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 regression equations and a criterion function, I_i, that determines which regime the agent faces[1]:

$$
\begin{aligned}
I_i &= 1 && \text{if } \gamma Z_i + u_i > 0 \\
I_i &= 0 && \text{if } \gamma Z_i + u_i \leq 0 \\
\text{Regime1}: \quad y_{1i} &= \beta_1 X_{1i} + \epsilon_{1i} && \text{if } I_i = 1 & (1) \\
\text{Regime2}: \quad y_{2i} &= \beta_2 X_{2i} + \epsilon_{2i} && \text{if } I_i = 0 & (2)
\end{aligned}
$$

Here, y_{ji} are the dependent variables in the continuous equations; X_{1i} and X_{2i} are vectors of weakly exogenous variables; and β_1, β_2, and γ are vectors of parameters. Assume that u_i, ϵ_{1i}, and ϵ_{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 σ_u^2 is a variance of the error term in the selection equation, and σ_1^2 and σ_2^2 are variances of the error terms in the continuous equations. σ_{1u} is a covariance of u_i and ϵ_{1i}, and σ_{2u} is a covariance of u_i and ϵ_{2i}. The covariance between ϵ_{1i} and ϵ_{2i} is not defined, as y_{1i} and y_{2i} are never observed simultaneously. We can assume that $\sigma_u^2 = 1$ (γ 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

$$
\begin{aligned}
\ln L = \sum_i \Bigg(I_i w_i \Big[\ln\big\{ F(\eta_{1i}) \big\} + \ln\big\{ f(\epsilon_{1i}/\sigma_1)/\sigma_1 \big\} \Big] + \\
(1 - I_i) w_i \Big[\ln\big\{ 1 - F(\eta_{2i}) \big\} + \ln\big\{ f(\epsilon_{2i}/\sigma_2)/\sigma_2 \big\} \Big] \Bigg)
\end{aligned}
$$

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

---

[1]The discussion in this section draws from Maddala (1983, 223–224).

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

where $\rho_1 = \sigma_{1u}^2/\sigma_u\sigma_1$ is the correlation coefficient between ϵ_{1i} and u_i and $\rho_2 = \sigma_{2u}^2/\sigma_u\sigma_2$ is the correlation coefficient between ϵ_{2i} and u_i. To make sure that estimated ρ_1 and ρ_2 are bounded between -1 and 1 and that estimated σ_1 and σ_2 are always positive, the maximum likelihood directly estimates $\ln\sigma_1$, $\ln\sigma_2$, and atanh ρ:

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

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

Unconditional expectations:

$$E(y_{1i}|x_{1i}) = x_{1i}\beta_1 \tag{3}$$
$$E(y_{2i}|x_{2i}) = x_{2i}\beta_2 \tag{4}$$

Conditional expectations:

$$E(y_{1i}|I_i = 1, x_{1i}) = x_{1i}\beta_1 + \sigma_1\rho_1 f(\gamma Z_i)/F(\gamma Z_i) \tag{5}$$
$$E(y_{1i}|I_i = 0, x_{1i}) = x_{1i}\beta_1 - \sigma_1\rho_1 f(\gamma Z_i)/\{1 - F(\gamma Z_i)\} \tag{6}$$
$$E(y_{2i}|I_i = 1, x_{2i}) = x_{2i}\beta_2 + \sigma_2\rho_2 f(\gamma Z_i)/F(\gamma Z_i) \tag{7}$$
$$E(y_{2i}|I_i = 0, x_{2i}) = x_{2i}\beta_2 - \sigma_2\rho_2 f(\gamma Z_i)/\{1 - F(\gamma Z_i)\} \tag{8}$$

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*], <u>sel</u>ect(*depvarₛ* = *varlistₛ*) [<u>robust</u> <u>cl</u>uster(*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

<u>p</u>sel	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_2	expected value in the second equation conditional on the dependent variable being observed
yc2_1	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

`select(`*depvar_s* `=` *varlist_s*`)` specifies the switching equation for I_i. *varlist_s* 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_s* must be specified as `select(`*depvar_s*`)`. 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.

