Italian Stata User Group – XVIII Conference
Florence, Italy
May 9-10, 2024

Structural Equation Modelling with Partial Least Squares using Stata

Sergio Venturini
sergio.venturini@unicatt.it

Università Cattolica del Sacro Cuore
Via Bissolati 74
26100 Cremona
Italy

Mehmet Mehmetoglu
mehmet.mehmetoglu@ntnu.no

NTNU – Norwegian University of Science and Technology
517, Dragvoll
NO-7491 Trondheim
Norway

Overview

1. What is Partial Least Squares Structural Equation Modeling (PLS-SEM)?
2. The PLS-SEM algorithm
3. The **`plssem`** *Stata* package
4. Future directions

What is PLS-SEM?

- PLS-SEM can be seen as:
  - The partial least squares (PLS) approach to structural equation modeling (SEM)
  - A statistical method for studying complex multivariate relationships among observed and latent variables
  - A data analysis approach for studying blocks of observed variables in which each block can be summarized by a latent variable and linear relations between the latent variables are assumed

What is PLS-SEM?

- PLS-SEM originates from the work of Herman Wold
- In the 1960s and 1970s Wold developed a set of iterative algorithms based on least squares that nowadays are referred to as **partial least squares** (PLS)
- PLS methods encompass a broad spectrum of both explanatory and exploratory multivariate techniques, ranging from regression to path modeling, and from principal component to multi-block data analysis

What is PLS-SEM?

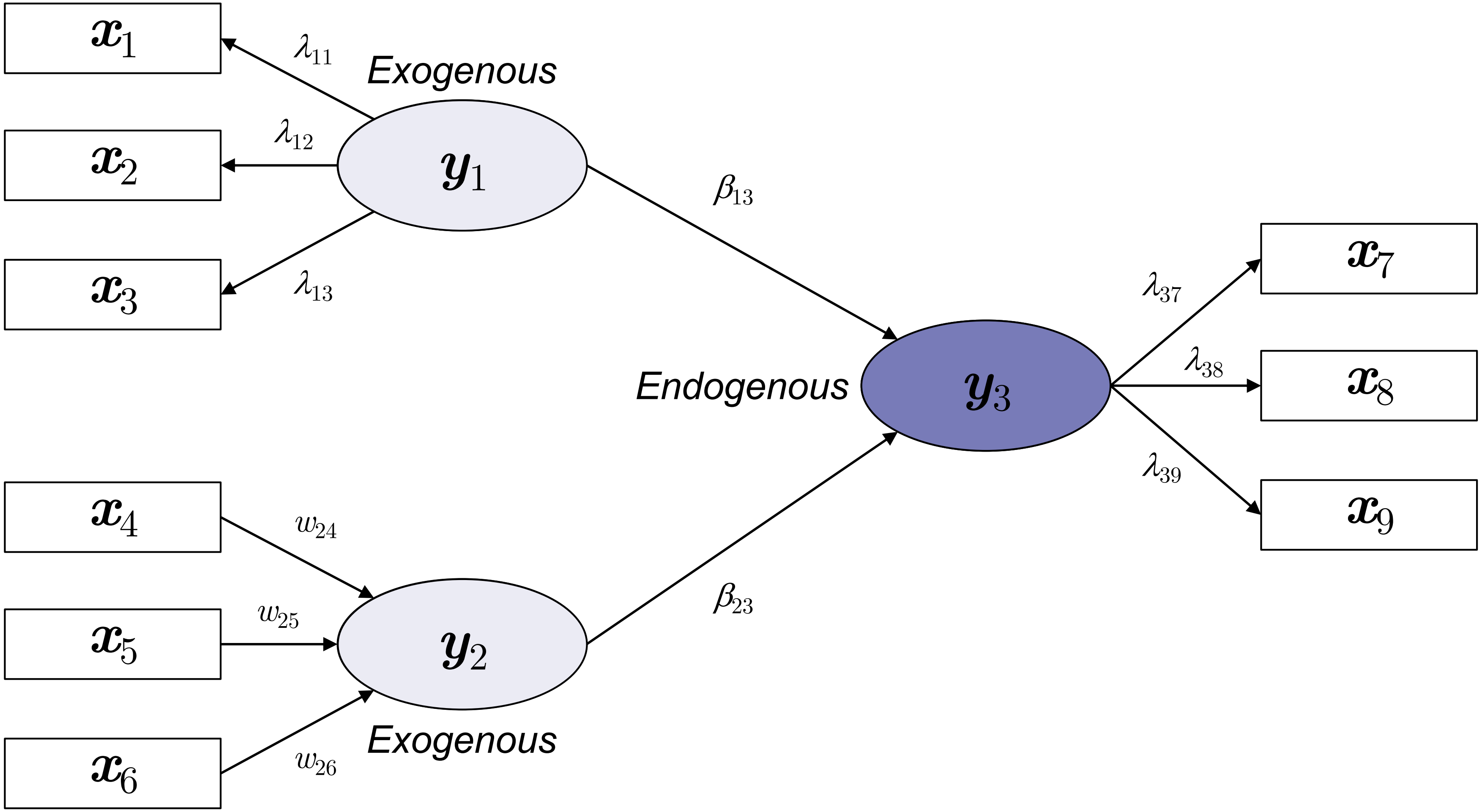
- PLS-SEM is frequently seen as an alternative approach to classical *covariance-based SEM* (COV-SEM):
  - they aim at studying the interdependencies among a set of *unobserved* latent variables (LVs), each of which is measured through a different set of *observed* (or *manifest*) variables (MVs)
  - they involve a *measurement* (or *outer*) *model* relating the latent variables to the corresponding manifest variables, and a *structural* (or *inner*) model providing the relations among the latent variables
  - both are typically specified using a *path diagram*

What is PLS-SEM?

- The main differences between the two approaches are:

COV-SEM	**PLS-SEM**
➢ it aims at reproducing the observed covariance matrix of the manifest variables	➢ it aims at maximizing the explained variance of the endogenous latent variables
➢ the model is estimated using maximum likelihood	➢ the model is estimated using an iterative algorithm that involves ordinary least squares
➢ it is typically used for theory testing	➢ it is typically used for predictive purposes

What is PLS-SEM?



