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 teaching, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

For more information on the Stata Journal, including information for authors, see the web page

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:

Subscriptions mailed to US and Canadian addresses:

- 3-year subscription (includes printed and electronic copy) $153
- 2-year subscription (includes printed and electronic copy) $110
- 1-year subscription (includes printed and electronic copy) $ 59
- 1-year student subscription (includes printed and electronic copy) $ 35

Subscriptions mailed to other countries:

- 3-year subscription (includes printed and electronic copy) $225
- 2-year subscription (includes printed and electronic copy) $158
- 1-year subscription (includes printed and electronic copy) $ 83
- 1-year student subscription (includes printed and electronic copy) $ 59
- 3-year subscription (electronic only) $153

Back issues of the Stata Journal may be ordered online at

http://www.stata.com/bookstore/sj.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 email sj@stata.com.
Maximum likelihood estimation of endogenous switching regression models

Michael Lokshin  
The World Bank  
mlokshin@worldbank.org

Zurab Sajaia  
The World Bank and Stanford University  
zsajaia@worldbank.org

Abstract. This article describes the movestay Stata command, which implements the maximum likelihood method to fit the endogenous switching regression model.

Keywords: st0071, movestay, endogenous variables, maximum likelihood, limited dependent variables, switching regression

1 Introduction

In this article, we describe the implementation of the maximum likelihood (ML) algorithm to fit the endogenous switching regression model. In this model, a switching equation sorts individuals over two different states (with one regime observed). The econometric problem of fitting a model with endogenous switching arises in a variety of settings in labor economics, the modeling of housing demand, and the modeling of markets in disequilibrium. For example,

- The union–nonunion model of Lee (1978) investigates the joint determination of the extent of unionism and the effects of unions on wage rates. The propensity to join a union depends on the net wage gains that might result from trade union membership. This paper explicitly models the interdependence between the wage-gain equation and the union-membership equation.

- Adamchik and Bedi (1983) use data from Poland to examine whether there are any wage differentials of workers in the public and private sectors. This paper interprets sectoral wage differentials in terms of expected benefits and the desirability of working in a particular sector.

- Thorst (1977) models the housing-demand problem by examining the expenditures on housing services in owner-occupied and rental housing. The study models the individual decision to own or rent a house and the amount spent on housing services.

Models with endogenous switching can be fitted one equation at a time by either two-step least squares or maximum likelihood estimation. However, both of these estimation methods are inefficient and require potentially cumbersome adjustments to derive consistent standard errors. The movestay command, on the other hand, implements the full-information ML method (FIML) to simultaneously fit binary and continuous parts.
of the model in order to yield consistent standard errors. This approach relies on joint
normality of the error terms in the binary and continuous equations.

2 Methods

Consider the following model, which describes the behavior of an agent with two re-
gression equations and a criterion function, $I_i$, that determines which regime the agent
faces:\footnote{The discussion in this section draws from Maddala (1983, 223–224).}

\[
\begin{align*}
I_i &= 1 & \text{if } \gamma Z_i + u_i > 0 \\
I_i &= 0 & \text{if } \gamma Z_i + u_i \leq 0
\end{align*}
\]

Regime 1: \quad \begin{aligned}
y_{1i} &= \beta_1 X_{1i} + \epsilon_{1i} & \text{if } I_i = 1 \\
y_{2i} &= \beta_2 X_{2i} + \epsilon_{2i} & \text{if } I_i = 0
\end{aligned}
\]

Regime 2: \quad \begin{aligned}
y_{1i} &= \beta_1 X_{1i} + \epsilon_{1i} & \text{if } I_i = 1 \\
y_{2i} &= \beta_2 X_{2i} + \epsilon_{2i} & \text{if } I_i = 0
\end{aligned}
\]

