# How can I perform factor analysis with missing data in Stata?

Authored by

**stats writer**

July 1, 2024

Factor analysis is a statistical method commonly used in data analysis to identify underlying factors or dimensions that explain the relationships among a set of observed variables. However, missing data can pose a challenge in performing factor analysis. Stata, a statistical software package, offers various techniques to handle missing data in factor analysis. One approach is to use the "missing data imputation" method, which replaces missing values with estimated values based on the available data. Another approach is the "direct maximum likelihood estimation" method, which uses all available data to estimate the factor structure. Additionally, Stata also offers the option to perform factor analysis on complete cases or to delete cases with missing data. With these techniques, users can effectively perform factor analysis with missing data in Stata and obtain reliable results for their data analysis.

## How can I do factor analysis with missing data in Stata? | Stata FAQ

**Trying to run factor analysis with missing data can be problematic. One issue is that traditional multiple imputation methods, such as mi estimate, don't work with Stata's factor command. Truxillo (2005) , Graham (2009), and Weaver and Maxwell (2014) have suggested an approach using maximum likelihood with the expectation-maximization (EM) algorithm to estimate of the covariance matrix. Stata's mi command computes an EM covariance matrix as part of the imputation process. We will demonstrate how to use this EM covariance matrix to obtain a factor solution.**

**To begin, we will load a Stata dataset fa_missing, get some descriptive statistics and compute the complete case covariance matrix.**

```
use https://stats.idre.ucla.edu/stat/data/fa_missing, clear

summarize

Variable | Obs Mean Std. Dev. Min Max
-------------+--------------------------------------------------------
item13 | 1419 4.450317 .7374944 1 5
item14 | 1428 4.518207 .7086049 1 5
item15 | 1424 4.434691 .7478835 1 5
item16 | 1420 4.270423 .8387034 1 5
item17 | 1423 4.158819 .8969815 1 5
-------------+--------------------------------------------------------
item18 | 1424 3.924157 1.032095 1 5
item19 | 1420 4.072535 .9665034 1 5
item20 | 1396 3.770774 .9137137 1 5
item21 | 1422 3.769339 .9863042 1 5
item22 | 1414 3.592645 1.122807 1 5
-------------+--------------------------------------------------------
item23 | 1423 3.800422 .9639492 1 5
item24 | 1417 3.653493 .9308223 1 5
```

item25 | 1398 2.285408 .9892487 1 5
item26 | 1414 2.077086 1.058313 1 5
item27 | 1420 1.496479 .7294192 1 5
-------------+---------------------------------------------------------
item28 | 1419 2.273432 .9677116 1 5

count /* count total number of observations */

1428

corr, cov /* complete case covariance matrix */

(obs=1331)

| item13 item14 item15 item16 item17 item18 item19 item20 item21 item22
-------------+---------------------------------------------------------
-------------------------
item13 | .536077
item14 | .338379 .48707
item15 | .321189 .321527 .536102
item16 | .345401 .289728 .305775 .681157
item17 | .376869 .34434 .382685 .435569 .799719
item18 | .310446 .313639 .348313 .350631 .52091 1.07943
item19 | .202336 .211924 .250935 .262903 .387361 .6311

.935065

item20 | .202417 .199169 .234828 .235136 .33403 .490213 .39099 .82419

item21 | .340947 .302693 .364249 .369725 .525901 .565762 .472584 .377907 .966934

item22 | .271303 .257165 .30236 .328247 .450612 .625471 .521898 .389491 .553577 1.24822

item23 | .39691 .374938 .405833 .35907 .526654 .567222 .406462 .352182 .558713 .52722

item24 | .305477 .28148 .290924 .325099 .432752 .457068 .332405 .30078 .449396 .456496

item25 | .008449 -.011696 -.038745 -.030012 -.039108 -.042074 -.064693 -.026436 -.025697 -.047378

item26 | .014954 -.024045 -.002687 -.019264 -.021647 -.018859 .018107 -.026555 .002384 -.019735

item27 | -.036163 -.045486 -.046055 -.065249 -.055178 -.070832 -.053228 -.036927 -.062904 -.099815

item28 | -.000554 -.013315 -.033624 -.048267 -.028426 -.051824 -.016597 -.044399 -.031681 -.07906

| item23 item24 item25 item26 item27 item28

----------+--------------------------------------------------

item23 | .913566

item24 | .618358 .848286

item25 | -.031721 -.043576 .976103
item26 | .014638 -.025494 .10275 1.10263
item27 | -.059988 -.063666 .123452 .170048 .51931
item28 | -.004233 -.049099 .23827 .210952 .353081 .941695

From the output above, you can see that there are a total of 1,428 observations with 1,365 complete cases. All of the variables have missing cases except for item14.
item20 has the most missing data with only 1,396 nonmissing cases.