`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 `pweight`s 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:

$$\ln w_{1i} = X_i \beta_1 + \epsilon_{1i} \tag{9}$$
$$\ln w_{2i} = X_i \beta_2 + \epsilon_{2i} \tag{10}$$
$$I_i^* = \delta(\ln w_{1i} - \ln w_{2i}) + Z_i \gamma + u_i \tag{11}$$

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. β_1, β_2, and γ are vectors of parameters, and u_i, ϵ_1, and ϵ_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:

$$
\begin{aligned}
I_i &= 1 \qquad \text{if } I_i^* > 0 \\
I_i &= 0 \qquad \text{otherwise}
\end{aligned}
\tag{12}
$$

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:

```
. use http://www.worldbank.org/research/projects/poverty/programs/
> movestay_example, clear
(Sample dataset to illustrate the use of movestay procedure)
```

```
. 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		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	-.4729906	.1573763
edu5	-.164094	.1300289	-1.26	0.207	-.4189461	.090758
reg2	-.2864941	.1097711	-2.61	0.009	-.5016416	-.0713466
reg3	.7076968	.1427093	4.96	0.000	.4279917	.987402
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	-.3110537
edu5	-.2973613	.0391875	-7.59	0.000	-.3741673	-.2205552
reg2	-.3780673	.0420359	-8.99	0.000	-.4604562	-.2956785
reg3	.7053256	.0532104	13.26	0.000	.601035	.8096161
reg4	-.2355433	.0474621	-4.96	0.000	-.3285673	-.1425193
_cons	9.322335	.2377244	39.21	0.000	8.856404	9.788267
private						
age	-.1455149	.025892	-5.62	0.000	-.1962622	-.0947676
age2	.0013623	.0003045	4.47	0.000	.0007655	.0019592
edu13	.0761837	.2457816	0.31	0.757	-.4055393	.5579068
edu4	.0690438	.1415167	0.49	0.626	-.2083238	.3464113
edu5	.2351346	.1063559	2.21	0.027	.026681	.4435883
reg2	-.4401675	.0958095	-4.59	0.000	-.6279508	-.2523843
reg3	-.5960669	.1187269	-5.02	0.000	-.8287674	-.3633664
reg4	-.6010513	.112781	-5.33	0.000	-.8220981	-.3800046
m_s1	.1569925	.0921425	1.70	0.088	-.0236035	.3375885
job_hold	.0551938	.0361721	1.53	0.127	-.0157022	.1260898
_cons	2.505474	.578989	4.33	0.000	1.370677	3.640272
/lns1	-.5903432	.0562427	-10.50	0.000	-.7005769	-.4801095
/lns2	-.4220208	.0186565	-22.62	0.000	-.4585869	-.3854546
/r1	.1456952	.3195504	0.46	0.648	-.480612	.7720024
/r2	1.353759	.0813975	16.63	0.000	1.194222	1.513295
sigma_1	.5541371	.0311662			.4962989	.6187156
sigma_2	.6557204	.0122335			.6321763	.6801414
rho_1	.144673	.3128621			-.4467336	.6480923
rho_2	.8749375	.0190864			.8318838	.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 `lmo_wage_1` section, and the wage regression in the public sector is reported in the `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`, `/lns1`, `/lns2`, `/r1`, and `/r2` 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 `lnsig` is its log. `/r1` and `/r2` are the transformation of the correlation between the errors from the two equations.

5 References

Adamchik, V. and V. Bedi. 1983. Wage differentials between the public and the private sectors: Evidence from an economy in transition. *Labour Economics* 7: 203–224.

Lee, L. 1978. Unionism and wage rates: A simultaneous equations model with qualitative and limited dependent variables. *International Economic Review* 19: 415–433.

Maddala, G. S. 1983. *Limited-Dependent and Qualitative Variables in Econometrics.* Cambridge: Cambridge University Press.

Mincer, J. and S. Polachek. 1974. Family investment in human capital: Earnings of women. *Journal of Political Economy (Supplement)* 82: S76–S108.

Thorst, R. 1977. *Demand for housing: A model based on inter-related choices between owning and renting.* Ph.D. dissertation, University of Florida.

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.