The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*), *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

Subscriptions are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

Subscription rates listed below include both a printed and an electronic copy unless otherwise mentioned.

U.S. and Canada		Elsewhere	
Printed & electronic		**Printed & electronic**	
1-year subscription	$115	1-year subscription	$145
2-year subscription	$210	2-year subscription	$270
3-year subscription	$285	3-year subscription	$375
1-year student subscription	$ 85	1-year student subscription	$115
1-year institutional subscription	$345	1-year institutional subscription	$375
2-year institutional subscription	$625	2-year institutional subscription	$685
3-year institutional subscription	$875	3-year institutional subscription	$965
Electronic only		**Electronic only**	
1-year subscription	$ 85	1-year subscription	$ 85
2-year subscription	$155	2-year subscription	$155
3-year subscription	$215	3-year subscription	$215
1-year student subscription	$ 55	1-year student subscription	$ 55

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.

The Stata Journal (2015)
15, Number 2, pp. 547–553

Transition matrix for a bivariate normal distribution in Stata

Marco Savegnago
Bank of Italy
Rome, Italy
marco.savegnago@bancaditalia.it

Abstract. `trabinor` calculates the population transition matrix between two discretized variables when the original continuous variables follow a bivariate normal distribution. The user can specify the five parameters of bivariate normal and how to discretize the two variables by choosing either a given number of quantiles or a set of absolute boundaries.

Keywords: st0392, trabinor, transition matrix, bivariate normal, binormal()

1 Introduction

A transition matrix is a useful tool to describe a stochastic process that involves a finite number of states. Examples include transitions among employment statuses (unemployed, employed part-time, and full-time) for a population of workers or among rating classes for companies and governments. Often the variable of interest is continuous, such as income; for instance, one might be interested in the association between an individual's income and the income of his or her father. After a proper discretization in a finite number of classes, a transition matrix can help answer questions like the following: given that the father earns less than x, what is the probability that a child earns more than y (that is, undergoes upward mobility) or less than x itself (that is, falls into a "poverty trap")?

`trabinor` calculates the transition matrix between two variables (after being appropriately discretized) if they follow a bivariate normal distribution. It can be used mostly for the following two purposes:

- It can be used when sample size is too small to have a reliable estimate of the transition matrix, especially when there are many cells to estimate. If one is willing to assume that the two variables are jointly normal, then sample moments can be computed from data and passed to `trabinor`.

 st0392

- It can be used in Monte Carlo simulations to study the properties of estimators related to a transition matrix. For example, in intergenerational mobility analysis, the trace of a matrix (that is, the sum of the elements of the main diagonal in a square matrix) is an important summary measure of persistence of socioeconomic status[1] (for a recent survey, see Black and Devereux [2011]). The behavior of this type of estimator when variables are contaminated by measurement errors is studied in O'Neill, Sweetman, and Van de gaer (2007). `trabinor` can be used to compute the "true" (that is, population) value of such estimators under different parameter values.

2 Statistical background

Assume that

$$\begin{bmatrix} Y \\ X \end{bmatrix} \backsim \mathcal{N}\left[\begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & \rho\sigma_Y\sigma_X \\ \rho\sigma_Y\sigma_X & \sigma_X^2 \end{pmatrix}\right]$$

where the parameters (μ_Y, σ_Y, μ_X, σ_X, ρ) are known or can be estimated from data.

Let $F_{Y,X}(y,x)$ also denote the joint normal cumulative distribution function (CDF), where $F_Y(\cdot)$ and $F_X(\cdot)$ are the marginal CDF. Assume that Y and X are discretized according to the rules $(-\infty < Y \leq y_1,\ y_1 < Y \leq y_2,\ \ldots,\ y_{K-1} < Y \leq y_K,\ y_K < Y < \infty)$ and $(-\infty < X \leq x_1,\ x_1 < X \leq x_2,\ \ldots,\ x_{J-1} < X \leq x_J,\ x_J < X < \infty)$.

Let $\mathbf{M}$ be the transition matrix between the discretized versions of Y and X,

$$\mathbf{M}_{J+1,K+1} = \begin{pmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,K+1} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,K+1} \\ \vdots & \vdots & \ddots & \vdots \\ p_{J+1,1} & p_{J+1,2} & \cdots & p_{J+1,K+1} \end{pmatrix}$$

where the generic element p_{jk} gives the probability that Y falls in the kth class given that X falls in the jth class.

$$\begin{aligned} p_{jk} &= \Pr(y_{k-1} < Y < y_k \mid x_{j-1} < X < x_j) \\ &= \frac{\Pr(y_{k-1} < Y < y_k,\ x_{j-1} < X < x_j)}{\Pr(x_{j-1} < X < x_j)} \\ &= \frac{F_{Y,X}(y_k, x_j) + F_{Y,X}(y_{k-1}, x_{j-1}) - F_{Y,X}(y_k, x_{j-1}) - F_{Y,X}(y_{k-1}, x_j)}{F_X(x_j) - F_X(x_{j-1})} \end{aligned}$$

The last equality exploits the Stata functions `binormal(`*h*`,` *k*`,` *r*`)` (for the numerator) and `normal(`*z*`)` (for the denominator).