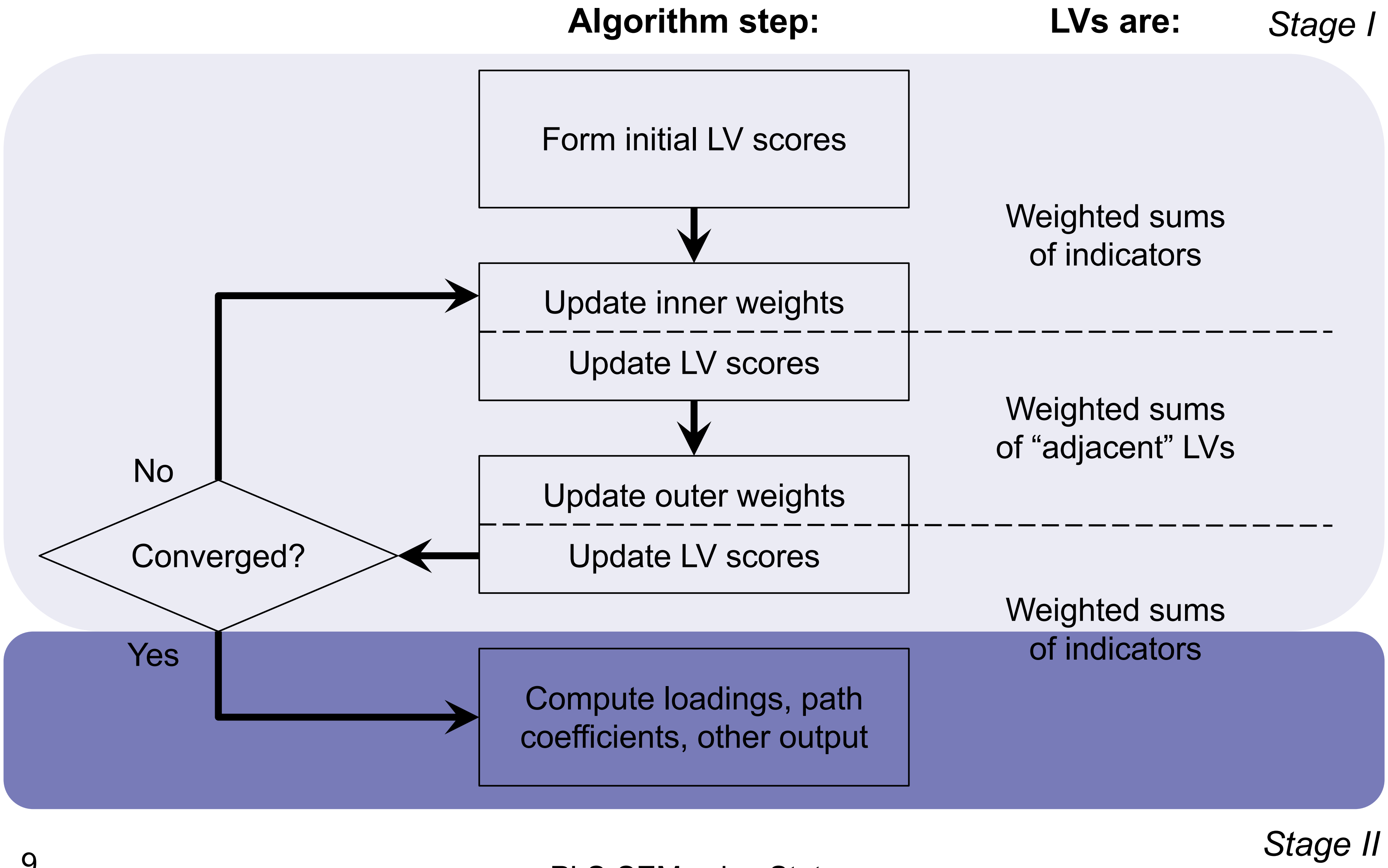
What is PLS-SEM?

- Both the structural and measurement models involve linear specifications:
  - In the structural model a generic endogenous LV y_j is linked to the corresponding latent predictors through the multiple linear regression model

$$y_j = \beta_{0j} + \sum_{m=1}^{M_j} \beta_{jm} y_{m \to j} + \delta_j$$

  - In the measurement model, the relation between each MV x_k and the corresponding LV is generally modeled as
    - *reflective* blocks ➔ $x_k = \lambda_{0k} + \lambda_{jk} y_j + \epsilon_k$
    - *formative* blocks ➔ $y_j = w_{0k} + \sum_h w_{jh} x_h + \zeta_j$