Here, $y_{ji}$ are the dependent variables in the continuous equations; $X_{1i}$ and $X_{2i}$ are
vectors of weakly exogenous variables; and $\beta_1$, $\beta_2$, and $\gamma$ are vectors of parameters.
Assume that $u_i$, $\epsilon_{1i}$, and $\epsilon_{2i}$ have a trivariate normal distribution with mean vector zero
and covariance matrix

\[
\Omega = \begin{bmatrix}
\sigma_u^2 & \sigma_{1u} & \sigma_{2u} \\
\sigma_{1u} & \sigma_1^2 & . \\
\sigma_{2u} & . & \sigma_2^2
\end{bmatrix}
\]

where $\sigma_u^2$ is a variance of the error term in the selection equation, and $\sigma_1^2$ and $\sigma_2^2$ are
variances of the error terms in the continuous equations. $\sigma_{1u}$ is a covariance of $u_i$ and
$\epsilon_{1i}$, and $\sigma_{2u}$ is a covariance of $u_i$ and $\epsilon_{2i}$. The covariance between $\epsilon_{1i}$ and $\epsilon_{2i}$ is not
defined, as $y_{1i}$ and $y_{2i}$ are never observed simultaneously. We can assume that $\sigma_u^2 = 1$
($\gamma$ is estimable only up to a scalar factor). The model is identified by construction
through nonlinearities. Given the assumption with respect to the distribution of the
disturbance terms, the logarithmic likelihood function for the system of (1–2) is

\[
\ln L = \sum_i \left( I_i w_i \left[ \ln \left\{ F(\eta_{1i}) \right\} + \ln \left\{ f(\epsilon_{1i}/\sigma_1)/\sigma_1 \right\} \right] + (1 - I_i) w_i \left[ \ln \left\{ 1 - F(\eta_{2i}) \right\} + \ln \left\{ f(\epsilon_{2i}/\sigma_2)/\sigma_2 \right\} \right] \right)
\]

where $F$ is a cumulative normal distribution function, $f$ is a normal density distribution
function, $w_i$ is an optional weight for observation $i$, and
\[ \eta_{ji} = \frac{(\gamma Z_i + \rho_j \epsilon_{ji} / \sigma_j)}{\sqrt{1 - \rho_j^2}} \quad j = 1, 2 \]

where \( \rho_1 = \sigma_{1u}^2 / \sigma_{u} \sigma_1 \) is the correlation coefficient between \( \epsilon_{1i} \) and \( u_i \) and \( \rho_2 = \sigma_{2u}^2 / \sigma_{u} \sigma_2 \) is the correlation coefficient between \( \epsilon_{2i} \) and \( u_i \). To make sure that estimated \( \rho_1 \) and \( \rho_2 \) are bounded between \(-1\) and \(1\) and that estimated \( \sigma_1 \) and \( \sigma_2 \) are always positive, the maximum likelihood directly estimates \( \ln \sigma_1 \), \( \ln \sigma_2 \), and \( \text{atanh} \, \rho \):

\[ \text{atanh} \, \rho_j = \frac{1}{2} \ln \left( \frac{1 + \rho_j}{1 - \rho_j} \right) \]

After estimating the model’s parameters, the following conditional and unconditional expectations could be calculated:

Unconditional expectations:

\[ E(y_{1i} | x_{1i}) = x_{1i} \beta_1 \]  
\[ E(y_{2i} | x_{2i}) = x_{2i} \beta_2 \]

Conditional expectations:

\[ E(y_{1i} | I_i = 1, x_{1i}) = x_{1i} \beta_1 + \sigma_1 \rho_1 f(\gamma Z_i) / F(\gamma Z_i) \]  
\[ E(y_{1i} | I_i = 0, x_{1i}) = x_{1i} \beta_1 - \sigma_1 \rho_1 f(\gamma Z_i) / \{1 - F(\gamma Z_i)\} \]  
\[ E(y_{2i} | I_i = 1, x_{2i}) = x_{2i} \beta_2 + \sigma_2 \rho_2 f(\gamma Z_i) / F(\gamma Z_i) \]  
\[ E(y_{2i} | I_i = 0, x_{2i}) = x_{2i} \beta_2 - \sigma_2 \rho_2 f(\gamma Z_i) / \{1 - F(\gamma Z_i)\} \]

### 3 The movestay command

#### 3.1 Syntax

`movestay` is implemented as a d2 ML evaluator that calculates the overall log likelihood along with its first and second derivatives. The command allows for weights and robust estimation, as well as the full set of options associated with Stata’s maximum likelihood procedures. The generic syntax for the command is as follows:

\`movestay (depvar1 = varlist1) [(depvar2 = varlist2)] [if exp] [in range] [weight], select(depvar = varlist) [robust cluster(varname) maximize_options]`  

`pweights`, `fweights`, and `iweights` are allowed.
When the explanatory variables in the regressions are the same and there is only one dependent variable, only one equation need be specified. Alternatively, both equations must be specified when the set of exogenous variables in the first regression is different from the set of exogenous variables in the second regression or when the dependent variables are different between the two regressions.