1. The higher the trace, the higher the elements on the first diagonal and the stronger the association between a father's and a son's socioeconomic statuses.

3 The trabinor command

3.1 Description

Besides the five parameters of the bivariate normal distribution (μ_Y, σ_Y, μ_X, σ_X, ρ),[2] Y and X can be discretized by choosing either a given number of quantiles or a set of absolute boundaries. If the user sets a number Q of quantiles, then the resulting transition matrix will obviously be of size $Q \times Q$. If the user sets K absolute boundaries for Y and J absolute boundaries for X, then the resulting transition matrix will be rectangular of size $(J+1) \times (K+1)$.[3] In this case, boundaries for X must be chosen such that all the states of X have a nonzero probability of occurrence; otherwise, we will be conditioning on an impossible event.[4] A warning message appears if any of the marginal probabilities of (the discretized) X are smaller than $5{,}000{,}000^{-1}$; a second message appears if this probability is so small that the corresponding row of the transition matrix cannot be computed.

If neither the number of quantiles nor the absolute boundaries are specified, the `trabinor` command computes a 5×5 quantile matrix by default.

Results are displayed in percentage points, so the row sums of the transition matrix sum to 100.

3.2 Syntax

The syntax of `trabinor` is the following:

`trabinor` [, `quant(`#`)` `y(`*numlist*`)` `x(`*numlist*`)` `muy(`#`)` `sy(`#`)` `mux(`#`)` `sx(`#`)` `rho(`#`)` `format(`*string*`)`]

3.3 Options

`quant(`#`)` specifies in how many quantiles Y and X should be divided. The default is `quant(5)`. `quant(`#`)` must be an integer greater than 1 and cannot be combined with the options `y()` and `x()`.

`y(`*numlist*`)` sets the boundaries of the marginal distribution of Y. It cannot be combined with `quant()`.

2. The default values are those of a standard bivariate normal with $\rho = 0.5$.
3. To see this, imagine that we set the boundary 0 for both Y and X; then, both variables will be divided in two levels, namely, $(-\infty, 0)$ and $(0, \infty)$.
4. From a theoretical point of view, the normal distribution assigns a nonzero probability on any interval defined on the real line, but these probabilities can be very small. For example, if $\Phi(\cdot)$ denotes the standard normal CDF, then $\Phi(-5) \approx 0.00000028665 < 3,000,000^{-1}$. Therefore, a wrong choice of the boundaries might induce some issue of numerical approximation, given that the conditional probabilities are computed as the ratios between the joint and marginal probabilities.

`x(`*numlist*`)` sets the boundaries of the marginal distribution of X. It cannot be combined with `quant()`.

If none of the options `quant()`, `y()`, or `x()` are specified, `trabinor` computes a 5×5 quantile transition matrix (as if `quant(5)` were invoked). If only `y()` is specified, the same boundaries are applied to X; analogously, if only `x()` is specified, the same boundaries are applied to Y.

`muy(#)` defines the mean of Y. The default is `muy(0)`.

`sy(#)` defines the standard deviation of Y. The default is `sy(1)`.

`mux(#)` defines the mean of X. The default is `mux(0)`.

`sx(#)` defines the standard deviation of X. The default is `sx(1)`.

`rho(#)` defines the correlation coefficient between Y and X. The default is `rho(.5)`. This must be between -1 and 1.

`format(`*string*`)` controls how to display results. The default is `format(%9.3f)`.

3.4 Stored results

`trabinor` stores the following in `r()`:

Scalars

`r(muy)`	mean of Y	`r(sy)`	standard deviation of Y
`r(mux)`	mean of X	`r(sx)`	standard deviation of X
`r(rho)`	correlation coefficient between Y and X		

Matrices

`r(M)`	resulting transition matrix	`r(chk100)`	consistency check on `r(M)` (that is, row sums are equal to 100)
`r(PY)`	matrix of marginal probabilities of Y	`r(J)`	matrix of joint probabilities
`r(PX)`	matrix of marginal probabilities of X		

4 Examples

Here I discuss three examples of `trabinor`. Let $\mathbf{\Theta}_1$, $\mathbf{\Theta}_2$, and $\mathbf{\Theta}_3$ represent the five parameters of the bivariate normal distribution in each of the three examples.

