Title

stata.com

ca postestimation — Postestimation tools for ca and camat

Postestimation commands

The following postestimation commands are of special interest after `ca` and `camat`:

Command	Description
`cabiplot`	biplot of row and column points
`caprojection`	CA dimension projection plot
`estat coordinates`	display row and column coordinates
`estat distances`	display χ^2 distances between row and column profiles
`estat inertia`	display inertia contributions of the individual cells
`estat loadings`	display correlations of profiles and axes
`estat profiles`	display row and column profiles
* `estat summarize`	estimation sample summary
`estat table`	display fitted correspondence table
`screeplot`	plot singular values

* `estat summarize` is not available after `camat`.

The following standard postestimation commands are also available:

Command	Description
* `estimates`	cataloging estimation results
† `predict`	fitted values, row coordinates, or column coordinates

* All `estimates` subcommands except `table` and `stats` are available.

† `predict` is not available after `camat`.

predict

Description for predict

`predict` creates a new variable containing predictions such as fitted values and row or column scores.

Menu for predict

Statistics > Postestimation

Syntax for predict

`predict` [*type*] *newvar* [*if*] [*in*] [`,` *statistic*]

statistic	Description
Main	
`fit`	fitted values; the default
`rowscore(#)`	row score for dimension #
`colscore(#)`	column score for dimension #

`predict` is not available after `camat`.

Options for predict

Main

`fit` specifies that fitted values for the correspondence analysis model be computed. `fit` displays the fitted values p_{ij} according to the correspondence analysis model. `fit` is the default.

`rowscore(#)` generates the row score for dimension #, that is, the appropriate elements from the normalized row coordinates.

`colscore(#)` generates the column score for dimension #, that is, the appropriate elements from the normalized column coordinates.

estat

Description for estat

`estat coordinates` displays the row and column coordinates.

`estat distances` displays the χ^2 distances between the row profiles and between the column profiles. Also, the χ^2 distances between the row and column profiles to the respective centers (marginal distributions) are displayed. Optionally, the fitted profiles rather than the observed profiles are used.

`estat inertia` displays the inertia (χ^2/N) contributions of the individual cells.

`estat loadings` displays the correlations of the row and column profiles and the axes, comparable to the loadings of principal component analysis.

`estat profiles` displays the row and column profiles; the row (column) profile is the conditional distribution of the row (column) given the column (row). This is equivalent to specifying the `row` and `column` options with the `tabulate` command; see [R] **tabulate twoway**.

`estat summarize` displays summary information about the row and column variables over the estimation sample.

`estat table` displays the fitted correspondence table. Optionally, the observed "correspondence table" and the expected table under independence are displayed.

Menu for estat

Statistics > Postestimation

Syntax for estat

Display row and column coordinates

`estat coordinates [, norow nocolumn format(%fmt)]`

Display chi-squared distances between row and column profiles

`estat distances [, norow nocolumn approx format(%fmt)]`

Display inertia contributions of cells

`estat inertia [, total noscale format(%fmt)]`

Display correlations of profiles and axes

`estat loadings [, norow nocolumn format(%fmt)]`

Display row and column profiles

`estat profiles [, norow nocolumn format(%fmt)]`

Display summary information

```
estat summarize [, labels noheader noweights]
```

Display fitted correspondence table

```
estat table [, fit obs independence noscale format(%fmt)]
```

options	Description
norow	suppress display of row results
nocolumn	suppress display of column results
format(%*fmt*)	display format; default is format(%9.4f)
approx	display distances between fitted (approximated) profiles
total	add row and column margins
noscale	display χ^2 contributions; default is inertias $= \chi^2/N$ (with estat inertia)
labels	display variable labels
noheader	suppress the header
noweights	ignore weights
fit	display fitted values from correspondence analysis model
obs	display correspondence table ("observed table")
independence	display expected values under independence
noscale	suppress scaling of entries to 1 (with estat table)

Options for estat

norow, an option used with estat coordinates, estat distances, and estat profiles, suppresses the display of row results.

nocolumn, an option used with estat coordinates, estat distances, and estat profiles, suppresses the display of column results.

format(%*fmt*), an option used with many of the subcommands of estat, specifies the display format for the matrix, for example, format(%8.3f). The default is format(%9.4f).

