix
δ and ε	: The error terms (Hayduk, 1987; Bollen, 1998; Schumacker and Lomax, 2004)

The structural model gives the potential causal relationships between endogenous and exogenous variables and the measurement model shows the relations between latent variables and their indicators (Hayduk, 1987). In specifying pathways structural equation modeling, there are two types of relationships. One is "free" pathways, which is not verified. Researchers hypothesize causal relationships between variables and test them. The other is "fixed" pathways which mean the relationships between variables are already estimated by previous studies. When model specification is completed, model identify must be estimated according to the number of data points (observing points) and the number of parameters. An identified model is a model which identifies unusually with a particular parameter value and no other parallel patterns can be given by a specific parameter value (Schumacker and Lomax, 2004). The model is unidentified, when there are not sufficiently data points to account for all the variance for the model, for instance, when there are fewer data points than the number of estimated parameters. In most of the literature and analysis tools of structural equation modeling, it is plotted as a path diagram. This diagram is composed of observed variables which are Roman letters enclosed in rectangles and unobserved variables which are Greek letters enclosed in ellipses and circles. The directed arrows which are labeled with Greek letters designate regression coefficients. And the bidirectional arrows signify covariance. The two bidirectional arrows with broken lines represent the covariance which is not included in an initial model specified for these data (Hayduk, 1987; Bollen, 1998).

Model specification in the SEM R package can be handled most conveniently via the "specify.Model". We give an example from Fox *et al*. (2012):

```
model.dhp<- specify Model () RIQ ->
ROccAsp, gam51, NARSES ->
ROccAsp, gam52, NAFSES ->
FOccAsp, gam63, NAFIQ ->
FOccAsp, gam64, NAFOccAsp ->
ROccAsp, beta56, NAROccAsp ->
FOccAsp, beta65, NAROccAsp<->
ROccAsp, ps55, NAFOccAsp<->
FOccAsp, ps66, NAROccAsp<->
FOccAsp, ps56, NA
```

This specification is concise and clear, but there are some difficulties should be noted, which can be read from Fox (2006). As general users, we just require to input these entries like above.

To give estimation, the covariance between the observed variables should be computed. The covariance matrix is entered in a lower-triangular form. Actually, SEM accepts a lower triangular, upper triangular and symmetric covariance matrix. The names of the observed variables can be entered as the first line of the codes. In the model specification that latent variables do not appear in the input covariance matrix. So we must be careful in naming the variables, if we choose the wrong name of observed variables, it may be interpreted as a latent variable, then we may produce an erroneous model (Fox *et al*., 2012; Fox, 2006).

After entering the "specify.Model" and standardized coefficients, the model can be calculated. The outcomes contain the model assessment results by a couple of guide lines and normalized residuals; parameter estimates results and so on. From these outcomes, we can give an introductory deduction for the structural equation modeling analysis (Fox *et al*., 2012; Fox, 2006).

Total, direct and indirect effects for structural equation models: The SEM method for the standard generic function effects computes total, direct and indirect effects for a fitted structural equation model according to the method illustrated in Fox *et al*. (2012). 'Effect.sem' returns an object of class 'semeffects' with 'Total', 'Direct' and 'Indirect elements'. After the operation of rearmost subsection, we can input codes to get output results about the total effects, direct effects and indirect effects.

All of the effects give us the relation of all the variables. According to the theory of structural equation modeling, Total effects = Direct effects + Indirect effects (Fox *et al*., 2012; Fox, 2006).

Estimation of free parameters: Parameter estimation is done through a comparison of the actual covariance matrices and the estimated covariance matrices (Chen and Portnoy, 1996; Enders, 2001). This is also can be accomplished by using R package of structural equation model. Categorical variables in structural equation models can be accommodated via the 'polycor' package (Fox, 2009). The 'systemfit' package implements a wider variety of estimators for observed-variables models, including non-linear simultaneous-equations models (Henningsen and Hamann, 2007). The package 'lavaan' can be used to estimate a large variety of multivariate statistical models, including path analysis, confirmatory factor analysis and structural equation modeling and growth curve models (Rosseel, 2010). It includes the 'lavaan' model syntax, which allows users to express their models in a compact way and allows for Maximum Likelihood (ML), Generalized Least Squares (GLS), Weighted Least Squares (WLS), robust Maximum Likelihood (robust ML) using Satorra-Bentler corrections and FIML for data with missing values (Rosseel, 2010). It completely supports for mean-structures and multiple group send reports standardized solutions, fit measures, modification indices and more as output (Rosseel, 2010).

The 'OpenMX' package, enabling estimation of a wide multiplicity of advanced multivariate statistical models, consists of a library of functions and optimizers that allow you to quickly and flexibly define a structural equation model and estimate parameters given observed data (Boker *et al*., 2011).

Assessment of fit: Assessment of fit is a fundamental task in structural equation modeling. The output of structural equation modeling analysis includes matrices of the estimated relationships between variables. And the fit assessment is to calculate the similarity of the foretold data to the actual data.