In the first example, `trabinor` computes the quantile transition matrix of size 4 for a bivariate standard normal distribution with $\rho = 0.5$. Moreover, it uses the output program to obtain the trace of the transition matrix. Interpreting the results for the quantile transition matrix is straightforward: the probability that Y is smaller than its first quartile, given that X is smaller than its first quartile, is 48.1%. Analogously, using a formal notation, $\Pr\{F_Y^{-1}(0.50) < Y \leq F_Y^{-1}(0.75) \mid X \leq F_X^{-1}(0.25); \mathbf{\Theta}_1\} = 16.8\%$, and so on.

```
. trabinor, quant(4) format(%9.1f)
    Mean      Std. Dev.      Corr(Y,X)
-----------------------------------------------------------------
Y    0           1            .5
X    0           1
    Discretization of Y and X
-----------------------------------------------------------------
Y    4 quantiles of the marginal distribution of Y
X    4 quantiles of the marginal distribution of X
    Population Transition Matrix of Y given X
-----------------------------------------------------------------
symmetric M[4,4]
      y1    y2    y3    y4
x1  48.1  27.8  16.8   7.2
x2  27.8  29.6  25.8  16.8
x3  16.8  25.8  29.6  27.8
x4   7.2  16.8  27.8  48.1
-----------------------------------------------------------------
. scalar mytrace = trace(r(M))
. display mytrace
155.32692
```

In the second example, let's specify different parameters for the bivariate distribution. The variable X is discretized in 4 classes $[(-\infty,-3],(-3,0],(0,3],(3,\infty)]$ and Y in 6 classes $[(-\infty,2],(2,3],(3,5],(5,7.5],(7.5,10],(10,\infty)]$. From the output, we observe that $\Pr(Y<2 \mid X<-3;\mathbf{\Theta}_2)=0.4\%$, or that $\Pr(7.5\leq Y<10 \mid -3\leq X<0;\mathbf{\Theta}_2)=23.0\%$, and so on. We then look at the marginal probabilities for both variables, presented as row vectors: $\Pr(7.5\leq Y<10;\mathbf{\Theta}_2)=15.5\%$ and $\Pr(-3\leq X<0;\mathbf{\Theta}_2)=43.3\%$.

```
. trabinor, y(2 3 5(2.5)10) x(-3(3)3) sx(2) muy(5) sy(3) rho(-0.6) f(%9.1f)
    Mean      Std. Dev.      Corr(Y,X)
-----------------------------------------------------------------
Y    5           3            -.6
X    0           2
    Discretization of Y and X
-----------------------------------------------------------------
Y    2 3 5 7.5 10
X    -3 0 3
    Population Transition Matrix of Y given X
-----------------------------------------------------------------
M[4,6]
      y1    y2    y3    y4    y5    y6
x1   0.4   0.9   6.7  26.7  38.1  27.1
x2   5.0   5.7  22.2  38.1  23.0   6.1
x3  22.7  13.6  30.9  25.4   6.7   0.7
x4  57.6  14.8  19.6   7.3   0.8   0.0
-----------------------------------------------------------------
. matrix define py=r(PY)´
. matrix define px=r(PX)´
. matrix list py
py[1,6]
            r1         r2         r3         r4         r5         r6
c1   15.865525  9.3837284  24.750746  29.767162  15.453803  4.7790352
```

```
. matrix list px
px[1,4]
           r1          r2          r3          r4
c1  6.6807201   43.31928   43.31928  6.6807201
```

In the last example, we pass the empirical sample moments to `trabinor` using a dataset containing measures of blood pressure in 2 time periods (`bp_before` and `bp_after`) for a sample of 120 patients characterized by high blood pressure. Then, we look at the conditional probability of being in hypertension (`bp_after` ≥ 140) given the level of `bp_before`. We note that this conditional probability is 80% given hypertension in the past and 69% given no hypertension in the past.

```
. sysuse bpwide, clear
(fictional blood-pressure data)
. quietly summarize bp_before
. scalar mu_0 = r(mean)
. scalar sd_0 = r(sd)
. quietly summarize bp_after
. scalar mu_1 = r(mean)
. scalar sd_1 = r(sd)
. quietly correlate bp_before  bp_after
. scalar corr = r(rho)
. trabinor, muy(`=mu_1´) sy(`=sd_1´) mux(`=mu_0´) sx(`=sd_0´) rho(`=corr´)
> x(140)
    Mean      Std. Dev.      Corr(Y,X)
-------------------------------------------------------------------
Y    151.36      14.178       .15912
X    156.45      11.39
    Discretization of Y and X
-------------------------------------------------------------------
Y    140
X    140
    Population Transition Matrix of Y given X
-------------------------------------------------------------------
M[2,2]
        y1      y2
x1  30.654  69.346
x2  20.389  79.611
-------------------------------------------------------------------
```

5 Acknowledgments

I thank the editor and an anonymous referee for helpful comments; any remaining error is my responsibility. Any views expressed in this article are those of the author and do not necessarily reflect those of the Bank of Italy.

6 References

Black, S. E., and P. J. Devereux. 2011. Recent developments in intergenerational mobility. In *Handbook of Labor Economics*, ed. O. Ashenfelter and D. Card, vol. 4B, 1487–1541. San Diego: Elsevier.

O'Neill, D., O. Sweetman, and D. Van de gaer. 2007. The effects of measurement error and omitted variables when using transition matrices to measure intergenerational mobility. *Journal of Economic Inequality* 5: 159–178.

About the author

Marco Savegnago is a junior economist at the Bank of Italy.