The command `mspredict` can follow `movestay` to calculate the predictive statistics. The statistics are available both in and out of sample; type `mspredict ... if e(sample) ...` if wanted only for the estimation sample.

```
mspredict newvarname [if exp] [in range], statistic
```

where statistic is

- `psel` probability of being in regime 1; the default
- `xb1` linear prediction in regime 1
- `xb2` linear prediction in regime 2
- `yc1_1` expected value in the first equation conditional on the dependent variable being observed
- `yc1_2` expected value in the first equation conditional on the dependent variable not being observed
- `yc2_1` expected value in the second equation conditional on the dependent variable being observed
- `yc2_2` expected value in the second equation conditional on the dependent variable not being observed
- `mill1` Mills’ ratio in regime 1
- `mill2` Mills’ ratio in regime 2

### 3.2 Options

The `select(depvar = varlist)` specifies the switching equation for \( I_i \). `varlist` includes the set of instruments that help identify the model. The selection equation is estimated based on all exogenous variables specified in the continuous equations and instruments. If there are no instrumental variables in the model, the `depvar` must be specified as `select(depvar)`. In that case, the model will be identified by nonlinearities, and the selection equation will contain all the independent variables that enter in the continuous equations.

The `robust` specifies that the Huber/White/sandwich estimator of variance be used in place of the conventional MLE variance estimator. `robust` combined with `cluster()` allows observations that are not independent within cluster, although they must be independent between clusters. Specifying `pweights` implies `robust`. See [U] 23.14 Obtaining robust variance estimates.
cluster(varname) specifies that the observations are independent across groups (clusters) but not necessarily within groups. varname specifies the group to which each observation belongs; e.g., cluster(personid) refers to data with repeated observations on individuals. cluster() affects the estimated standard errors and variance–covariance matrix of the estimators (VCE) but not the estimated coefficients. cluster() can be used with pweights to produce estimates for unstratified cluster-sampled data. Specifying cluster() implies robust.

maximize_options control the maximization process; see [R] maximize. With the possible exception of iterate(0) and trace, you should only have to specify them if the model is unstable.

3.3 Options for mspredict

One of the following statistics can be specified with the mspredict command:

psel calculates the probability of being in regime 1. This is the default statistic.

xb1 calculates the linear prediction for the regression equation in regime 1. This is the unconditional prediction referred to in Methods (3).

xb2 calculates the linear prediction for the regression equation in regime 2. This is the unconditional prediction referred to in Methods (4).

yc1.1 calculates the expected value of the dependent variable in the first equation conditional on the dependent variable being observed ((5) in Methods).

yc1.2 calculates the expected value of the dependent variable in the first equation conditional on the dependent variable not being observed ((6) in Methods).

yc2.2 calculates the expected value of the dependent variable in the second equation conditional on the dependent variable being observed ((7) in Methods).

yc2.1 calculates the expected value of the dependent variable in the second equation conditional on the dependent variable not being observed ((8) in Methods).

mills1 and mills2 calculate corresponding Mills’ ratios for the two regimes.

4 Example

We will illustrate the use of the movestay command by looking at the problem of estimating individual earnings in the public and private sectors. A typical specification might be the following:

\[
\begin{align*}
\ln w_{1i} &= X_i \beta_1 + \epsilon_{1i} \\
\ln w_{2i} &= X_i \beta_2 + \epsilon_{2i} \\
I_i^* &= \delta(\ln w_{1i} - \ln w_{2i}) + Z_i \gamma + u_i
\end{align*}
\]
Here $I_i^*$ is a latent variable that determines the sector in which individual $i$ is employed; $w_{ij}$ is the wage of individual $i$ in sector $j$; $Z_i$ is a vector of characteristics that influences the decision regarding sector of employment. $X_i$ is a vector of individual characteristics that is thought to influence individual wage. $\beta_1$, $\beta_2$, and $\gamma$ are vectors of parameters, and $u_i$, $\epsilon_1$, and $\epsilon_2$ are the disturbance terms. The observed dichotomous realization $I_i$ of latent variable $I_i^*$ of whether the individual $i$ is employed in a particular sector has the following form:

$$I_i = 1 \quad \text{if} \quad I_i^* > 0$$
$$I_i = 0 \quad \text{otherwise} \quad (12)$$