approx, an option used with estat distances, computes distances between the fitted profiles. The default is to compute distances between the observed profiles.

total, an option used with estat inertia, adds row and column margins to the table of inertia or χ^2 (χ^2/N) contributions.

noscale, as an option used with estat inertia, displays χ^2 contributions rather than inertia ($= \chi^2/N$) contributions. (See below for the description of noscale with estat table.)

labels, an option used with estat summarize, displays variable labels.

noheader, an option used with estat summarize, suppresses the header.

noweights, an option used with estat summarize, ignores the weights, if any. The default when weights are present is to perform a weighted summarize on all variables except the weight variable itself. An unweighted summarize is performed on the weight variable.

fit, an option used with estat table, displays the fitted values for the correspondence analysis model. fit is implied if obs and independence are not specified.

`obs`, an option used with `estat table`, displays the observed table with nonnegative entries (the "correspondence table").

`independence`, an option used with `estat table`, displays the expected values p_{ij} assuming independence of the rows and columns, $p_{ij} = r_i c_j$, where r_i is the mass of row i and c_j is the mass of column j.

`noscale`, as an option used with `estat table`, normalizes the displayed tables to the sum of the original table entries. The default is to scale the tables to overall sum 1. (See above for the description of `noscale` with `estat inertia`.)

Remarks and examples

stata.com

Remarks are presented under the following headings:

Postestimation statistics

After you conduct a correspondence analysis, there are several additional tables to help you understand and interpret your results. Some of these tables resemble tables produced by other Stata commands but are provided as part of the `ca` postestimation suite of commands for a unified presentation style.

▷ Example 1: estat profiles, estat distances, estat table

We continue with the classic example of correspondence analysis, namely, the data on smoking in organizations. We extract only one dimension.

```
. use http://www.stata-press.com/data/r15/ca_smoking
. ca rank smoking, dim(1)
Correspondence analysis                   Number of obs      =       193
                                          Pearson chi2(12)   =     16.44
                                          Prob > chi2        =    0.1718
                                          Total inertia      =    0.0852
    5 active rows                         Number of dim.     =         1
    4 active columns                      Expl. inertia (%)  =     87.76
```

Dimension	singular value	principal inertia	chi2	percent	cumul percent
dim 1	.2734211	.0747591	14.43	87.76	87.76
dim 2	.1000859	.0100172	1.93	11.76	99.51
dim 3	.0203365	.0004136	0.08	0.49	100.00
total		.0851899	16.44	100	

Statistics for row and column categories in symmetric normalization

Categories	overall mass	overall quality	overall %inert	dimension_1 coord	dimension_1 sqcorr	dimension_1 contrib
rank						
senior mngr	0.057	0.092	0.031	0.126	0.092	0.003
junior mngr	0.093	0.526	0.139	-0.495	0.526	0.084
senior empl	0.264	0.999	0.450	0.728	0.999	0.512
junior empl	0.456	0.942	0.308	-0.446	0.942	0.331
secretary	0.130	0.865	0.071	0.385	0.865	0.070
smoking						
none	0.316	0.994	0.577	0.752	0.994	0.654
light	0.233	0.327	0.083	-0.190	0.327	0.031
medium	0.321	0.982	0.148	-0.375	0.982	0.166
heavy	0.130	0.684	0.192	-0.562	0.684	0.150

CA analyzes the similarity of row and of column categories by comparing the row profiles and the column profiles—some may prefer to talk about conditional distributions for a two-way frequency distribution, but CA is not restricted to this type of data.

```
. estat profiles
```

Row profiles (rows normalized to 1)

	none	light	medium	heavy	mass
senior mngr	0.3636	0.1818	0.2727	0.1818	0.0570
junior mngr	0.2222	0.1667	0.3889	0.2222	0.0933
senior empl	0.4902	0.1961	0.2353	0.0784	0.2642
junior empl	0.2045	0.2727	0.3750	0.1477	0.4560
secretary	0.4000	0.2400	0.2800	0.0800	0.1295
mass	0.3161	0.2332	0.3212	0.1295	

Column profiles (columns normalized to 1)

	none	light	medium	heavy	mass
senior mngr	0.0656	0.0444	0.0484	0.0800	0.0570
junior mngr	0.0656	0.0667	0.1129	0.1600	0.0933
senior empl	0.4098	0.2222	0.1935	0.1600	0.2642
junior empl	0.2951	0.5333	0.5323	0.5200	0.4560
secretary	0.1639	0.1333	0.1129	0.0800	0.1295
mass	0.3161	0.2332	0.3212	0.1295	