We will use the mlong format for mi set but this approach will work with any of the mi data formats. When you register variables to be imputed (mi register imputed) you should also include the variables without missing values, such as item14, so that they will be included in the EM covariance matrix.
Next, run the mi impute mvn command with the emonly option. Notice that there are no variables to the right of the equal sign. In fact, there is no equal sign at all.

**After running mi impute, the EM covariance matrix can be found in the saved results in r(Sigma_em) which we will then save to the matrix cov_em for use in factormat.**

```
mi set mlong

mi register imputed item13-item28

(97 m=0 obs. now marked as incomplete)

mi impute mvn item13-item28, emonly

note: variable item14 contains no soft missing (.) values; imputing nothing

Iteration 0: Observed log likelihood = -9021.7844
Iteration 1: Observed log likelihood = -4116.7934
Iteration 2: Observed log likelihood = -4113.8728
Iteration 3: Observed log likelihood = -4113.8685
Iteration 4: Observed log likelihood = -4113.8685
Iteration 5: Observed log likelihood = -4113.8685

Expectation-maximization estimation Number obs = 1428
Number missing = 167
```

**Number patterns = 34**

**Prior: uniform Obs per pattern: min = 1**

**avg = 42**

**max = 1331**

**Observed log likelihood = -4113.8685 at iteration 5**

**------------------------------------------------------------------------------**

**| item13 item14 item15 item16 item17 item18 item19 item20**

**-------------+----------------------------------------------------------------**

**Coef |**

**_cons | 4.451285 4.518207 4.435308 4.268804 4.156375 3.922213 4.070152 3.767296**

**-------------+----------------------------------------------------------------**

**Sigma |**

**item13 | .5430297 .348556 .3350714 .3450292 .3822566 .3104201 .2059764 .2103791**

**item14 | .348556 .5017693 .3455828 .3003632 .355927 .3212322 .2256632 .2100503**

**item15 | .3350714 .3455828 .5584064 .3186633 .400073**

.3582909 .2667201 .2515707

item16 | .3450292 .3003632 .3186633 .705225 .4391928 .3453138 .2643872 .2404926

item17 | .3822566 .355927 .400073 .4391928 .8085157 .5141493 .3895897 .3465985

item18 | .3104201 .3212322 .3582909 .3453138 .5141493 1.068469 .6312669 .4935741

item19 | .2059764 .2256632 .2667201 .2643872 .3895897 .6312669 .9371905 .4000026

item20 | .2103791 .2100503 .2515707 .2404926 .3465985 .4935741 .4000026 .8360225

item21 | .3460404 .3144694 .3764157 .3618431 .5214286 .5644802 .4791261 .3844396

item22 | .2817349 .2724658 .3191765 .3361343 .4594059 .6214813 .5312196 .3984811

item23 | .4075676 .3925719 .4292744 .3747293 .5356915 .567577 .4185394 .3681236

item24 | .3165909 .3003985 .3143 .334972 .4410231 .4564384 .3410914 .3120366

item25 | -.004998 -.0312026 -.0558893 -.0478347 -.049516 -.0489516 -.0731514 -.0307672

item26 | .0131501 -.0212114 .0033069 -.0091686 -.0192239 -.0085362 .0180015 -.0241016

item27 | -.0416338 -.054581 -.0551466 -.0723354 -

.0615237 -.0827423 -.0596916 -.0426282
item28 | -.0053221 -.0246628 -.0425885 -.0575029 -.0363375 -.0585845 -.0250774 -.0471011
---
---

---
---
| item21 item22 item23 item24 item25 item26 item27 item28
---+---
---
Coef |
_cons | 3.770048 3.593149 3.79814 3.655047 2.285293 2.077139 1.49686 2.273121
---+---
---
Sigma |
item13 | .3460404 .2817349 .4075676 .3165909 -.004998 .0131501 -.0416338 -.0053221
item14 | .3144694 .2724658 .3925719 .3003985 -.0312026 -.0212114 -.054581 -.0246628
item15 | .3764157 .3191765 .4292744 .3143 -.0558893 .0033069 -.0551466 -.0425885

item16 | .3618431 .3361343 .3747293 .334972 -.0478347 -.0091686 -.0723354 -.0575029

item17 | .5214286 .4594059 .5356915 .4410231 -.049516 -.0192239 -.0615237 -.0363375

item18 | .5644802 .6214813 .567577 .4564384 -.0489516 -.0085362 -.0827423 -.0585845

item19 | .4791261 .5312196 .4185394 .3410914 -.0731514 .0180015 -.0596916 -.0250774

item20 | .3844396 .3984811 .3681236 .3120366 -.0307672 -.0241016 -.0426282 -.0471011

item21 | .970728 .5586832 .5705437 .4616899 -.0378778 .0058816 -.0728921 -.0373168