The PLS-SEM algorithm



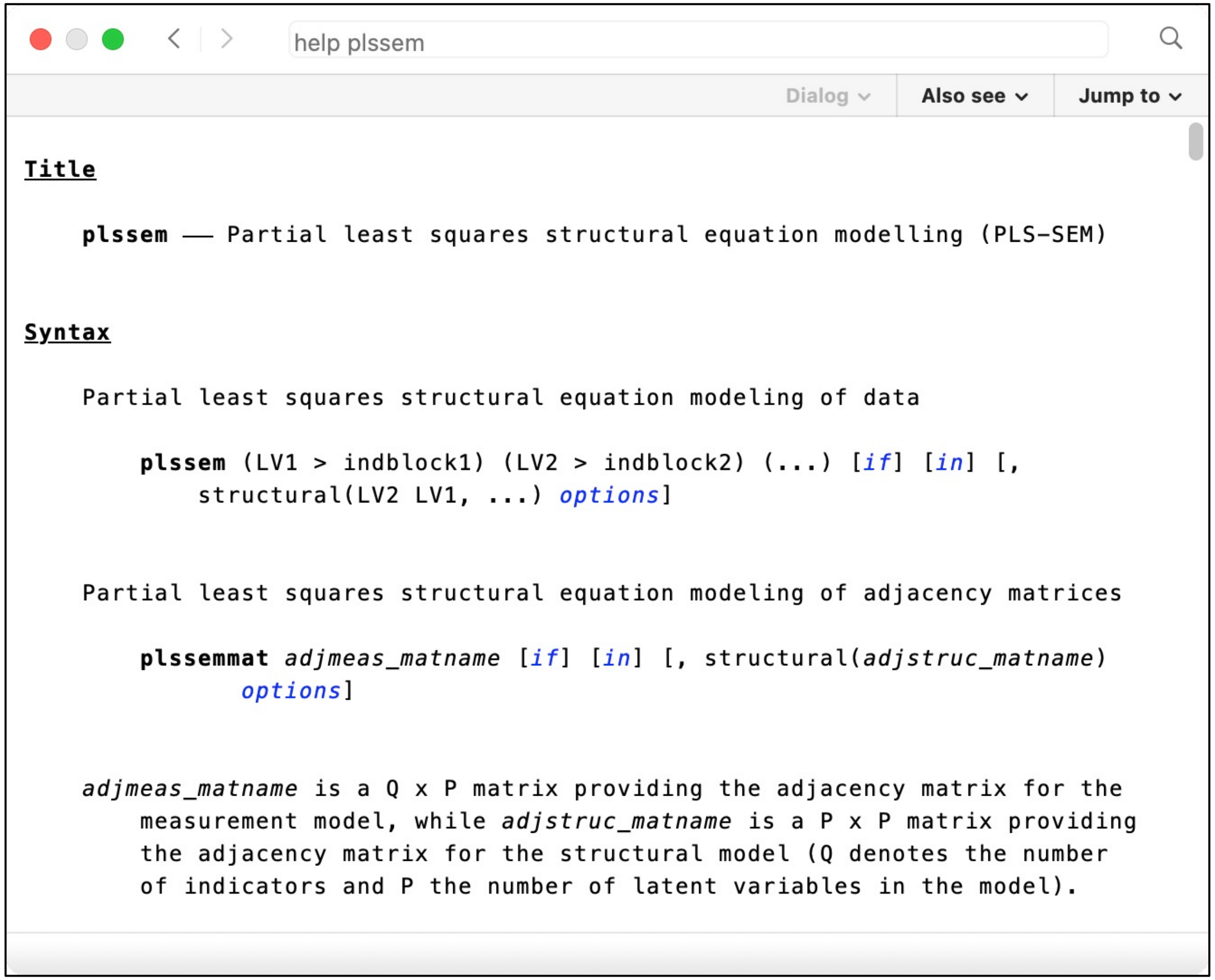
The `plssem` *Stata* package

- Different software packages are available for fitting PLS-SEM models, both commercial (e.g. SmartPLS, ADANCO) and open-source (e.g. `cSEM`, `SEMinR`)
- While Stata has a very nice suite of commands for COV-SEM, nothing is available for PLS-SEM
- To fill the gap, some years ago we started the development of a Stata package for PLS-SEM called **`plssem`**
- The project is open-source and it can be installed from one of the author's GitHub account (https://github.com/sergioventurini/plssem)

The **`plssem`** *Stata* package

- The package provides:
  - estimation commands
    - **`plssem`** ➔ implements the standard PLS-SEM algorithm
    - **`plssemc`** ➔ implements the consistent PLS-SEM (PLSc) algorithm
    - **`plssemmat`** ➔ matrix-based version of **`plssem`**
    - **`plssemcmat`** ➔ matrix-based version of **`plssemc`**
  - post-estimation commands
    - **`estat`** ➔ computes many goodness of fit and diagnostic measures
    - **`plssemplot`** ➔ creates some graphs for visualizing the results
    - **`predict`** ➔ computes the predicted values and residuals

The `plssem` *Stata* package



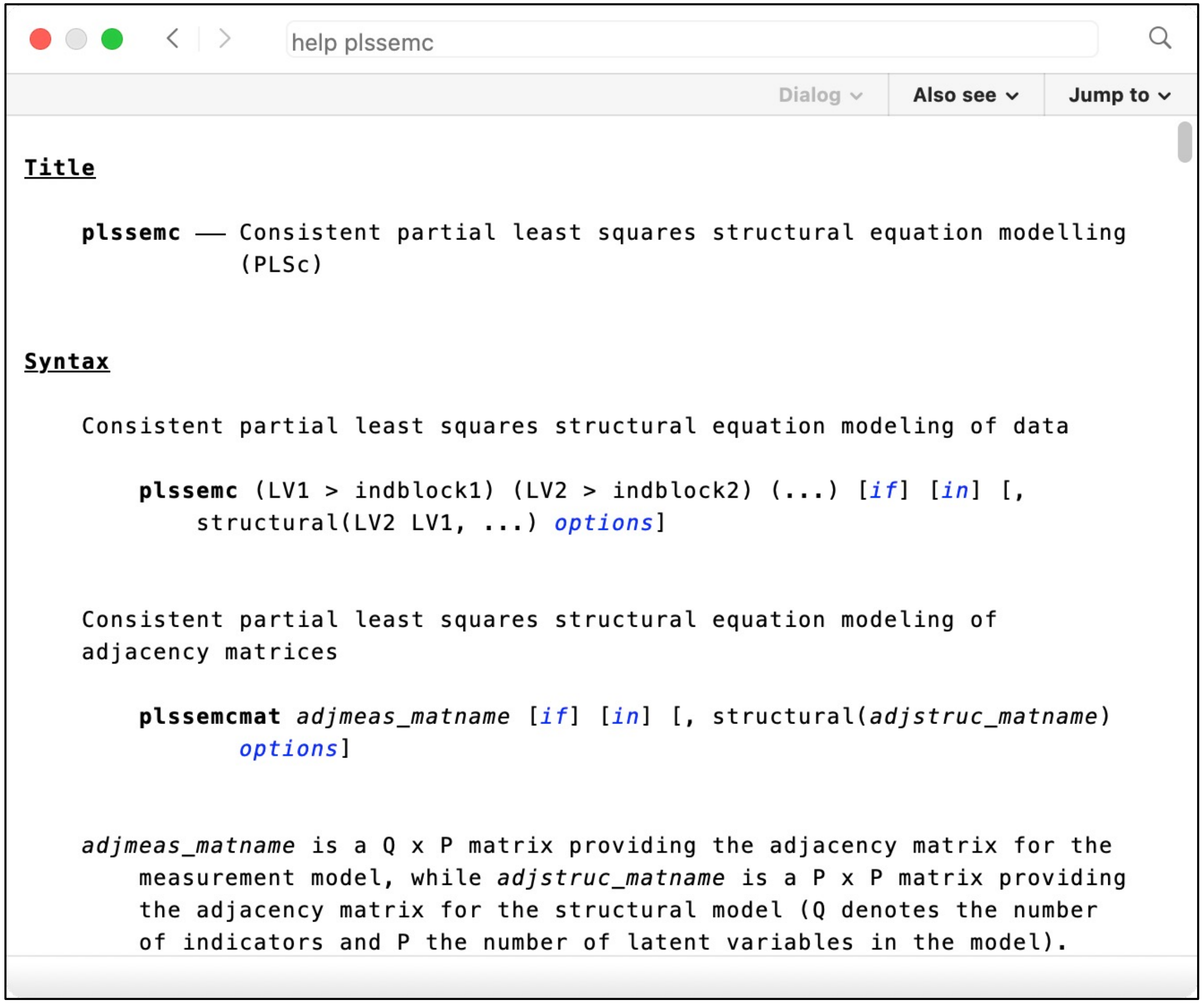
Title

```
    plssem —— Partial least squares structural equation modelling (PLS-SEM)
```