The assumption that is often made in this type of model is that the sector of employment is endogenous to wages. Some unobserved characteristics that influence the probability to choose a particular sector of employment could also influence the wages the individual receives once he is employed. Neglecting these selectivity effects is likely to give a false picture of the relative earning positions in both the public and private sectors. The simultaneous ML estimation (9–12) corrects for the selection bias in sectoral wage estimates.

In our example, the sector choice indicator private takes value 1 if the individual is employed in the private sector and 0 if in the public sector. The wage equations (9–10) estimate log of monthly individual earnings, lmo.earn. The exogenous variables in the wage regressions (9–10) are based on a typical Mincer’s type specification (Mincer and Polachek 1974) and include such individual characteristics as age, age$^2$, education, and regional dummies. In addition to these variables, the sector selection equation (11) includes two variables to improve identification. An individual’s marital status and the number of jobholders in the household are believed to influence an individual’s choice of the sector of employment but not affect the wages. The ML estimation of this specification using the movestay command and the dataset movestay_example.dta is shown below:

> movestay_example, clear
(Sample dataset to illustrate the use of movestay procedure)

(Continued on next page)
**ML estimation of endogenous switching regression models**

```stata
. movestay lmo_wage age age2 edu13 edu4 edu5 reg2 reg3 reg4, 
> select(private = m_s1 job_hold)
```

Fitting initial values ......

```
Iteration 0: log likelihood = -2504.2563
(iteration output omitted)
```

Endogenous switching regression model

| Coef. Std. Err. z P>|z|  [95% Conf. Interval] |
|-----------------|-----------------|--------|------------------|-----------------|-----------------|
| **lmo_wage_1**  |                 |        |                  |                  |                  |
| age             | 0.0423471       | 0.0291874 | 1.45         | 0.147         | -0.0148592 - 0.0995534 |
| age2            | -0.0005007      | 0.0003227  | -1.55       | 0.121         | -0.0011332 - 0.0000139 |
| edu13           | 0.9437058       | 0.2793217  | 3.36        | 0.001         | 0.6699977 - 1.2174174 |
| edu4            | -1.579071       | 0.1629109  | -9.75       | 0.000         | -2.909966 - -0.2481769 |
| edu5            | -1.640599       | 0.1302278  | -12.69      | 0.000         | -3.002625 - -1.278573 |
| reg2            | -0.2864941      | 0.1099711  | -2.59       | 0.010         | -0.5016416 - -0.0713466 |
| reg3            | 0.7079698       | 0.1427093  | 4.96        | 0.000         | 0.42799170 - 0.9874021 |
| reg4            | -1.333714       | 0.1414171  | -9.45       | 0.000         | -1.615438 - -1.05199 |
| _cons           | 7.415686        | 0.4808005  | 15.42       | 0.000         | 6.473334 - 8.358037  |

| **lmo_wage_0**  |                 |        |                  |                  |                  |
| age             | -0.0370404      | 0.0111445 | -3.32      | 0.001         | -0.0588832 - -0.0151976 |
| age2            | 0.0030517       | 0.0002827  | 1.08        | 0.278         | 0.0002128 - 0.0000625 |
| edu13           | -0.5066122      | 0.0858002  | -5.94       | 0.000         | -0.680694 - -0.3331459 |
| edu4            | -0.410602       | 0.0507909  | -8.08       | 0.000         | -0.5101503 - -0.3110537 |
| edu5            | -0.2973613      | 0.0391875  | -7.59       | 0.000         | -0.3741673 - -0.2205552 |
| reg2            | -0.3780673      | 0.0420395  | -8.99       | 0.000         | -0.4604562 - -0.2956785 |
| reg3            | 0.7053256       | 0.0532104  | 13.26       | 0.000         | 0.601035 - 0.8096161  |
| reg4            | -2.355433       | 0.0474621  | -49.60      | 0.000         | -3.285673 - -1.425193  |
| _cons           | 9.322335        | 0.2372244  | 39.21       | 0.000         | 8.856404 - 9.788267  |