There are a number of formal statistical tests and fit indices for these purposes which are used popularly. SEM tests are similar to all statistical hypothesis tests which are based on assumption. Until now, a lot of studies have discussed about fit and lead to different suggestions for the application of the various fit indices and hypothesis tests. There are some fit indices we usually mentioned in SEM R project, for instance Chi-square, RMSEA index, Bentler CFI, SRMR, AIC (Bertossi and Bertossi, 2012). Chi-square is a basic test of model fit which is used for calculating fit measures. Theoretical, it is a function on the sample size and the difference between the observed covariance matrix and the model covariance matrix (Fox *et al*., 2012; Fox, 2006). Root Mean Square Error of Approximation (RMSEA index) is another measure of fit, desirable models are considered to have a RMSEA index of 0.05 or less. Models whose RMSEA index is 0.1 or more have a poor fit. Comparative Fit Index (Bentler CFI) is also a popular fit indicator, when making baseline comparisons examining, which relies on huge part on the mean size of the correlations in the data. The mean correlation between variables and the Bentler CFI are positive correlation. A Bentler CFI value of 0.9 or higher is good. The Standardized Root Mean Residual (SRMR) is also a model fit, which needs to be smaller than 0.05. Akaike Information Criterion (AIC) is a relative measure of fit, which focuses on how little the fitted values diverge from a saturated model (Bertossi and Bertossi, 2012).

In the last few years, various researchers study on the fit assessment and some of them have enhanced R project with packages for structural equation modeling. SEMGOF is an expanded R package which supplies 14 goodness-of-fit indices for SEM (Bertossi and Bertossi, 2012). R package SEMModComp conducts tests of difference in fit for mean and covariance structure models as in SEM (Levy, 2010). General paradigm for carrying on statistical tests on competing mean and covariance structure models is provided by Levy and Hancock (2007). The framework they proposed is appropriate for hierarchically related models as well as non-hierarchically related models. R package SEMModComp is used to statistically compare models in accordance with the framework. R package SEMPLS fits SEM using Partial Least Squares (PLS). The PLS method is mentioned as 'soft-modeling' technique which needs no distributional assumptions on the observed data (Monecke and Leisch, 2012; Graf *et al*., 2012). PLS methods with emphasizes on SEM with latent variables. They are given in 'plspm' within which also includes 'pathmox' serve as a companion package with methods of segmentation trees in PLS path modeling (Sanchez and Trinchera, 2012). The 'plsRglm' package is designed to provide PLS regression and PLS generalized linear regression. It includes various criteria in order to select the number of components, repeated k-fold cross-validation, bootstrap confidence intervals and significance testing (Bertrand *et al*., 2010). R package 'semdiag' is considered as outlier and leverage

diagnostics regarding structural equation modeling (Zhang and Yuan, 2012).

Model modification: In order to improve the fit after assessment, the introductory model needs to be modified. Our objective is to estimate the most likely relationships between all the variables. Modifications that improve the fit for a model are deemed to potential changes. The most important thing in this step is that although the model fit should be improved, being very cautious about every modification is required. Science even a small wrong modification may cause the unconformity between the new model the theoretical sense. Modification indices and estimated parameter changes are calculated by 'mod.indices' for the fixed and constrained parameters which are fitted by multi-normal maximum likelihood (Fox, 2009). As is representative of R programs, SEM returns an object rather than an impressed report (Fox *et al*., 2012; Fox, 2006). The object is returned by 'mod.indices' are simply imprinted, which produce a brief report. The output results of SEM have been shown beforehand. One can perform supplemental calculations on SEM objects (Fox *et al*., 2012; Fox, 2006).

There is a more complete report from summarizing method for these objects. The report shows all modification indexes as well as approximations and estimates. The results would be each omitted parameter contained in the model (Fox, 2009).

After the 'mod.indices' operations, the computer output the results of 5 largest modification indices. The A matrix contains regression coefficients, whereas the P matrix contains covariance. The modification indexes may illuminate us to release some of the covariance in the latent endogenous variables' measuring errors for achieving a better fit to the data. We modify the initial parameters according to these results, then the steps before should be repeated.

CONCLUSION

Compare with those other software for SEM as AMOS, EQS, LISREL, or Mplus, the latent variable modeling facility provided by R project SEM function is relatively basic. Fox (2006) conceive that one possible future direction for the SEM package, consequently, would be to spread out capacities in areas such as multiple-group models and alternative fitting functions. It is displayed that the packages of R project for structural equation modeling are deficient, until now. A lot of further work needs doing. Up to now, there is no user interface for R package SEM, but a path diagram is needed to intuitively illustrate the model. Although Model specification in the SEM package is picked up most conveniently by means of the function "specify Model", software program not as instinctive as the path diagram volunteered by other narrates software (Fox, 2006). To puzzle out this problem, a graphical interface is obligatory to be energized (Fox *et al*., 2012).

REFERENCES

Bentler, P., 1985. Theory and Implementation of EQS: A Structural Educational Program. Manual for Program Version 2.0, BMDP Statistical Software, Los Angeles.

Bentler, P., 1995. EQS Structural Equations Program Manual. Program Version 5.0, Multivariate Software Encino, CA.