Syntax

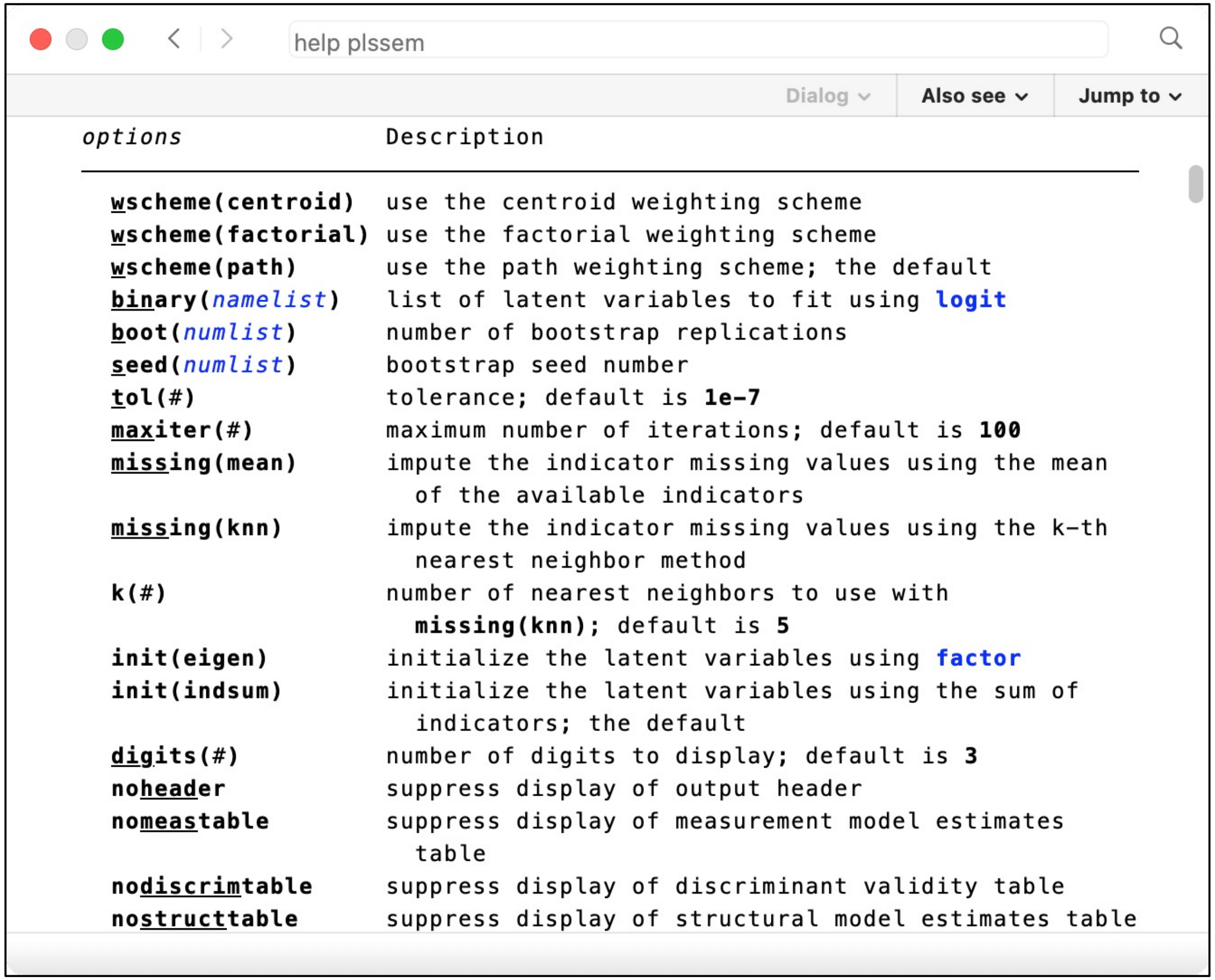
```
    Partial least squares structural equation modeling of data

        plssem (LV1 > indblock1) (LV2 > indblock2) (...) [if] [in] [,
            structural(LV2 LV1, ...) options]


    Partial least squares structural equation modeling of adjacency matrices

        plssemmat adjmeas_matname [if] [in] [, structural(adjstruc_matname)
               options]


    adjmeas_matname is a Q x P matrix providing the adjacency matrix for the
        measurement model, while adjstruc_matname is a P x P matrix providing
        the adjacency matrix for the structural model (Q denotes the number
        of indicators and P the number of latent variables in the model).
```