The tables also include the row and column masses—marginal probabilities. Two row categories are similar to the extent that their row profiles (that is, their distribution over the columns) are the same. Similar categories could be collapsed without distorting the information in the table. In CA, similarity or dissimilarity of the row categories is expressed in terms of the χ^2 distances between the rows. These are sums of squares, weighted with the inverse of the column masses. Thus a difference is counted "heavier" (inertia!) the smaller the respective column mass. In the table, we also add the χ^2 distances of the rows to the row centroid, that is, to the marginal distribution. This allows us to easily see which row categories are similar to each other as well as which row categories are similar to the population.

```
. estat distances, nocolumn
Chi2 distances between the row profiles
```

rank	junior_~r	senior_~l	junior_~l	secretary	center
senior_mngr	0.3448	0.3721	0.3963	0.3145	0.2166
junior_mngr		0.6812	0.3044	0.5622	0.3569
senior_empl			0.6174	0.2006	0.3808
junior_empl				0.4347	0.2400
secretary					0.2162

We see that senior employees are especially dissimilar from junior managers in terms of their smoking behavior but are rather similar to secretaries. Also the senior employees are least similar to the average staff member among all staff categories.

One of the goals of CA is to come up with a low-dimensional representation of the rows and columns in a common space. One way to see the adequacy of this representation is to inspect the implied approximation for the χ^2 distances—are the similarities between the row categories and between the column categories adequately represented in lower dimensions?

```
. estat distances, nocolumn approx
Chi2 distances between the dim=1 approximations of the row profiles
```

rank	junior_~r	senior_~l	junior_~l	secretary	center
senior_mngr	0.3247	0.3148	0.2987	0.1353	0.0658
junior_mngr		0.6396	0.0260	0.4600	0.2590
senior_empl			0.6135	0.1795	0.3806
junior_empl				0.4340	0.2330
secretary					0.2011

Some of the row distances are obviously poorly approximated, whereas the quality of other approximations is hardly affected. The dissimilarity in smoking behavior between junior managers and junior employees is particularly poorly represented in one dimension. From the CA with two dimensions, the second dimension is crucial to adequately represent the senior managers and the junior managers. By itself, this does not explain where the one-dimensional approximation fails; for this, we would have to take a closer look at the representation of the smoking categories as well.

A correspondence analysis can also be seen as equivalent to fitting the model

$$P_{ij} = r_i c_j (1 + R_{i1} C_{j1} + R_{i2} C_{j2} + \cdots)$$

to the correspondence table $\mathbf{P}$ by some sort of least squares, with parameters r_i, c_j, R_{ij}, and C_{jk}. We may compare the (observed) table $\mathbf{P}$ with the fitted table $\widehat{\mathbf{P}}$ to assess goodness of fit informally. Here we extract only one dimension, and so the fitted table is

$$\widehat{P}_{ij} = r_i c_j (1 + \widehat{R}_{i1} \widehat{C}_{j1})$$

with $\mathbf{R}$ and $\mathbf{C}$ the coordinates in symmetric (or row principal or column principal) normalization. We display the observed and fitted tables.

```
. estat table, fit obs
Correspondence table (normalized to overall sum = 1)
```

	none	light	medium	heavy
senior_mngr	0.0207	0.0104	0.0155	0.0104
junior_mngr	0.0207	0.0155	0.0363	0.0207
senior_empl	0.1295	0.0518	0.0622	0.0207
junior_empl	0.0933	0.1244	0.1710	0.0674
secretary	0.0518	0.0311	0.0363	0.0104

```
Approximation for dim = 1 (normalized to overall sum = 1)
```

	none	light	medium	heavy
senior_mngr	0.0197	0.0130	0.0174	0.0069
junior_mngr	0.0185	0.0238	0.0355	0.0154
senior_empl	0.1292	0.0531	0.0617	0.0202
junior_empl	0.0958	0.1153	0.1710	0.0738
secretary	0.0528	0.0280	0.0356	0.0132

Interestingly, some categories (for example, the junior employees, the nonsmokers, and the medium smokers) are very well represented in one dimension, whereas the quality of the fit of other categories is rather poor. This can, of course, also be inferred from the quality column in the `ca` output. We would consider the fit unsatisfactory and would refit the model with a second dimension.

