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

Keywords: 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 estimate 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 estimating 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. The paper explicitly models the interdependence between the wage gain equation and the union membership equation.
- Adamchik and Bedi (2000) use data from Poland to examine whether there are any wage differentials of workers in the public and private sectors. The 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 estimated one equation at a time either by two-step least square or maximum likelihood estimation. However, both of these estimation methods are inefficient. In addition, these approaches 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 estimate 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 regression equations, and a criterion function Ii that determines which regime the agent faces:

\[
I_i = 1 \quad \text{if} \quad \gamma Z_i + u_i > 0 \\
I_i = 0 \quad \text{if} \quad \gamma Z_i + u_i \leq 0
\]

Regime 1: \( y_{1i} = \beta_1 X_{1i} + \varepsilon_{1i} \) \( \text{if} \ I_i = 1 \) \hspace{1cm} (2.2)

Regime 2: \( y_{2i} = \beta_2 X_{2i} + \varepsilon_{2i} \) \( \text{if} \ I_i = 0 \) \hspace{1cm} (2.3)

Here, \( y_{ji} \) are the dependent variables in the continuous equations, \( X_1 \) and \( X_2 \) are vectors of weakly exogenous variables, and \( \beta_1, \beta_2, \) and \( \gamma \) are vectors of parameters. Assume that \( u_i, \varepsilon_{1i} \) and \( \varepsilon_{2i} \) have a trivariate normal distribution, with mean vector zero and covariance matrix:

\[
\Omega = \begin{bmatrix}
\sigma_u^2 & \cdots \\
\sigma_{21} & \sigma_1^2 & \cdots \\
\sigma_{31} & \cdots & \sigma_2^2
\end{bmatrix}
\]

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. \( \sigma_{21} \) is a covariance of \( u_i \) and \( \varepsilon_{1i} \) and \( \sigma_{31} \) is a covariance of \( u_i \) and \( \varepsilon_{2i} \). The covariance between \( \varepsilon_{1i} \) and \( \varepsilon_{2i} \) 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 equations (2.2-2.3) is:

\[
\ln L = \sum_{i=1} \{ I_i w_i [\ln(F(\eta_{1i})) + \ln(f(\varepsilon_{1i}/\sigma_1)/\sigma_1) + \\
(1-I_i) w_i [\ln(1-F(\eta_{2i})) + \ln(f(\varepsilon_{2i}/\sigma_2)/\sigma_2)]] \}
\]

(2.4)

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

\[
\eta_{ji} = \frac{(\gamma Z_i + \rho_j \varepsilon_{ji}/\sigma_j)}{\sqrt{1-\rho_j^2}} \quad j = 1,2
\]

\[\text{1 The discussion in this section draws from Maddala (1983) p. 223-224.}\]
where $\rho_1 = \frac{\sigma_{11}}{\sigma_1\sigma_1}$ is the coefficient of correlation between $\varepsilon_1$ and $u$ and $\rho_2 = \frac{\sigma_{31}}{\sigma_1\sigma_2}$ are the coefficients of correlation between $\varepsilon_2$ and $u$. To make sure that estimated $\rho_1$, $\rho_2$ are bounded between $-1$ and $1$ and estimated $\sigma_1$, $\sigma_2$ 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:

$$E(y_{it} \mid x_{it}) = x_{it}\beta_i \quad (2.5)$$

$$E(y_{2i} \mid x_{2i}) = x_{2i}\beta_2 \quad (2.6)$$

Conditional expectations:

$$E(y_{it} \mid I_i = 1, x_{it}) = x_{it}\beta_i + \sigma_1 \rho_1 f(\gamma Z_i)/F(\gamma Z_i) \quad (2.7)$$

$$E(y_{it} \mid I_i = 0, x_{it}) = x_{it}\beta_2 - \sigma_2 \rho_2 f(\gamma Z_i)/(1 - F(\gamma Z_i)) \quad (2.8)$$

$$E(y_{2i} \mid I_i = 1, x_{2i}) = x_{2i}\beta_i + \sigma_2 \rho_2 f(\gamma Z_i)/F(\gamma Z_i) \quad (2.9)$$

$$E(y_{2i} \mid I_i = 0, x_{2i}) = x_{2i}\beta_2 - \sigma_2 \rho_2 f(\gamma Z_i)/(1 - F(\gamma Z_i)) \quad (2.10)$$

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, robust estimation, as well as the full set of options associated with Stata’s maximum likelihood procedures. The generic syntax for the command is:

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

Pweights, fweights and iweights are allowed.

In cases when the explanatory variables in the regressions are the same and there is only one dependent variable, only one equation need be specified. Alternatively, when the set of exogenous variables in the first regression is different from the set of exogenous variables in the second regression and/or the dependent variables are different between the two regressions, both equations must be specified.
The command `mspredict` can follow `movestay` to calculate the predictive statistics. The statistics could be both in and out of the sample; type "`mspredict ... if e(sample)`" if statistics are wanted only for the estimation sample.

`mspredict newvarname [if] [in range], statistics`

### 3.2 General options

`select(depvar_s=varlist_s)` gives the specification of switching equation for $I$, `varlist_s` includes the set of instruments that help identify the model. It is an integral part of the `movestay` estimation and is not optional. The selection equation is estimated based on all exogenous variables specified in the continuous equations plus instruments. If there are no instrumental variables in the model, the `depvar_s` must be specified as `select(depvar_s)`. In that case the model will be identified by non-linearities and the selection equation will contain all the independent variables that enter in the continuous equations.