The **plssem** *Stata* package



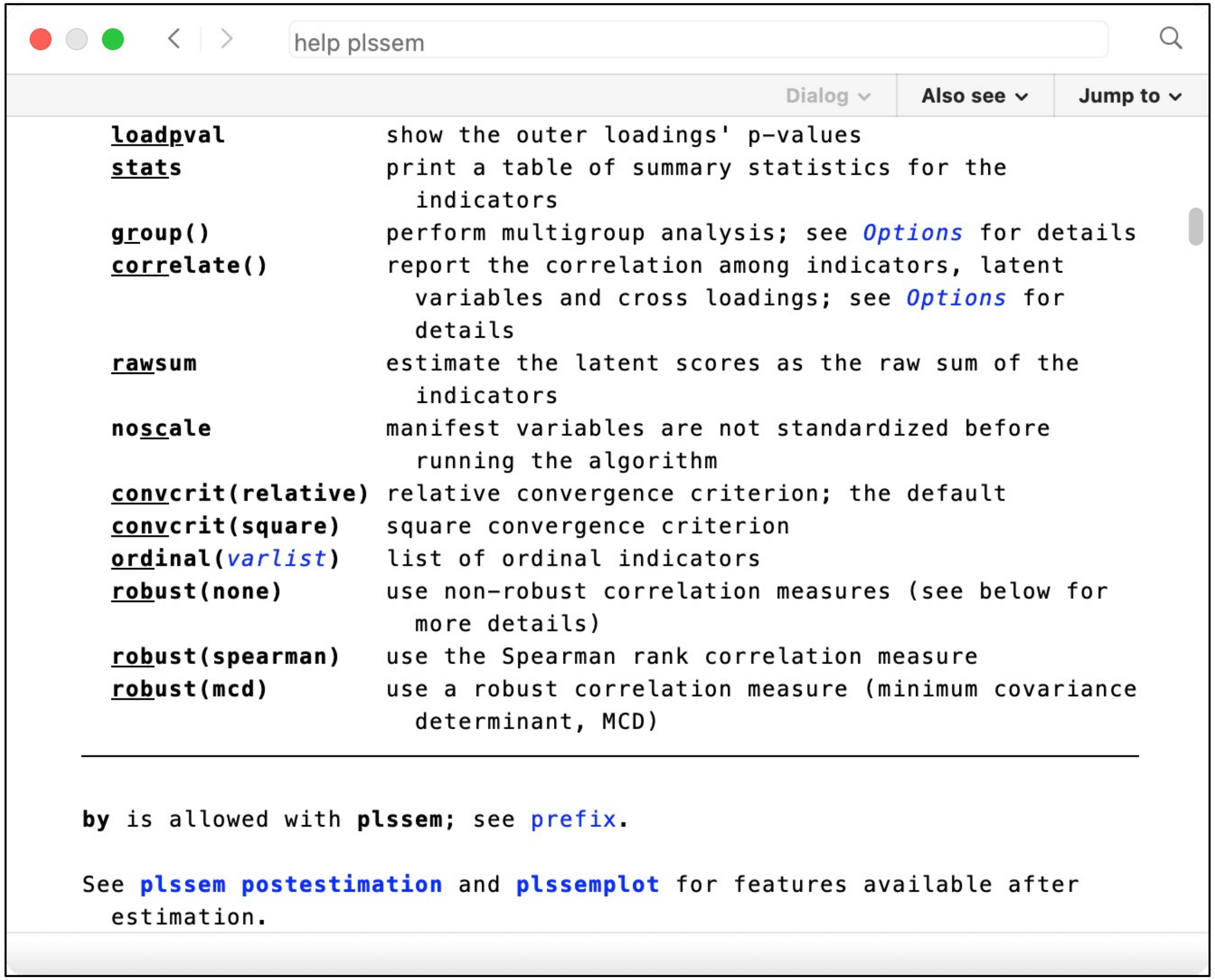
```
Title

    plssemc —— Consistent partial least squares structural equation modelling
               (PLSc)


Syntax

    Consistent partial least squares structural equation modeling of data

        plssemc (LV1 > indblock1) (LV2 > indblock2) (...) [if] [in] [,
            structural(LV2 LV1, ...) options]


    Consistent partial least squares structural equation modeling of
    adjacency matrices

        plssemcmat adjmeas_matname [if] [in] [, structural(adjstruc_matname)
               options]


    adjmeas_matname is a Q x P matrix providing the adjacency matrix for the
        measurement model, while adjstruc_matname is a P x P matrix providing
        the adjacency matrix for the structural model (Q denotes the number
        of indicators and P the number of latent variables in the model).
```

The `plssem` *Stata* package



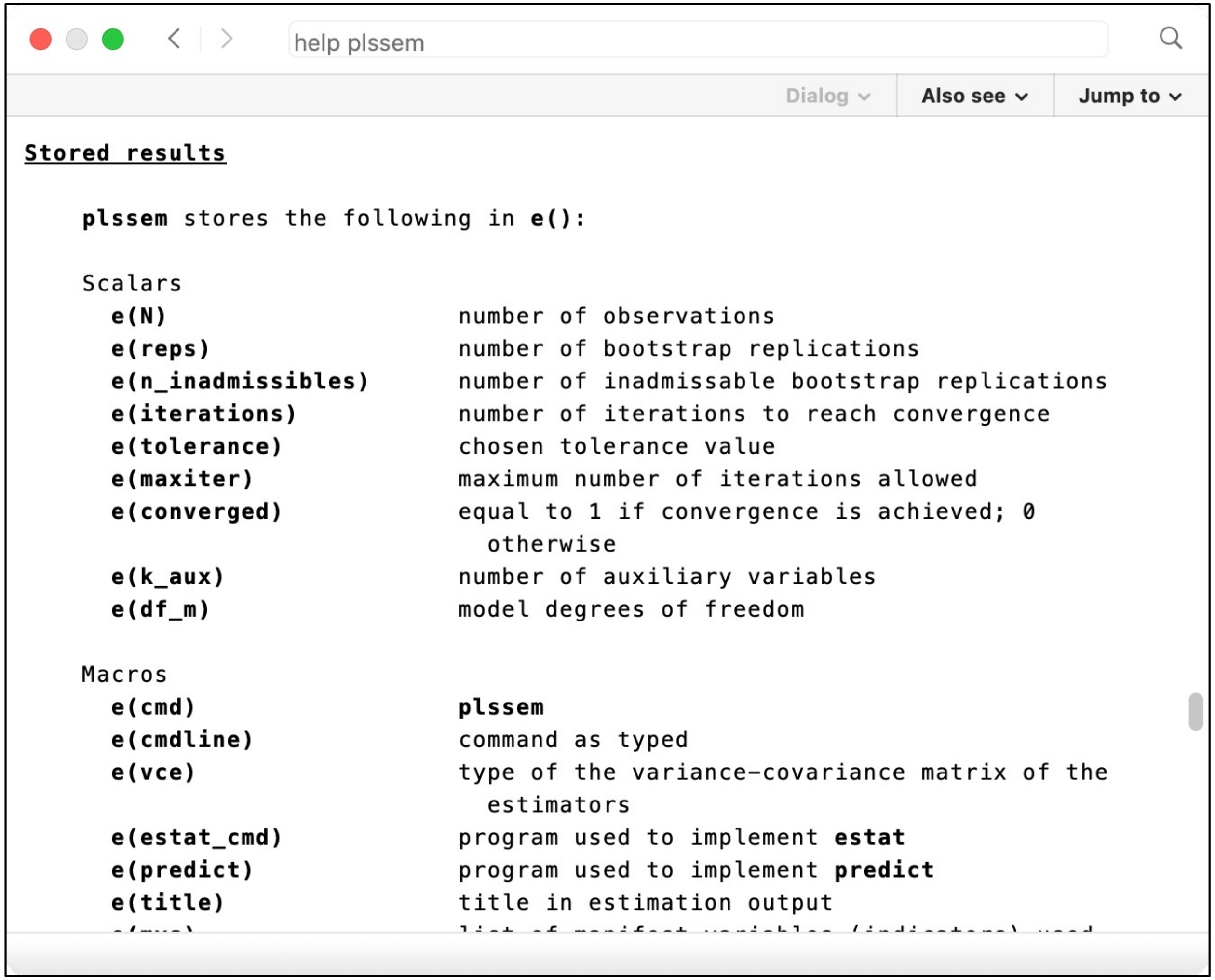
```
options                 Description
---------------------------------------------------------------------------
  wscheme(centroid)   use the centroid weighting scheme
  wscheme(factorial)  use the factorial weighting scheme
  wscheme(path)       use the path weighting scheme; the default
  binary(namelist)    list of latent variables to fit using logit
  boot(numlist)       number of bootstrap replications
  seed(numlist)       bootstrap seed number
  tol(#)              tolerance; default is 1e-7
  maxiter(#)          maximum number of iterations; default is 100
  missing(mean)       impute the indicator missing values using the mean
                        of the available indicators
  missing(knn)        impute the indicator missing values using the k-th
                        nearest neighbor method
  k(#)                number of nearest neighbors to use with
                        missing(knn); default is 5
  init(eigen)         initialize the latent variables using factor
  init(indsum)        initialize the latent variables using the sum of
                        indicators; the default
  digits(#)           number of digits to display; default is 3
  noheader            suppress display of output header
  nomeastable         suppress display of measurement model estimates
                        table
  nodiscrimtable      suppress display of discriminant validity table
  nostructtable       suppress display of structural model estimates table
```