◁

❑ Technical note

If the data are two-way cross-classified frequencies, as with `ca`, it may make sense to assume that the data are multinomial distributed, and the parameters can be estimated by maximum likelihood. The estimator has well-established properties in contrast to the estimation method commonly used in CA. One advantage is that sampling variability, for example, in terms of standard errors of the parameters, can be easily assessed. Also, the likelihood-ratio test against the saturated model may be used to select the number of dimensions to be extracted. See Van der Heijden and de Leeuw (1985).

❑

Predicting new variables

If you use `ca` to obtain the optimal scaling positions for the rows and columns, you may use `predict` to obtain the corresponding scores in the normalization used.

▷ Example 2: Predictions

First, we obtain scores for the first dimension.

```
. quietly ca rank smoking, normalize(symmetric) dim(2)
. predict r1, row(1)
. predict c1, col(1)
. describe r1 c1
              storage   display    value
variable name   type    format     label      variable label
------------------------------------------------------------------------------
r1              float   %9.0g                 rank score(1) in symmetric norm.
c1              float   %9.0g                 smoking score(1) in symmetric
                                                norm.
. correlate r1 c1
(obs=193)
             |       r1       c1
-------------+------------------
          r1 |   1.0000
          c1 |   0.2734   1.0000
```

The correlation of `r1` and `c1` is 0.2734, which equals the first singular value reported in the first panel by `ca`. In the same way, we may obtain scores for the second dimension.

```
. predict r2, row(2)
. predict c2, col(2)
. correlate r1 r2 c1 c2
(obs=193)
             |       r1       r2       c1       c2
-------------+------------------------------------
          r1 |   1.0000
          r2 |  -0.0000   1.0000
          c1 |   0.2734   0.0000   1.0000
          c2 |   0.0000   0.1001   0.0000   1.0000
```

The correlation between the row and column scores `r2` and `c2` for the second dimension is 0.1001, which is the same as the second singular value. Moreover, the row scores for dimensions 1 and 2 are not correlated, nor are the column scores.

◁

Obtaining the fitted values of the CA model is also possible,

$$\pi_{ij} = r_i c_j (1 + R_{i1} C_{i1} + R_{i2} C_{i2})$$

where $\mathbf{R}$ and $\mathbf{C}$ are the row and column scales in symmetric normalization. These may be used, say, to compute fit measures, for instance, from the Cressie–Read power family to analyze the fit of the CA model (Weesie 1997).

Stored results

`estat distances` stores the following in `r()`:

Matrices

`r(Dcolumns)`	χ^2 distances between the columns and between the columns and the column center
`r(Drows)`	χ^2 distances between the rows and between the rows and the row center

`estat inertia` stores the following in `r()`:

Matrices

`r(Q)`	matrix of (squared) inertia (or χ^2) contributions

`estat loadings` stores the following in `r()`:

Matrices

`r(LC)`	column loadings
`r(LR)`	row loadings

`estat profiles` stores the following in `r()`:

Matrices

`r(Pcolumns)`	column profiles (columns normalized to 1)
`r(Prows)`	row profiles (rows normalized to 1)

`estat table` stores the following in `r()`:

Matrices

`r(Fit)`	fitted (reconstructed) values
`r(Fit0)`	fitted (reconstructed) values, assuming independence of row and column variables
`r(Obs)`	correspondence table

Methods and formulas

See *Methods and formulas* in [MV] **ca** for information.

References

Van der Heijden, P. G. M., and J. de Leeuw. 1985. Correspondence analysis used complementary to loglinear analysis. *Psychometrika* 50: 429–447.

Weesie, J. 1997. sg68: Goodness-of-fit statistics for multinomial distributions. *Stata Technical Bulletin* 36: 26–28. Reprinted in *Stata Technical Bulletin Reprints*, vol. 6, pp. 183–186. College Station, TX: Stata Press.

Also see *References* in [MV] **ca**.

Also see

[MV] **ca** — Simple correspondence analysis

[MV] **ca postestimation plots** — Postestimation plots for ca and camat

[MV] **screeplot** — Scree plot

[U] **20 Estimation and postestimation commands**