How can I perform a factor analysis with categorical (or categorical and continuous) variables in Stata?

Authored by

stats writer

July 1, 2024

Factor analysis is a statistical method used to identify underlying latent factors from a set of observed variables. In Stata, there are various techniques available to perform factor analysis, including principal component analysis and maximum likelihood estimation. To perform factor analysis with categorical or mixed categorical and continuous variables in Stata, one can use the command "factormat" or "factominate". These commands allow for the inclusion of categorical variables by converting them into dummy variables and using polychoric or tetrachoric correlations. This allows for the identification of common underlying factors among both categorical and continuous variables, providing insights into the relationships between these variables. The result of factor analysis can be interpreted and used for further analysis or model building.

How can I perform a factor analysis with categorical (or categorical and continuous) variables? | Stata FAQ

Standard methods of performing factor analysis (i.e., those based on a matrix of Pearson's correlations) assume that the variables are continuous and follow a multivariate normal distribution. If the model includes variables that are dichotomous or ordinal a factor analysis can be performed using a polychoric correlation matrix. In Stata we can generate a matrix of polychoric correlations using the user-written command polychoric. You can find and install the polychoric command by typing search polychoric in the Stata command window and following the directions the

screen. For more information on
locating and installing user-written commands see our FAQ:
How do I use search to search for programs and additional help?. Note that variables used with polychoric may be
binary (0/1), ordinal, or continuous, but cannot be nominal (unordered
categories). Also note that the correlations in the matrix produced by the
polychoric command are not all polychoric correlations. When both
variables have 10 or fewer observed values, a polychoric correlation is
calculated, when only one of the variables takes on 10 or fewer values (i.e., one
variable is continuous and the other categorical) a polyserial correlation is
calculated, and if both variables take on more than 10 values a Pearson's
correlation is calculated. Once we have a polychoric correlation matrix, we can use
the factormat command to perform an exploratory factor analysis using the

matrix as input, rather than raw variables.

The dataset for this example includes data on 1428 college students
and their instructors. The example analysis includes dichotomous variables, including
faculty sex (facsex) and faculty nationality (US citizen or foreign citizen,
facnat); ordered categorical variables, including faculty rank (facrank), student
rank (studrank) and grade (A, B, C, etc., grade);
and the continuous variables faculty salary (salary), years teaching at
the University of Texas (yrsut), and number of students in the class (nstud) in this analysis. These
variables were selected to represent a range of types of variables (i.e., dichotomous, ordered categorical, and continuous),
and do not necessarily form substantively meaningful factors.

Below we open the dataset and generate the polychoric correlation matrix for the
eight variables in our analysis. You may notice that the

polychoric command
runs somewhat more slowly than Stata's correlate and pwcorr
commands, this is normal.

use https://stats.idre.ucla.edu/stat/stata/output/m255, clearpolychoric facsex facnat facrank studrank grade salary yrsut nstud

Polychoric correlation matrix

facsex facnat facrank studrank grade salary
facsex 1
facnat -.08153951 1
facrank -.33496545 -.54985327 1
studrank .14701719 -.04503906 -.0006882 1
grade -.05250522 -.07768724 .03336171 .21606134 1
salary -.24422069 -.31687704 .75225252 .04830565 -.0073763 1
yrsut -.09789967 -.49838303 .68902129 .00459421 .01994406 .53046614
nstud -.46151997 .2795961 -.17809723 -.33304524 -.13713578 -.08439606

yrsut nstud

yrsut 1
nstud -.31031949 1

The polychoric command does not does not display the number of cases (with listwise deletion)
used to generate the matrix,
but it does store the n in r(sum_w) so we can use the display command to view it. Then we use the
matrix command to store the polychoric correlation matrix (saved in r(R)
by the polychoric command) as r, so that we can use it with the
factormat command. The factormat command is followed by the name
of the matrix we wish to use for the analysis (i.e., r). The n(...)
"option" gives the sample size, and is required. We have used the factors(...)
option to indicate that we wish to retain three factors.

display r(sum_w)
1338

global N = r(sum_w)

```
matrix r = r(R)
factormat r, n($N) factors(3)
(obs=1338)

Factor analysis/correlation Number of obs = 1338
Method: principal factors Retained factors = 3
Rotation: (unrotated) Number of params = 21

--------------------------------------------------------------------------
Factor | Eigenvalue Difference Proportion Cumulative
-------------+------------------------------------------------------------
Factor1 | 2.42705 1.26666 0.6995 0.6995
Factor2 | 1.16039 0.84359 0.3344 1.0340
Factor3 | 0.31680 0.18808 0.0913 1.1253
Factor4 | 0.12871 0.16060 0.0371 1.1624
Factor5 | -0.03189 0.08326 -0.0092 1.1532
Factor6 | -0.11515 0.05212 -0.0332 1.1200
Factor7 | -0.16727 0.08181 -0.0482 1.0718
Factor8 | -0.24908 . -0.0718 1.0000
--------------------------------------------------------------------------
LR test: independent vs. saturated: chi2(28) = 3824.64
Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances
```

```
-------------------------------------------------------------
Variable | Factor1 Factor2 Factor3 | Uniqueness
-------------+------------------------------+--------------
facsex | -0.1902 -0.6651 -0.2171 | 0.4744
facnat | -0.5913 0.2174 0.1465 | 0.5816
facrank | 0.9183 0.1642 0.0173 | 0.1295
studrank | 0.0645 -0.3558 0.3430 | 0.7516
grade | 0.0636 -0.1380 0.3316 | 0.8670
salary | 0.7365 0.1822 0.0751 | 0.4187
yrsut | 0.7520 -0.0762 -0.1107 | 0.4165
nstud | -0.2861 0.6777 -0.0493 | 0.4565
-------------------------------------------------------------
```

The above factor analysis output can be interpreted in a manner similar to a standard factor analysis model, including the use of rotation methods to increase interpretability.

See also

Stata Annotated Output: Factor Analysis