| **private**     |                 |        |                  |                  |                  |
| age             | -1.1455149      | 0.2589292 | -5.62      | 0.000         | -1.696222 - -0.9476767 |
| age2            | 0.0013623       | 0.0003054  | 4.47        | 0.000         | 0.0007655 - 0.0019592 |
| edu13           | 0.7081378       | 0.2547816  | 2.79        | 0.006         | 0.4055393 - 0.5790685 |
| edu4            | 0.0690438       | 0.1415167  | 0.49        | 0.626         | 0.2083238 - 0.3464113 |
| edu5            | 0.2353146       | 0.1063559  | 2.21        | 0.027         | 0.026681 - 0.4358833 |
| reg2            | -0.4401675      | 0.0958095  | -4.59       | 0.000         | -0.6279508 - -0.2523843 |
| reg3            | -0.5960669      | 0.1187269  | -5.02       | 0.000         | -0.8287674 - -0.3635664 |
| reg4            | -0.6010513      | 0.1127811  | -5.33       | 0.000         | -0.8220981 - -0.3800046 |
| m_s1            | 0.1569925       | 0.0921425  | 1.70        | 0.088         | 0.0236035 - 0.3758858 |
| job_hold        | 0.0519388       | 0.0367121  | 1.53        | 0.127         | -0.015702 - 0.1260898 |
| _cons           | 2.5054744       | 0.5789999  | 4.33        | 0.000         | 1.370677 - 3.640272  |

| **/lns1**       |                 |        |                  |                  |                  |
|                 | -0.5903432      | 0.0562427  | -10.50      | 0.000         | -0.7005769 - -0.4801095 |
| **/lns2**       |                 |        |                  |                  |                  |
|                 | -0.4220208      | 0.0186565  | -22.62      | 0.000         | -0.4585869 - -0.3854546 |
| **/r1**         |                 |        |                  |                  |                  |
|                 | 0.1456952       | 0.0395504  | 3.66        | 0.000         | -0.208612 - 0.7720024 |
| **/r2**         |                 |        |                  |                  |                  |
|                 | 1.353759        | 0.0813975  | 16.63       | 0.000         | 1.194222 - 1.513295  |

| **sigma_1**     |                 |        |                  |                  |                  |
|                 | 0.5541371       | 0.0311662  | 17.44        | 0.000         | 0.4962989 - 0.6187156 |
| **sigma_2**     |                 |        |                  |                  |                  |
|                 | 0.6557204       | 0.0122335  | 53.93        | 0.000         | 0.631763 - 0.6801414 |
| **rho_1**       |                 |        |                  |                  |                  |
|                 | 0.144673        | 0.0328621  | 4.47        | 0.000         | -0.4467336 - 0.6480923 |
| **rho_2**       |                 |        |                  |                  |                  |
|                 | 0.8749375       | 0.0190864  | 46.93        | 0.000         | 0.8318838 - 0.907522 |

LR test of indep. eqns. : chi2(1) = 86.94 Prob > chi2 = 0.0000
The results of the sector selection equation are reported in the section of the output headed private. The results of the wage regression in the private sector are reported in the \texttt{lmo\_wage\_1} section, and the wage regression in the public sector is reported in the \texttt{lmo\_wage\_0} section.

The correlation coefficients $\rho_1$ and $\rho_2$ are both positive but are significant only for the correlation between the sector choice equation and the public sector wage equation. Since $\rho_2$ is positive and significantly different from zero, the model suggests that individuals who choose to work in the public sector earn lower wages in that sector than a random individual from the sample would have earned, and those working in the private sector do no better or worse than a random individual. The likelihood-ratio test for joint independence of the three equations is reported in the last line of the output.

The variables $\sigma_1$, $\ln s_1$, $\ln s_2$, $r_1$, and $r_2$ are ancillary parameters used in the maximum likelihood procedure. $\sigma_1$ and $\sigma_2$ are the square roots of the variances of the residuals of the regression part of the model, and $\ln s_1$ is its log. $r_1$ and $r_2$ are the transformation of the correlation between the errors from the two equations.

5 References


About the Authors

Michael Lokshin is a Senior Economist at the Research Department of the World Bank.

Zurab Sajaia is working on his PhD in Economics at Stanford University.