The **plssem** *Stata* package



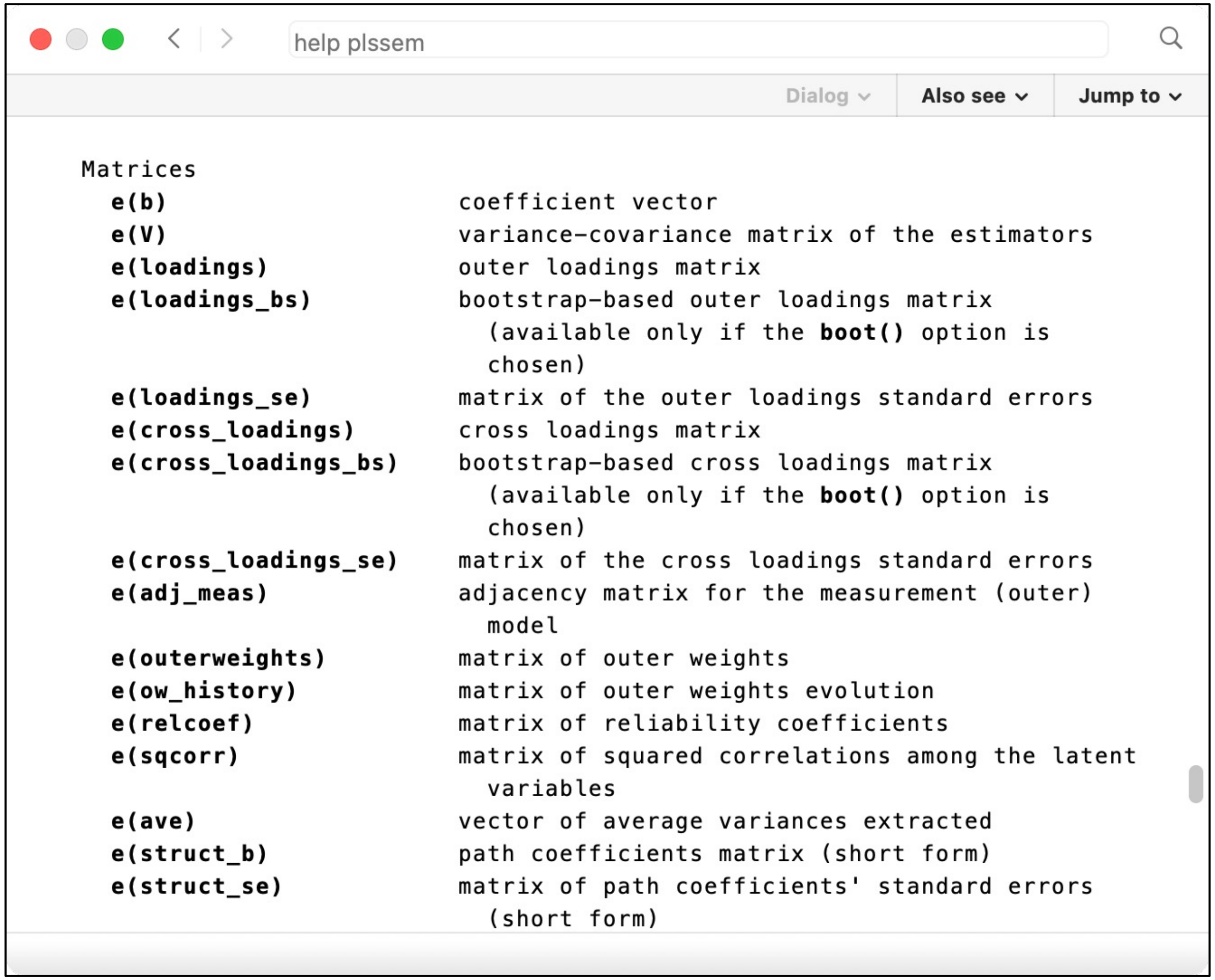
```
    loadpval            show the outer loadings' p-values
    stats               print a table of summary statistics for the
                          indicators
    group()             perform multigroup analysis; see Options for details
    correlate()         report the correlation among indicators, latent
                          variables and cross loadings; see Options for
                          details
    rawsum              estimate the latent scores as the raw sum of the
                          indicators
    noscale             manifest variables are not standardized before
                          running the algorithm
    convcrit(relative)  relative convergence criterion; the default
    convcrit(square)    square convergence criterion
    ordinal(varlist)    list of ordinal indicators
    robust(none)        use non-robust correlation measures (see below for
                          more details)
    robust(spearman)    use the Spearman rank correlation measure
    robust(mcd)         use a robust correlation measure (minimum covariance
                          determinant, MCD)
  ------------------------------------------------------------------------

  by is allowed with plssem; see prefix.

  See plssem postestimation and plssemplot for features available after
    estimation.
```