item22 | .5586832 1.261583 .5445269 .4737303 -.0510983 -.0252185 -.0972336 -.0784298

item23 | .5705437 .5445269 .9342235 .6368121 -.0503701 .0146983 -.066706 -.009803

item24 | .4616899 .4737303 .6368121 .8657582 -.0596746 -.0287266 -.0696632 -.0554101

item25 | -.0378778 -.0510983 -.0503701 -.0596746 .9778334 .0906248 .1282772 .2393721

item26 | .0058816 -.0252185 .0146983 -.0287266 .0906248 1.118956 .1671461 .196613

item27 | -.0728921 -.0972336 -.066706 -.0696632 .1282772 .1671461 .5314816 .3568273

item28 | -.0373168 -.0784298 -.009803 -.0554101 .2393721 .196613 .3568273 .9353024

---

Note: no imputation performed.

matrix cov_em = r(Sigma_em)

matrix list cov_em

symmetric cov_em
item13 item14 item15 item16 item17 item18 item19 item20
item13 .54302971
item14 .348556 .50176934
item15 .33507137 .34558277 .55840641
item16 .34502917 .3003632 .31866333 .705225
item17 .38225661 .35592696 .40007302 .43919281 .80851571
item18 .31042005 .32123222 .35829089 .34531382 .51414933 1.0684692
item19 .20597637 .22566319 .26672014 .26438724 .38958966 .63126692 .93719047
item20 .21037912 .21005025 .25157069 .24049261

.34659848 .49357406 .40000261 .83602251
item21 .34604038 .31446937 .37641575 .3618431 .52142858 .56448016 .47912613 .38443958
item22 .28173489 .27246583 .31917653 .33613426 .45940591 .62148129 .53121959 .39848113
item23 .40756762 .39257192 .42927439 .37472927 .53569152 .56757702 .41853944 .36812361
item24 .3165909 .30039846 .3143 .33497204 .4410231 .45643844 .34109143 .31203658
item25 -.00499796 -.03120263 -.05588928 -.04783472 -.04951603 -.04895159 -.07315143 -.03076718
item26 .01315008 -.02121138 .00330686 -.0091686 -.0192239 -.00853618 .01800149 -.02410163
item27 -.04163384 -.054581 -.05514657 -.07233544 -.06152372 -.08274233 -.05969162 -.04262815
item28 -.00532215 -.02466275 -.0425885 -.05750295 -.03633748 -.05858451 -.02507744 -.04710109

item21 item22 item23 item24 item25 item26 item27 item28
item21 .97072797
item22 .55868325 1.2615831
item23 .57054369 .54452686 .93422349
item24 .46168988 .4737303 .63681207 .86575819

item25 -.03787779 -.05109826 -.05037005 -.05967465 .97783345
item26 .00588162 -.02521849 .01469832 -.02872665 .09062481 1.1189565
item27 -.07289205 -.09723358 -.06670603 -.0696632 .1282772 .16714608 .53148159
item28 -.03731678 -.07842976 -.00980302 -.05541012 .23937206 .19661302 .35682735 .93530244

We will use the factormat command with the EM estimate of the covariance matrix to obtain
our factor solution. The factormat is for use with a correlation or covariance matrix. The command requires that the sample size, n, be entered along with the
name of the covariance matrix. In her paper, Truxillo discusses three methods for specifying nominal
sample size, 1) column-wise minimum, 2) column-wise average and 3) pairwise minimum.
Column-wise minimum is just the number of complete cases for the variables with the most missing
values which is the value we will use for this example. If you will recall from above that value is

1,396.

factormat cov_em, n(1396) fact(4) ml

(obs=1396)
Iteration 0: log likelihood = -236.78484
Iteration 1: log likelihood = -85.766521
(...omitted...)
Iteration 90: log likelihood = -85.345691
Iteration 91: log likelihood = -85.345691

Factor analysis/correlation Number of obs = 1396
Method: maximum likelihood Retained factors = 4
Rotation: (unrotated) Number of params = 58
Schwarz's BIC = 590.691
Log likelihood = -85.34569 (Akaike's) AIC = 286.691

---
Factor | Eigenvalue Difference Proportion Cumulative
---+---
Factor1 | 5.83062 4.66732 0.7030 0.7030
Factor2 | 1.16329 0.34933 0.1403 0.8432
Factor3 | 0.81396 0.32778 0.0981 0.9414
Factor4 | 0.48619 . 0.0586 1.0000
---

LR test: independent vs. saturated: chi2(120) = 9652.01
Prob>chi2 = 0.0000
LR test: 4 factors vs. saturated: chi2(62) = 169.61
Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances

---

| Variable | Factor1 | Factor2 | Factor3 | Factor4 | Uniqueness |
|---|---|---|---|---|---|
| item13 | 0.7069 | 0.0779 | -0.3640 | 0.1850 | 0.3274 |
| item14 | 0.7087 | 0.0258 | -0.3268 | 0.1776 | 0.3588 |
| item15 | 0.7260 | -0.0043 | -0.2333 | 0.1634 | 0.3919 |
| item16 | 0.6158 | -0.0577 | -0.2016 | 0.2151 | 0.5306 |
| item17 | 0.7618 | -0.0252 | -0.0548 | 0.1732 | 0.3860 |
| item18 | 0.7099 | -0.1273 | 0.3400 | 0.1912 | 0.3276 |
| item19 | 0.5856 | -0.1310 | 0.4220 | 0.2273 | 0.4102 |
| item20 | 0.5331 | -0.0986 | 0.2382 | 0.1632 | 0.6227 |
| item21 | 0.7143 | -0.0444 | 0.1319 | 0.0987 | 0.4606 |
| item22 | 0.6024 | -0.1166 | 0.2541 | 0.0639 | 0.5549 |
| item23 | 0.8857 | 0.1157 | 0.0221 | -0.2968 | 0.1134 |
| item24 | 0.7246 | 0.0169 | 0.0385 | -0.2133 | 0.4276 |
| item25 | -0.0717 | 0.2901 | 0.0348 | 0.0744 | 0.9039 |
| item26 | -0.0061 | 0.2680 | 0.0632 | 0.0585 | 0.9207 |

item27 | -0.1399 0.6198 0.1288 0.1428 | 0.5593
item28 | -0.0586 0.7349 0.1431 0.1606 | 0.4102
---

rotate, varimax normalize blanks(.3)

Factor analysis/correlation Number of obs = 1396
Method: maximum likelihood Retained factors = 4
Rotation: orthogonal varimax (Kaiser on) Number of params = 58
Schwarz's BIC = 590.691
Log likelihood = -85.34569 (Akaike's) AIC = 286.691

---
Factor | Variance Difference Proportion Cumulative
---+---
Factor1 | 3.25325 0.32888 0.3922 0.3922
Factor2 | 2.92437 1.70143 0.3526 0.7448
Factor3 | 1.22294 0.32944 0.1474 0.8923
Factor4 | 0.89350 . 0.1077 1.0000
---
LR test: independent vs. saturated: chi2(120) = 9652.01
Prob>chi2 = 0.0000
LR test: 4 factors vs. saturated: chi2(62) = 169.61
Prob>chi2 = 0.0000

## Rotated factor loadings (pattern matrix) and unique variances

---

Variable | Factor1 Factor2 Factor3 Factor4 | Uniqueness

---+---

item13 | 0.7843 | 0.3274

item14 | 0.7534 | 0.3588

item15 | 0.6961 0.3120 | 0.3919

item16 | 0.6114 | 0.5306

item17 | 0.6040 0.4695 | 0.3860

item18 | 0.3043 0.7487 | 0.3276

item19 | 0.7454 | 0.4102

item20 | 0.5561 | 0.6227

item21 | 0.4237 0.5581 | 0.4606

item22 | 0.5829 | 0.5549

item23 | 0.5049 0.4350 0.6651 | 0.1134

item24 | 0.4010 0.3935 0.5019 | 0.4276

item25 | 0.3044 | 0.9039

item26 | | 0.9207

item27 | 0.6562 | 0.5593

item28 | 0.7676 | 0.4102

---

(blanks represent abs(loading)<.3)

## Factor rotation matrix

```
-------------------------------------------------
| Factor1 Factor2 Factor3 Factor4
-------------+-----------------------------------
Factor1 | 0.6778 0.5954 -0.0605 0.4270
Factor2 | 0.1006 -0.1816 0.9512 0.2282
Factor3 | -0.6581 0.7251 0.1934 0.0609
Factor4 | 0.3120 0.2945 0.2326 -0.8728
-------------------------------------------------
```

**Almost identical results to these were obtain using SAS proc mi with proc factor and**
**using Mplus with the missing data option.**

## Reference

**Truxillo, C. (2005). Maximum likelihood parameter estimation with incomplete data.**
**Proceedings of the Thirtieth Annual SAS(r) Users Group International Conference.**
**<http://www2.sas.com/proceedings/sugi30/111-30.pdf >**

**Graham, J. W. (2009). Missing data analysis: Making it work in the real world. Annu. Rev. Psychol., 60, 549-576.**

https://www.personal.psu.edu/jxb14/M554/articles/Graham2009.pdf

Weaver, B., & Maxwell, H. (2014). Exploratory factor analysis and reliability analysis with missing data: A simple method for SPSS users. The Quantitative Methods for Psychology, 10 (2), 143-152. https://www.tqmp.org/RegularArticles/vol10-2/p143/p143.pdf