Bentler, P.M. and E.C.J. Wu, 1993. EQS-Windows user's guide. BMDP Statistical Software, Los Angeles.

Bertossi, E. and M.E. Bertossi, 2012. Goodness-of-fit Indeces for Structural Equations Models. Package 'semGOF'. Retrieved from: http://www.r-project.org,http://sites.google.com/site/bertossielena.

Bertrand, F., M. Maumy-Bertrand and N. Meyer, 2010. plsRglm, PLS generalised linear models for R. inria-00494857, Version 1, Jun. 24, 2010.

Boker, S., M. Neale, H. Maes, M. Wilde, M. Spiegel, T. Brick, J. Spies, R. Estabrook, S. Kenny, T. Bates, *et al*., 2011. Open MX: An open source extended structural equation modeling framework. Psychometrika, 76: 306-317.

Bollen, K., 1996. An alternative two stage least squares (2sls) estimator for latent variable equations. Psychometrika, 61: 109-121.

Bollen, K., 1998. Structural Equation Models. In: Armitage, P. and T. Colton (Eds.), Encyclopedia of Biostatistics. John Wiley, Sussex, England, pp: 4363-4372.

Chen, L. and S. Portnoy, 1996. Two-stage regression quantiles and two-stage trimmed least squares estimators for structural equation models. Commun. Stat. Theory, 25: 1005-1032.

Enders, C., 2001. The impact of nonnormality on full information maximum-likelihood estimation for structural equationmodels with missing data. Psychol. Methods, 6: 352.

Fornell, C. and D. Larcker, 1981. Evaluating structural equation models with unobservable variables and measurement error. J. Marketing Res., 18: 39-50.

Fox, J., 2006. Teacher's corner: Structural equation modeling with the sem package in R. Struct. Equ. Modeling, 13: 465-486.

Fox, J., 2009. Polycor: Polychoric and Polyserial Correlations. R Package Version 0.7-7. Retrieved from: http://cran.r-project.org/web/packages/polycor/.

Fox, J., Z. Nie, J. Byrnes, M. Culbertson, M. Friendly, A. Kramer and G. Monette, 2012. Package Sem: Structural Equation Models. Version 3.0-0. Retrieved from: http://www.r-project.org, http://socserv.socsci.mcmaster.ca/jfox/.

Graf, A., S. Kaiser and F. Leisch, 2012. semPLS: An R package for structural equation models using partial least squares.

Hatcher, L., 1994. A Step-by-step Approach to Using the SAS System for Factor Analysis and Structural Equation Modeling. SAS Publishing, Cary, Car. du N.

Hayduk, L.A., 1987. Structural Equation Modeling with LISREL: Essentials and Advances. Johns Hopkins University Press, Baltimore.

Henningsen, A. and J.D. Hamann, 2007. Systemfit: A package for estimating systems of simultaneous equations in R. J. Stat. Softw., 23: 1-40.

Ihaka, R. and R. Gentleman, 1996. R: A language for data analysis and graphics. J. Comput. Graph. Stat., 5: 299-314.

Levy, R., 2010. SEMModComp: An R package for calculating likelihood ratio tests for mean and covariance structure models. Appl. Psychol. Meas., 34: 370-371.

Levy, R. and G.R. Hancock, 2007. A framework of statistical tests for comparing mean and covariance structure models. Multivar. Behav. Res., 42(1): 33-66.

Mair, P., E. Wu and P. Bentler, 2010. EQS goes R: Simulations for SEM using the package REQS. Struct. Equ. Modeling, 17: 333-349.

Monecke, A. and F. Leisch, 2012. semPLS: Structural equation modeling using partial least squares. J. Stat. Softw., 48(3): 1-32.

Mueller, R., 1996. Basic Principles of Structural Equation Modeling: An Introduction to LISREL and EQS. Springer, New York.

Ripley, B., 2001. The R project in statistical computing. MSOR Connections, 1: 23-25.

Rosseel, Y., 2010. Lavaan: An R Package for Structural Equation Modeling and More. Version 0.3-1. Ghent University, Belgium.

Sanchez, G. and L. Trinchera, 2012. Package plspm: Title Partial Least Squares Data Analysis Methods. Version 0.2-2. Retrieved from: http://www.plsmodeling.com.

Schumacker, R.E. and R.G. Lomax, 2004. A Beginner's Guide to Structural Equation Modeling. 2nd Edn., Erlbaum., Mahwah, NJ.

Sobel, M., 1982. Asymptotic confidence intervals for indirect effects in structural equation models. Sociol. Methodol., 13: 290-312.

Tomas, J., J. Melia and A. Oliver, 1999. A cross-validation of a structural equation model of accidents: Organizational and psychological variables as predictors of work safety. Work Stress, 13: 49-58.

Waelbroeck, C., L. Labeyrie, J. Duplessy, J. Guiot, M. Labracherie, H. Leclaire and J. Duprat, 1998. Improving past sea surface temperature estimates based on planktonic fossil faunas. Paleoceanography, 13: 272-283.

Zhang, Z. and K.H. Yuan, 2012. Package semdiag: Title Structural Equation Modeling Diagnostics. Version 0.1.2.