`robust` specifies that the Huber/White/sandwich estimator of the variance is to be used in place of the conventional MLE variance estimator. `robust` combined with `cluster()` further allows observations which are not independent within cluster (although they must be independent between clusters). If you specify `pweights`, `robust` is implied. 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 to which group 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 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 the Methods section (Equation 2.5).
xb2 calculates the linear prediction for the regression equation in regime 2. This is the unconditional prediction referred to in the Methods section (Equation 2.6).

yc1_1 calculates the expected value of the dependent variable in the first equation conditional on the dependent variable being observed (Equation 2.7).

yc1_2 calculates the expected value of the dependent variable in the first equation conditional on the dependent variable not being observed (Equation 2.8).

yc2_2 calculates the expected value of the dependent variable in the second equation conditional on the dependent variable being observed (Equation 2.9).

yc2_1 calculates the expected value of the dependent variable in the second equation conditional on the dependent variable not being observed (Equation 2.10).

mills1 and mills2 calculate corresponding Mill’s 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:

\[
\ln w_{ji} = X_i \beta_1 + \varepsilon_{1i} \\
\ln w_{2i} = X_i \beta_2 + \varepsilon_{2i} \\
I_i^* = \delta (\ln w_{ji} - \ln w_{2i}) + Z_i \gamma + u_i
\]

Here \( I_i^* \) is a latent variable that determines the sector in which individual \( i \) is employed; \( w_{ji} \) 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 \( u_i, \varepsilon_1 \) and \( \varepsilon_2 \) 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:

\[
I_i = 1 \text{ if } I_i^* > 0 \\
I_i = 0 \text{ otherwise }
\]

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 of equations (4.1-4.4) 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 she is employed in the public sector. The wage equations (4.1-4.2) estimate log of monthly individual earnings: `lmo_earn`. The set of exogenous variables in the wage regressions (4.1-4.2) are based on typical Minser’s type specification (Minser and Polachek, 1974) and includes such individual characteristics as age, age², educational, and regional dummies. In addition to these variables, the sector selection equation (4.3) includes two variables to improve identification. An individual’s marital status and the number of jobholders in the household are believed to influence individual’s choice of the sector of employment, but not to affect the wages. The ML estimation of this specification using `movestay` command and the dataset `movestay_example.dta` is shown below:

```stata
. use movestay_example, clear
. local str age age2 edu13 edu4 edu5 reg2 reg3 reg4
. movestay (lmo_wage = `str'), select(private= m_s1 job_hold)

Fitting initial values ......
Iteration 0:   log likelihood = -2504.2563
. . . .Iteration output omitted . . . .
. . . .Iteration 6:   log likelihood = -2470.9304

Endogenous switching regression model                  Number of obs   =       2094
Wald chi2(8)    =     102.43
Log likelihood = -2470.9304                         Prob > chi2     =     0.0000

------------------------------------------------------------------------------
|      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lmo_wage_1   |
    age |   .0423471   .0291874     1.45   0.147    -.0148592    .0995534
   age2 |  -.0005007   .0003227    -1.55   0.121    -.0011332    .0001319
  edu13 |   .3437058   .2793217     1.23   0.219    -.2037546    .8911661
   edu4 |  -.1578071   .1608109    -0.98   0.326    -.4729756    .1573654
   edu5 |  -.164094    .1300289    -1.26   0.207    -.4189461    .0907580
   reg2 |  -.2864941   .1097711    -2.61   0.009    -.5016416   -.0713466
   reg3 |   .7076968   .1427093     4.96   0.000     .4279917     .9874020
   reg4 |  -.1383714   .1414171    -0.98   0.328    -.4155438    .1388009
    _cons |   7.415686   .4808005    15.42   0.000     6.473334    8.358037
-------------+----------------------------------------------------------------
lmo_wage_0   |
    age |  -.0370404   .0111445    -3.32   0.001    -.0588832   -.0151976
   age2 |   .0003735   .0001285     2.91   0.004      .0001216    .0006255
  edu13 |  -.5066122   .0885002    -5.72   0.000    -.6800694   -.3331549
   edu4 |  -.410602    .0507909    -8.08   0.000    -.5101503    -.3105372
   edu5 |  -.2973613   .1300289    -2.26   0.024    -.5516762   -.0430571
   reg2 |  -.3780673   .1427093    -2.66   0.007    -.6546146   -.1015641
   reg3 |   .7053256   .0532104    13.26   0.000     .5980967    .8125545
   reg4 |  -.2355433   .0474621    -4.96   0.000    -.3285673   -.1425193
    _cons |   9.322335   .4808005    19.41   0.000     8.383334    10.26134
-------------+----------------------------------------------------------------
        private   |
    age |  -.1455149   .0258922    -5.62   0.000    -.1962622   -.0947676
   age2 |   .0013623   .0003024     4.47   0.000      .0007655    .0019592
  edu13 |   .0761837   .2457811     0.31   0.757    -.4055393    .5579068
   edu4 |   .0690438   .1415167     0.49   0.626    -.2083238    .3464113
-------------+----------------------------------------------------------------
```

6
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 “lmo_wage_1” section and the wage regression in the private sector is outputted in the “lmo_wage_0” section.

The correlation coefficients $\rho_1$ and $\rho_2$ are both positive, but 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 $\sigma$, $\ln s_1$, $\ln s_2$, $r_1$, 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 $\ln s_i$ is its log. $r_1$ and $r_2$ are the transformation of the correlation between the errors from the two equations.

5 References