The **plssem** *Stata* package



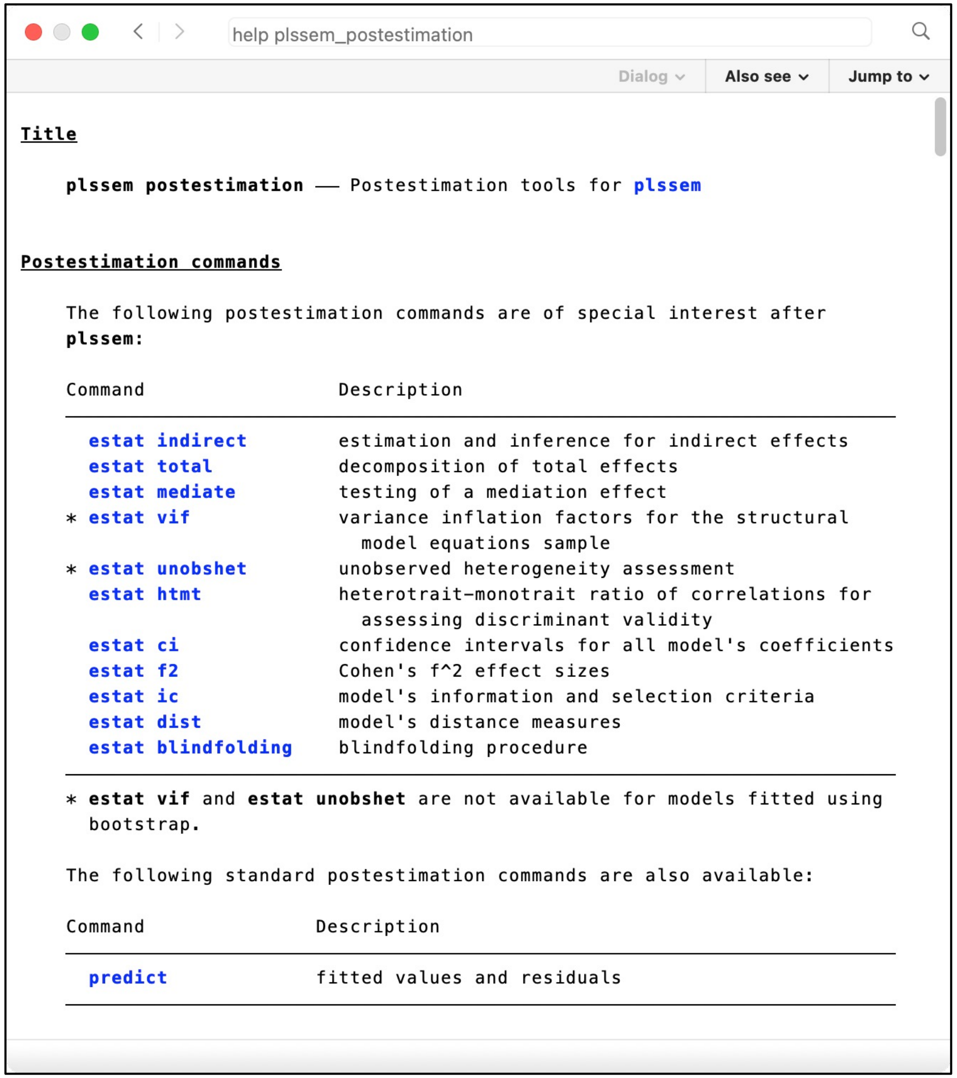
```
Stored results

    plssem stores the following in e():

    Scalars
      e(N)                    number of observations
      e(reps)                 number of bootstrap replications
      e(n_inadmissibles)      number of inadmissable bootstrap replications
      e(iterations)           number of iterations to reach convergence
      e(tolerance)            chosen tolerance value
      e(maxiter)              maximum number of iterations allowed
      e(converged)            equal to 1 if convergence is achieved; 0
                                otherwise
      e(k_aux)                number of auxiliary variables
      e(df_m)                 model degrees of freedom

    Macros
      e(cmd)                  plssem
      e(cmdline)              command as typed
      e(vce)                  type of the variance-covariance matrix of the
                                estimators
      e(estat_cmd)            program used to implement estat
      e(predict)              program used to implement predict
      e(title)                title in estimation output
```

The `plssem` *Stata* package



```
Matrices
  e(b)                   coefficient vector
  e(V)                   variance-covariance matrix of the estimators
  e(loadings)            outer loadings matrix
  e(loadings_bs)         bootstrap-based outer loadings matrix
                           (available only if the boot() option is
                           chosen)
  e(loadings_se)         matrix of the outer loadings standard errors
  e(cross_loadings)      cross loadings matrix
  e(cross_loadings_bs)   bootstrap-based cross loadings matrix
                           (available only if the boot() option is
                           chosen)
  e(cross_loadings_se)   matrix of the cross loadings standard errors
  e(adj_meas)            adjacency matrix for the measurement (outer)
                           model
  e(outerweights)        matrix of outer weights
  e(ow_history)          matrix of outer weights evolution
  e(relcoef)             matrix of reliability coefficients
  e(sqcorr)              matrix of squared correlations among the latent
                           variables
  e(ave)                 vector of average variances extracted
  e(struct_b)            path coefficients matrix (short form)
  e(struct_se)           matrix of path coefficients' standard errors
                           (short form)
```

The `plssem` *Stata* package



help plssem_postestimation

Title

plssem postestimation — Postestimation tools for **plssem**

Postestimation commands

The following postestimation commands are of special interest after **plssem**:

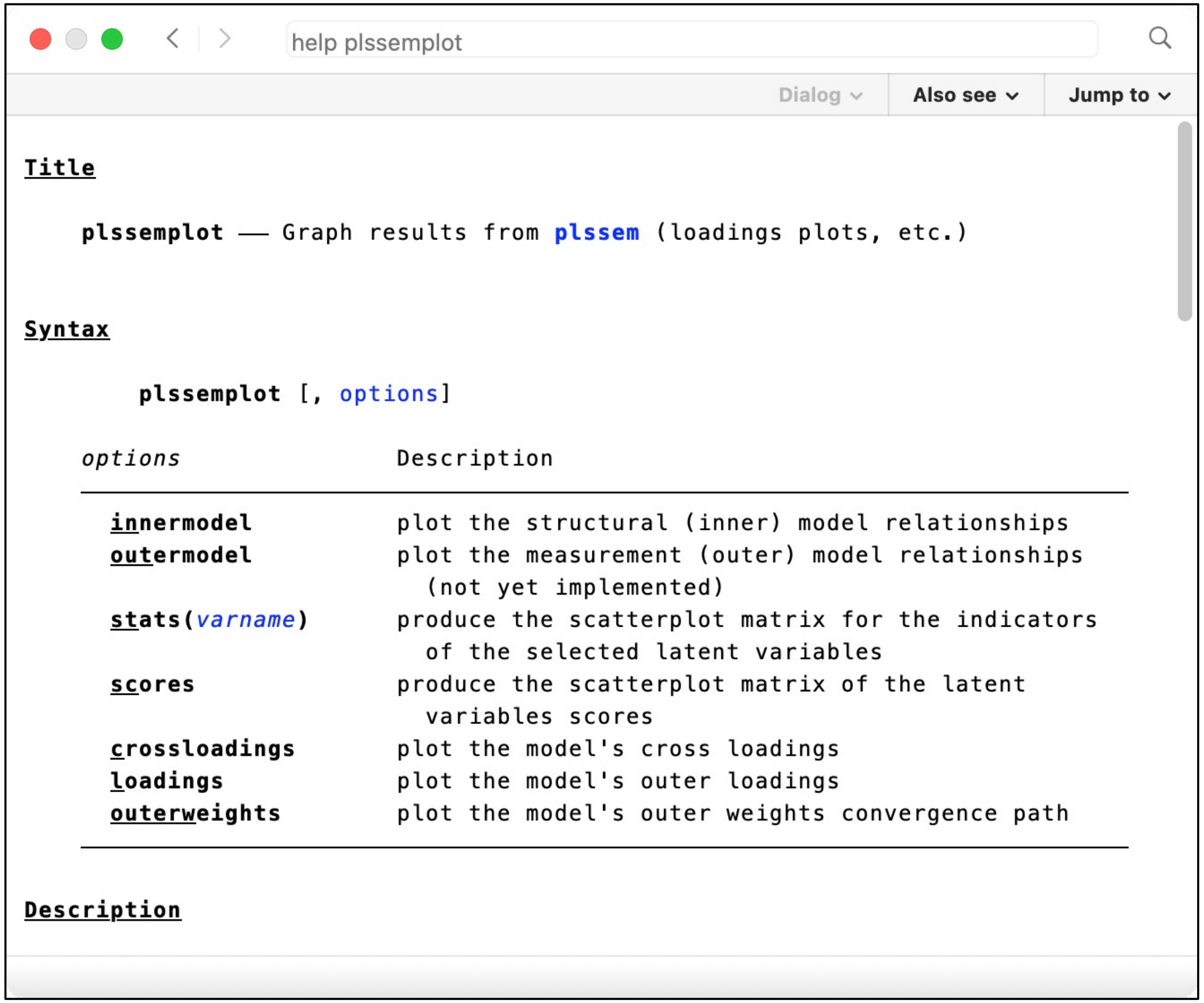
Command	Description
estat indirect	estimation and inference for indirect effects
estat total	decomposition of total effects
estat mediate	testing of a mediation effect
* **estat vif**	variance inflation factors for the structural model equations sample
* **estat unobshet**	unobserved heterogeneity assessment
estat htmt	heterotrait-monotrait ratio of correlations for assessing discriminant validity
estat ci	confidence intervals for all model's coefficients
estat f2	Cohen's f^2 effect sizes
estat ic	model's information and selection criteria
estat dist	model's distance measures
estat blindfolding	blindfolding procedure

* **estat vif** and **estat unobshet** are not available for models fitted using bootstrap.

The following standard postestimation commands are also available:

Command	Description
predict	fitted values and residuals

The **plssem** *Stata* package



Title

plssemplot — Graph results from **plssem** (loadings plots, etc.)

Syntax

```
        plssemplot [, options]

    options                    Description
    ------------------------------------------------------------------------
      innermodel               plot the structural (inner) model relationships
      outermodel               plot the measurement (outer) model relationships
                                 (not yet implemented)
      stats(varname)           produce the scatterplot matrix for the indicators
                                 of the selected latent variables
      scores                   produce the scatterplot matrix of the latent
                                 variables scores
      crossloadings            plot the model's cross loadings
      loadings                 plot the model's outer loadings
      outerweights             plot the model's outer weights convergence path
    ------------------------------------------------------------------------
```

Description

Future directions

- We continue actively developing the package and we are planning to expand it in different directions:
  - moderated mediation
  - nonlinear effects in the structural model
  - multiple imputation
  - graphical interface to interactively specify the entire model, similar to Stata's **`sembuilder`** for COV-SEM ➔ *call for collaborations!!!*

References

1. Esposito Vinzi, V., Russolillo, G. 2013. Partial least squares algorithms and methods. *WIREs Computational Statistics*, 5, 1-19.
2. Esposito Vinzi, V., Trinchera, L., Squillacciotti, S., Tenenhaus, M. 2008. REBUS-PLS: a response-based procedure for detecting unit segments in PLS path modeling. *Applied Stochastic Models in Business and Industry*, 24, 439-458.
3. Hair, J. F., Hult, G. T. M., Ringle, C. M., Sarstedt, M. 2017. *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*. 2nd edition. Sage.
4. Hair, J. F., Sarstedt, M., Ringle, C. M., Gudergan, S. P. 2017. *Advanced Issues in Partial Least Squares Structural Equation Modeling*. Sage.
5. Mehmetoglu, M., Venturini, S. 2021. *Structural Equation Modelling with Partial Least Squares Using Stata and R*. CRC Press
6. Monecke, A., Leisch, F. 2012. **`semPLS`**: Structural Equation Modeling Using Partial Least Squares. *Journal of Statistical Software*, 48, 3, 1-32.
7. Sanchez, G. 2013. *PLS Path Modeling with R*. Trowchez Editions.
8. Sanchez, G., Trinchera, L., Russolillo, G. 2015. **`plspm`**: Tools for Partial Least Squares Path Modeling (PLS-PM). R package version 0.4.7.
9. Venturini, S., Mehmetoglu, M. 2019 **`plssem`**: A Stata Package for Structural Equation Modeling with Partial Least Squares. *Journal of Statistical Software*, 88, 8, 1-35.